Discrete Structures

California State University, Fullerton Department of Mathematics

Last Revised: April 23, 2018

This text is a compilation of material to be used in the Math 270A: Discrete Structures I course at California State University, Fullerton.

The material in this text is taken from the following sources:

A Spiral Workbook for Discrete Mathematics, by Harris Kwong. Licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 3.0 Unported license. Abbreviated in this text as **SWDM**.

Discrete Mathematics: An Open Introduction, by Oscar Levin, 2nd edition. Licensed under a Creative Commons Attribution-ShareAlike 4.0 license. Abbreviated in this text as **DMOI**.

Mathematics for Computer Science, by MIT OpenCourseware, as taught in Fall 2010 by Prof. Tom Leighton and Dr. Marten van Dijk. Licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 4.0 license. Abbreviated in this text as **MFCS**.

Contents

1 SWDM 2: Logic 5

2 SWDM 3: Proof Techniques 41

3 SWDM 4: Sets 81

4 SWDM 6: Functions 119

5 SWDM 7: Relations 161

6 DMOI 1: Counting 185

7 DMOI 2: Sequences 235

8 MFCS 10: Recurrences 261

9 MFCS 11: The Pigeonhole Principle 279

10 MFCS 5: Graph Theory 285

11 MFCS 6: Directed Graphs 309

12 MFCS 14: Events and Probability Spaces 315

13 MFCS 15: Conditional Probability 341

14 MFCS 16: Independence 349

15 MFCS 17: Random Variables and Distributions 355

Chapter 1

SWDM 2: Logic

Chapter 2

Logic

2.1 Propositions

The rules of logic allow us to distinguish between valid and invalid arguments. Besides mathematics, logic has numerous applications in computer science, including the design of computer circuits and the construction of computer programs. To analyze whether a certain argument is valid, we first extract its syntax.

Example 2.1.1 These two arguments:

- If $x + 1 = 5$, then $x = 4$. Therefore, if $x \neq 4$, then $x + 1 \neq 5$.
- If I watch Monday night football, then I will miss the following Tuesday 8 A.M. class. Therefore, if I do not miss my Tuesday 8 A.M. class, then I did not watch football the previous Monday night.

use the same format:

If p then q. Therefore, if q is false then p is false.

If we can establish the validity of this type of argument, then we have proved *at once* that both arguments are legitimate. In fact, we have also proved that any argument using the same format is also credible. ▲

Hands-On Exercise 2.1.1 Can you give another argument that uses the same format in the last example?

△

In mathematics, we are interested in statements that can be proved or disproved. We define a ***proposition*** (sometimes called a ***statement***, or an ***assertion***) to be a sentence that is either true or false, but not both.

Example 2.1.2 The following sentences:

- Barack Obama is the president of the United States.
- $2 + 3 = 6$.

are propositions, because each of them is either true or false (but not both). ▲

Example 2.1.3 These two sentences:

- Ouch!
- What time is it?

are not propositions because they do not proclaim anything; they are exclamation and question, respectively. ▲

Example 2.1.4 Explain why the following sentences are *not* propositions:

(a) $x + 1 = 2$.

(b) $x - y = y - x$.

(c) $A^2 = 0$ implies $A = 0$.

Solution: (a) This equation is not a statement because we cannot tell whether it is true or false unless we know the value of x. It is true when $x = 1$; it is false for other x-values. Since the sentence is sometimes true and sometimes false, it cannot be a statement.

(b) For the same reason, since $x - y = y - x$ is sometimes true and sometimes false, it cannot be a statement.

(c) This looks like a statement because it appears to be true all the time. Yet, this is *not* a statement, because we never say what A represents. The claim is true if A is a real number, but it is not always true if A is a matrix[1]. Thus, it is not a proposition. ▲

Hands-On Exercise 2.1.2 Explain why these sentences are not propositions:

(a) He is the quarterback of our football team.

(b) $x + y = 17$.

(c) $AB = BA$.

△

Example 2.1.5 Although the sentence "$x + 1 = 2$" is not a statement, we can change it into a statement by adding some condition on x. For instance, the following is a true statement:

For some real number x, we have $x + 1 = 2$.

and the statement

For all real numbers x, we have $x + 1 = 2$.

is false. The parts of these two statements that say "for some real number x" and "for all real numbers x" are called quantifiers. We shall study them in Section 2.6. ▲

[1]Some students may not be familiar with matrices. A matrix is rectangular array of numbers. Matrices are important tools in mathematics. The product of two matrices of appropriate sizes is defined in a rather unusual way. It is the peculiar way that two matrices are multiplied that makes matrices so useful in mathematics. The square of a matrix is of course the product of the matrix with itself. It is well-defined only when the matrix is a square matrix. As it turns out, the order of multiplication of two matrices is important. In other words, given any two matrices A and B, it is not always true that $AB = BA$.

Example 2.1.6 Saying that

"A statement is not a proposition if we *cannot* decide whether it is true or false."

is different from saying that

"A statement is not a proposition if we do not know *how* to verify whether it is true or false."

The more important issue is whether the truth value of the statement can be determined in theory. Consider the sentence

Every even integer greater than 2 can be written as the sum of two primes.

Nobody has ever proved or disproved this claim, so we do not know whether it is true or false, even though computational data suggest it is true. Nevertheless, it *is* a proposition because it is either true or false but not both. It is impossible for this sentence to be true sometimes, and false at other times. With the advancement of mathematics, someone may be able to either prove or disprove it in the future. The example above is the famous ***Goldbach Conjecture***, which dates back to 1742. ▲

We usually use the lowercase letters p, q and r to represent propositions. This can be compared to using variables x, y and z to denote real numbers. Since the truth values of p, q, and r vary, they are called ***propositional variables***. A proposition has only two possible values: it is either true or false. We often abbreviate these values as T and F, respectively.

Given a proposition p, we form another proposition by changing its truth value. The result is called the ***negation*** of p, and is denoted $\neg p$ or $\sim p$, both of which are pronounced as "not p." The similarity between the notations $\neg p$ and $-x$ is obvious.

We can also write the negation of p as $\overline{p}$, which is pronounced as "p bar." The truth value of $\overline{p}$ is opposite of that of p. Hence, if p is true, then $\overline{p}$ would be false; and if p is false, then $\overline{p}$ would be true. We summarize these results in a ***truth table***:

p	$\overline{p}$
T	F
F	T

Example 2.1.7 Find the negation of the following statements:

(a) George W. Bush is the president of the United States.
(b) It is not true that New York is the largest state in the United States.
(c) x is a real number such that $x = 4$.
(d) x is a real number such that $x < 4$.

If necessary, you may rephrase the negated statements, and change a mathematical notation to a more appropriate one.

Solution: (a) George W. Bush is not the president of the United States.

(b) It is true that New York is the largest state in the United States.

(c) The phrase "x is a real number" describes what kinds of numbers we are considering. The main part of the proposition is the proclamation that $x = 4$. Hence, we only need to negate "$x = 4$". The answer is:

x is a real number such that $x \neq 4$.

(d) x is a real number such that $x \geq 4$. ▲

Hands-On Exercise 2.1.3 Negate the following statements:

(a) x is an integer greater than 7.

(b) We can factor 144 into a product of prime numbers.

(c) The number 64 is a perfect square.

△

Since we will be studying numbers throughout this course, it is convenient to introduce some notations to facilitate our discussion. Let

$$\begin{aligned}
\mathbb{N} &= \text{the set of natural numbers (positive integers)},\\
\mathbb{Z} &= \text{the set of integers},\\
\mathbb{R} &= \text{the set of real numbers, and}\\
\mathbb{Q} &= \text{the set of rational numbers.}
\end{aligned}$$

Recall that a rational number is a number that can be expressed as a ratio of two integers. Hence, a rational number can be written as $\frac{m}{n}$ for some integers m and n, where $n \neq 0$. If you use a word processor, and cannot find, for example, the symbol $\mathbb{N}$, you may use bold face $\mathbf{N}$ as a replacement.

We usually use uppercase letters such as A, B, C, S and T to represent sets, and denote their elements by the corresponding lowercase letters a, b, c, s, and t, respectively. To indicate that b is an element of the set B, we adopt the notation

$$b \in B \qquad \text{[pronounced as “}b\text{ belongs to }B\text{”]}.$$

Occasionally, we also use the notation

$$B \ni b \qquad \text{[pronounced as “}B\text{ contains }b\text{”]}.$$

Consequently, saying $x \in \mathbb{R}$ is another way of saying x is a real number.

Denote the set of positive real numbers, the set of negative real numbers, and the set of nonzero real numbers, by inserting the appropriate sign in the superscript:

$$\begin{aligned}
\mathbb{R}^+ &= \text{the set of all positive real numbers},\\
\mathbb{R}^- &= \text{the set of all negative real numbers},\\
\mathbb{R}^* &= \text{the set of all nonzero real numbers.}
\end{aligned}$$

The same convention applies to $\mathbb{Z}$ and $\mathbb{Q}$. Notice that $\mathbb{Z}^+$ is same as $\mathbb{N}$.

The meaning of kS. In addition, if S is a set of numbers, and k is a number, we sometimes use the notation kS to indicate the set of numbers obtained by multiplying k to every number in S.

Example 2.1.8 The notation $2\mathbb{Z}$ denotes the set of all even integers. Take note that an even integer can be positive, negative, or even zero. ▲

Summary and Review

- A proposition (statement or assertion) is a sentence which is either always true or always false.
- The negation of the statement p is denoted $\neg p$, $\sim p$, or $\overline{p}$.
- We can describe the effect of a logical operation by displaying a truth table which covers all possibilities (in terms of truth values) involved in the operation.

- The notations $\mathbb{R}$, $\mathbb{Q}$, $\mathbb{Z}$, and $\mathbb{N}$ represent the set of real numbers, rational numbers, integers, and natural numbers (positive integers), respectively.
- If S denotes a set of numbers, S^+ means the set of positive numbers in S, S^- means the set of negative numbers in S, and S^* means the set of nonzero numbers in S.
- If S denotes a set of numbers, and k is a real number, then kS means the set of numbers obtained by multiplying k to every number in S.

Exercises 2.1

1. Indicate which of the following are propositions (assume that x and y are real numbers).
   (a) The integer 36 is even.
   (b) Is the integer $3^{15} - 8$ even?
   (c) The product of 3 and 4 is 11.
   (d) The sum of x and y is 12.
   (e) If $x > 2$, then $x^2 \geq 3$.
   (f) $5^2 - 5 + 3$.
2. Which of the following are propositions (assume that x is a real number)?
   (a) $2\pi + 5\pi = 7\pi$.
   (b) The product of x^2 and x^3 is x^6.
   (c) It is not possible for $3^{15} - 7$ to be both even and odd.
   (d) If the integer x is odd, is x^2 odd?
   (e) The integer $2^{524287} - 1$ is prime.
   (f) $1.7 + .2 = 4.0$.
3. Determine the truth values of these statements:
   (a) The product of x^2 and x^3 is x^6 for any real number x.
   (b) $x^2 > 0$ for any real number x.
   (c) The number $3^{15} - 8$ is even.
   (d) The sum of two odd integers is even.
4. Determine the truth values of these statements:
   (a) $\pi \in \mathbb{Z}$.
   (b) $1^3 + 2^3 + 3^3 = 3^2 \cdot 4^2/4$.
   (c) u is a vowel.
   (d) This statement is both true and false.
5. Negate the statements in Problem 4.
6. Determine the truth values of these statements:
   (a) $\sqrt{2} \in \mathbb{Z}$ (b) $-1 \notin \mathbb{Z}^+$ (c) $0 \in \mathbb{N}$
   (d) $\pi \in \mathbb{R}$ (e) $\frac{4}{2} \in \mathbb{Q}$ (f) $1.5 \in \mathbb{Q}$
7. Determine whether these statements are true or false:
   (a) $0 \in \mathbb{Q}$ (b) $0 \in \mathbb{Z}$ (c) $-4 \in \mathbb{Z}$
   (d) $-4 \in \mathbb{N}$ (e) $2 \in 3\mathbb{Z}$ (f) $-18 \in 3\mathbb{Z}$
8. Negate the following statements about the real number x:
   (a) $x > 0$ (b) $x \leq -5$ (c) $7 \leq x$
9. Explain why $7\mathbb{Q} = \mathbb{Q}$. Is it still true that $0\mathbb{Q} = \mathbb{Q}$?
10. Find the number(s) k such that $k\mathbb{Z} = \mathbb{Z}$.

2.2 Conjunctions and Disjunctions

Given two real numbers x and y, we can form a new number by means of addition, subtraction, multiplication, or division, denoted $x+y$, $x-y$, $x \cdot y$, and x/y, respectively. The symbols $+$, $-$, $\cdot$, and $/$ are ***binary operators*** because they all work on two ***operands***. In fact, the negative sign in $-x$ can be regarded as a ***unary operator*** that changes the sign of x.

In a similar manner, from one or more logical statements, we can form a ***compound statement*** by joining them with ***logical operators***, which are also called ***logical connectives*** because they are used to connect logical statements. Obviously, negation is a unary operation.

Since a compound statement is itself a statement, it is either true or false. Therefore, we define a logical operation by describing the truth value of the resulting compound statement. The first two binary operations we shall study are conjunction and disjunction. They perform the "and" and "or" operations, respectively.

name	meaning	notation	truth value
conjunction	p and q	$p \wedge q$	true if both p and q are true, false otherwise
disjunction	p or q	$p \vee q$	false if both p and q are false, true otherwise

Their truth values are summarized in the following truth table:

p	q	$p \wedge q$	$p \vee q$
T	T	T	T
T	F	F	T
F	T	F	T
F	F	F	F

Example 2.2.1 Do not use mathematical notations as abbreviation in writing. For example, do *not* write "$x \wedge y$ are real numbers" if you want to say "x and y are real numbers."

In fact, the phrase "$x \wedge y$ are real numbers" is syntactically incorrect. Since $\wedge$ is a binary *logical* operator, it is used to connect two logical statements. Here, the "x" before $\wedge$ is not a logical statement. Therefore we cannot write "$x \wedge y$ are real numbers."

Incidentally, the statement "x and y are real numbers" is actually a conjunction. It means "x is a real number and y is a real number," or symbolically,

$$(x \in \mathbb{R}) \wedge (y \in \mathbb{R}).$$

It is wrong to write "$x \wedge y \in \mathbb{R}$." Can you explain why? ▲

Hands-On Exercise 2.2.1 Write "x and y are rational" as a conjunction, first in words, then in mathematical symbols.

△

Example 2.2.2 The statement "New York is the largest state in the United States and New York City is the state capital of New York" is clearly a conjunction. A conjunction of two statements is true only when both statements are true. Since New York is not the largest state in the United States, the conjunction is false.

In general, in a conjunction of two statements, if the first statement is false, no further consideration of the second statement is necessary since we know the conjunction must be false. In computer science, this is referred to as the ***short circuit evaluation***. ▲

Example 2.2.3 The statement "$\sqrt{30}$ is greater than 6 or $\sqrt{30}$ is less than 5" can be expressed symbolically as

$$\left(\sqrt{30} > 6\right) \vee \left(\sqrt{30} < 5\right).$$

Both statements "$\sqrt{30} > 6$" and "$\sqrt{30} < 5$" are false. Hence, their disjunction is also false. ▲

Example 2.2.4 Determine the truth values of the following statements:

(a) $\left(\sqrt{30} > 5\right) \wedge \left(\sqrt{30} > 7\right)$

(b) Either $\left(\sqrt{30} < 5\right)$ or $\left(\sqrt{30} > 7\right)$

Solution: (a) Since $\sqrt{30} > 5$ is true, but $\sqrt{30} > 7$ is false, their conjunction is false.

(b) Since $\sqrt{30} < 5$ is false, and $\sqrt{30} > 7$ is also false, their disjunction is false. ▲

Hands-On Exercise 2.2.2 Determine the truth values of the following statements:

(a) $\left(\sqrt{30} < 5\right)$ and $\left(\sqrt{30} > 7\right)$.

(b) $\left(\sqrt{30} > 5\right) \vee \left(\sqrt{30} < 7\right)$.

Be sure to show your reasons. △

Example 2.2.5 What does "$0 \leq x \leq 1$" really mean, logically?

Solution: It means the conjunction "$(0 \leq x) \wedge (x \leq 1)$." Hence, given a real number x, to test whether $0 \leq x \leq 1$, we have to check whether $0 \leq x$ *and* $x \leq 1$. ▲

"$0 \leq x \leq 1$" *means* "$0 \leq x$ *and* $x \leq 1$."

Hands-On Exercise 2.2.3 Write $5 < x < 8$ as a conjunction.

△

Hands-On Exercise 2.2.4 Many students assume that they can negate "$0 \leq x \leq 1$" by reversing the signs. However, neither "$0 \geq x \geq 1$" nor "$0 > x > 1$" is the correct negation. For example, what does "$0 \geq x \geq 1$" really mean? Actually, the statement "$0 \geq x \geq 1$" is syntactically correct, and it is always false. Can you explain why?

The negation of "$0 \leq x \leq 1$" *is* **not** "$0 \geq x \geq 1$."

△

In the everyday usage of most languages, when we say "p or q," we normally mean ***exclusive or***, which means either p or q is true, but not both. An example is "I either pass or fail this course," which really means

Either I pass this course or I fail this course.

Sometimes, as illustrated in the statement

Either you pass this course, or I pass this course.

the connective "or" can be interpreted as an ***inclusive or***. The actual meaning of "or" in human languages depends on the context. In mathematics, however, "or" *always* means inclusive or.

Summary and Review

- The conjunction "p and q" is denoted "$p \wedge q$". It is true only when both p and q are true.
- The disjunction "p or q" is denoted "$p \vee q$". It is false only when both p and q are false.
- The inequality "$a < x < b$" is actually a conjunction, it means "$(a < x) \wedge (x < b)$".
- Likewise, the phrase "x and y are rational" is also a conjunction, it means "x is rational and y is rational." Symbolically, we can write "$x \in \mathbb{Q} \wedge y \in \mathbb{Q}$."

Exercises 2.2

1. Let p, q, and r represent the following statements:

   p: Sam had pizza last night.
   q: Chris finished her homework.
   r: Pat watched the news this morning.

   Give a formula (using appropriate symbols) for each of these statements:

   (a) Sam had pizza last night and Chris finished her homework.
   (b) Chris did not finish her homework and Pat watched the news this morning.
   (c) Sam did not have pizza last night or Chris did not finish her homework.
   (d) Either Chris finished her homework or Pat watched the news this morning, but not both.

2. Define the propositional variables p, q, and r as in Problem 1. Express, in words, the statements represented by the following formulas:

   (a) $p \vee q$
   (b) $q \wedge r$
   (c) $(p \wedge q) \vee r$
   (d) $\overline{p} \vee r$

3. Consider the following statements:

   p: Niagara Falls is in New York.
   q: New York City is the state capital of New York.
   r: New York City will have more than 40 inches of snow in 2525.

   The statement p is true, but the statement q is false. Represent each of the following statements by a formula. What are their truth values if r is true? What if r is false?

   (a) Niagara Falls is in New York and New York City is the state capital of New York.
   (b) Niagara Falls is in New York or New York City is the state capital of New York.
   (c) Either Niagara Falls is in New York and New York City is the state capital of New York, or New York City will have more than 40 inches of snow in 2525.
   (d) New York City is not the state capital of New York and New York City will have more than 40 inches of snow in 2525.

4. Determine the truth values of these statements:

   (a) $(0 \in \mathbb{Q}) \wedge (-4 \in \mathbb{Z})$ (b) $(-4 \in \mathbb{N}) \vee (3 \in 2\mathbb{Z})$

5. Determine the truth values of these statements:

   (a) $(-3 > -2) \wedge (\sqrt{3} > 2)$ (b) $(4^2 - 5^2 \leq 0) \vee (\sqrt{3^2 + 4^2} = 3 + 4)$

6. Construct the truth tables for the following formulas:

   (a) $p \wedge \overline{q}$ (b) $\overline{p} \vee q$ (c) $\overline{p \wedge q}$

7. Rewrite the following expressions as conjunction:

   (a) $4 \leq x \leq 7$ (b) $4 < x \leq 7$ (c) $4 \leq x < 7$

8. In words, the inequality $0 < x < 1$ means "x is between 0 and 1." Its negation means x is outside this range. Hence, the negation is "$x \leq 0$ or $x \geq 1$." Find the negation of the following inequalities:

   (a) $0 \leq x \leq 4$ (b) $-2 < x \leq 5$ (c) $1.76 \leq x < \sqrt{5}$

9. In volleyball it is important to know which team is serving, because a team scores a point only if that team is serving and wins a volley. If the serving team loses the volley, then the other team gets to serve. Thus, to keep score in a volleyball game between teams A and B, it may be useful to define propositional variables p and q, where p is true if team A is serving (hence false if team B is serving); and q is true if team A wins the current volley (hence false if team B wins it).

   (a) Give a formula that is true if team A scores a point and is false otherwise.
   (b) Give a formula that is true if team B scores a point and is false otherwise.
   (c) Give a formula that is true if the serving team loses the current volley and is false otherwise.
   (d) Give a formula whose truth value determines whether the serving team will serve again.

10. The exclusive or operation, denoted $p \veebar q$, means "p or q, but not both."

    (a) Express $p \veebar q$ as a logic statement.
    (b) Construct the truth table for $p \veebar q$.

2.3 Implications

Most theorems in mathematics appear in the form of compound statements called conditional and biconditional statements. We shall study biconditional statement in the next section. Conditional statements are also called implications.

Definition. An ***implication*** is the compound statement of the form "if p, then q." It is denoted $p \Rightarrow q$, which is read as "p implies q." It is false only when p is true and q is false, and is true in all other situations.

p	q	$p \Rightarrow q$
T	T	T
T	F	F
F	T	T
F	F	T

The statement p in an implication $p \Rightarrow q$ is called its ***hypothesis***, ***premise***, or ***antecedent***, and q the ***conclusion*** or ***consequence***. $\diamondsuit$

Implications come in many disguised forms. There are several alternatives for saying $p \Rightarrow q$. The most common ones are

- p implies q,
- p only if q,
- q if p,
- q, provided that p.

All of them mean $p \Rightarrow q$.

Implications play a key role in logical argument. If an implication is known to be true, then whenever the hypothesis is met, the consequence must be true as well. This is why an implication is also called a ***conditional statement***.

Example 2.3.1 The quadratic formula asserts that

$$b^2 - 4ac > 0 \quad \Rightarrow \quad ax^2 + bx + c = 0 \text{ has two distinct real solutions.}$$

Consequently, the equation $x^2 - 3x + 1 = 0$ has two distinct real solutions because its coefficients satisfy the inequality $b^2 - 4ac > 0$. ▲

Hands-On Exercise 2.3.1 More generally,

- If $b^2 - 4ac > 0$, then the equation $ax^2 + bx + c = 0$ has two distinct real solutions. In fact, $ax^2 + bx + c = a(x - r_1)(x - r_2)$, where $r_1 \neq r_2$ are the two distinct roots.
- If $b^2 - 4ac = 0$, then the equation $ax^2 + bx + c = 0$ has only one real solution r. In such an event, $ax^2 + bx + c = a(x - r)^2$. Consequently, we call r a repeated root.
- If $b^2 - 4ac = 0$, then the equation $ax^2 + bx + c = 0$ has no real solution.

Use these results to determine how many solutions these equations have:

(a) $4x^2 + 12x + 9 = 0$

(b) $2x^2 - 3x - 4 = 0$

(c) $x^2 + x = -1$

△

Example 2.3.2 We have remarked earlier that many theorems in mathematics are in the form of implications. Here is an example:

$$\text{If } |r| < 1, \text{ then } 1 + r + r^2 + r^3 + \cdots = \tfrac{1}{1-r}.$$

It means, symbolically, $|r| < 1 \Rightarrow 1 + r + r^2 + r^3 + \cdots = \frac{1}{1-r}$. ▲

Hands-On Exercise 2.3.2 Express the following statement in symbol:

$$\text{If } x > y > 0, \text{ then } x^2 > y^2.$$

△

Example 2.3.3 If a father promises his kids, "If tomorrow is sunny, we will go to the beach," the kids will take it as a true statement. Consequently, if they wake up the next morning and find it sunny outside, they expect they will go to the beach. The father breaks his promise (hence making the implication false) only when it is sunny but he does not take his kids to the beach.

If it is cloudy outside the next morning, they do not know whether they will go to the beach, because no conclusion can be drawn from the implication (their father's promise) if the weather is bad. Nonetheless, they may still go to the beach, even if it rains! Since their father does not contradict his promise, the implication is still true. ▲

Many students are bothered by the validity of an implication even when the hypothesis is false. It may help if we understand how we use an implication. Assume we want to show that a certain statement q is true.

(i) First, we find a result of the form $p \Rightarrow q$. If we cannot find one, we have to prove that $p \Rightarrow q$ is true.
(ii) Next, show that the hypothesis p is fulfilled.
(iii) These two steps together allow us to draw the conclusion that q must be true.

Consequently, if p is false, we are not expected to use the implication $p \Rightarrow q$ at all. Since we are not are going to use it, we can define its truth value to anything we like. Nonetheless, we have to maintain consistency with other logical connectives. We will give a justification of our choice at the end of the next section.

Example 2.3.4 To show that "if $x = 2$, then $x^2 = 4$" is true, we need not worry about those x-values that are not equal to 2, because the implication is immediately true if $x \neq 2$. It suffices to assume that $x = 2$, and try to prove that we *will* get $x^2 = 4$. Since we do have $x^2 = 4$ when $x = 2$, the validity of the implication is established.

In contrast, to determine whether the implication "if $x^2 = 4$, then $x = 2$" is true, we assume $x^2 = 4$, and try to determine whether x *must* be 2. Since $x = -2$ makes $x^2 = 4$ true but $x = 2$ false, the implication is false.

We can use a counterexample to disprove *a claim.*

In general, to disprove an implication, it suffices to find a counterexample that makes the hypothesis true and the conclusion false. ▲

Hands-On Exercise 2.3.3 Determine whether these two statements are true or false:

(a) If $(x-2)(x-3) = 0$, then $x = 2$.

(b) If $x = 2$, then $(x-2)(x-3) = 0$.

Explain.

△

Example 2.3.5 Although we said examples can be used to disprove a claim, examples alone can *never* be used as proofs. If you are asked to show that

Examples cannot be used as proofs.

$$\text{if } x > 2, \text{ then } x^2 > 4,$$

you cannot prove it by *checking* just a few values of x, because you may find a counterexample after trying a few more calculations. Therefore, examples are only for illustrative purposes, they are *not* acceptable as proofs. ▲

Example 2.3.6 The statement

"If a triangle PQR is isosceles, then two of its angles have equal measure."

takes the form of an implication $p \Rightarrow q$, where

p : The triangle PQR is isosceles
q : Two of the angles of the triangle PQR have equal measure

In this example, we have to rephrase the statements p and q, because each of them should be a stand-alone statement. If we leave q as "two of its angles have equal measure," it is not clear what "its" is referring to. In addition, it is a good habit to spell out the details. It helps us focus our attention on what we are investigating. ▲

Example 2.3.7 The statement

"A square must also be a parallelogram."

can be expressed as an implication: "if the quadrilateral $PQRS$ is a square, then the quadrilateral $PQRS$ is a parallelogram."

Likewise, the statement

"All isosceles triangles have two equal angles."

can be rephrased as "if the triangle PQR is isosceles, then the triangle PQR has two equal angles." Since we have expressed the statement in the form of an implication, we no longer need to include the word "all." ▲

Hands-On Exercise 2.3.4 Rewrite each of these logical statements:

(a) Any square is also a parallelogram.

(b) A prime number is an integer.

(c) All polynomials are differentiable.

as an implication $p \Rightarrow q$. Specify what p and q are. △

Example 2.3.8 What does "p unless q" translate into, logically speaking? We know that p is true, provided that q does not happen. It means, in symbol, $\overline{q} \Rightarrow p$. Therefore,

The quadrilateral $PQRS$ is not a square
unless the quadrilateral $PQRS$ is a parallelogram

is the same as saying

If a quadrilateral $PQRS$ is not a parallelogram,
then the quadrilateral $PQRS$ is not a square.

Equivalently, "p unless q" means $\overline{p} \Rightarrow q$, because q is a necessary condition that prevents p from happening. ▲

Given an implication $p \Rightarrow q$, we define three related implications:

- Its ***converse*** is defined as $q \Rightarrow p$.
- Its ***inverse*** is defined as $\overline{p} \Rightarrow \overline{q}$.
- Its ***contrapositive*** is defined as $\overline{q} \Rightarrow \overline{p}$.

Among them, the contrapositive $\overline{q} \Rightarrow \overline{p}$ is the most important one. We shall study it again in the next section.

Example 2.3.9 The converse, inverse, and contrapositive of "$x > 2 \Rightarrow x^2 > 4$" are listed below.

$$\begin{array}{lrcl} \text{converse:} & x^2 > 4 & \Rightarrow & x > 2, \\ \text{inverse:} & x \leq 2 & \Rightarrow & x^2 \leq 4, \\ \text{contrapositive:} & x^2 \leq 4 & \Rightarrow & x \leq 2. \end{array}$$

We can change the notation when we negate a statement. If it is appropriate, we may even rephrase a sentence to make the negation more readable. ▲

Hands-On Exercise 2.3.5 List the converse, inverse, and contrapositive of the statement "if p is prime, then $\sqrt{p}$ is irrational."

△

The inverse of an implication is seldom used in mathematics, so we will only study the truth values of the converse and contrapositive.

p	q	$p \Rightarrow q$	$q \Rightarrow p$	$\overline{q}$	$\overline{p}$	$\overline{q} \Rightarrow \overline{p}$
T	T	T	T	F	F	T
T	F	F	T	T	F	F
F	T	T	F	F	T	T
F	F	T	T	T	T	T

An implication and its contrapositive always have the same truth value, but this is not true for the converse. What this means is, even though we know $p \Rightarrow q$ is true, there is no guarantee that $q \Rightarrow p$ is also true. This is an important observation, especially when we have a theorem stated in the form of an implication. So let us say it again:

$$\boxed{\text{The converse of a theorem in the form of an implication may not be true.}}$$

Accordingly, if you only know that $p \Rightarrow q$ is true, do not assume that its converse $q \Rightarrow p$ is also true. Likewise, if you are asked to prove that $p \Rightarrow q$ is true, do not attempt to prove $q \Rightarrow p$, because these two implications are not the same.

Example 2.3.10 We know that $p \Rightarrow q$ does not necessarily mean we also have $q \Rightarrow p$. This important observation explains the invalidity of the "proof" of $21 = 6$ in Example 1.4.3.

$$\boxed{\begin{array}{rcl} 21 & = & 6 \\ 6 & = & 21 \\ 27 & = & 27 \end{array}}$$

The argument we use here consists of three equations, but they are not individual unrelated equations. They are connected by implication.

$$\boxed{\begin{array}{rrcl} & 21 & = & 6 \\ \Rightarrow & 6 & = & 21 \\ \Rightarrow & 27 & = & 27 \end{array}}$$

Since implications are not reversible, even though we do have $27 = 27$, we cannot use this fact to prove that $21 = 6$. After all, an implication is true if its hypothesis is false. Therefore, having a true implication does not mean that its hypothesis must be true. In this example, the logic is sound, but it does not prove that $21 = 6$. ▲

There are two other ways to describe an implication $p \Rightarrow q$ in words. They are completely different from the ones we have seen thus far. They focus on whether we can tell one of the two components p and q is true or false if we know the truth value of the other.

- p is a ***sufficient condition*** for q
- q is a ***necessary condition*** for p.

They are difficult to remember, and can be easily confused. You may want to visualize it pictorially:

$$\boxed{\text{sufficient condition} \Rightarrow \text{necessary condition.}}$$

The idea is, assuming that $p \Rightarrow q$ is true, then

- For q to be true, it is enough to know or show that p is true. Hence, knowing p is true alone is sufficient for us to draw the conclusion the q must also be true.

- For p to be true, it is necessary to have q be true as well. Thus, knowing q is true does not necessarily mean that p must be true.

Example 2.3.11 Consider the implication

$$x = 1 \Rightarrow x^2 = 1.$$

If $x = 1$, we must have $x^2 = 1$. So, knowing $x = 1$ is enough for us to conclude that $x^2 = 1$. We say that $x = 1$ is a sufficient condition for $x^2 = 1$.

If $x = 1$, it is necessarily true that $x^2 = 1$, because, for example, it is impossible to have $x^2 = 2$. Nonetheless, knowing $x^2 = 1$ alone is not enough for us to decide whether $x = 1$, because x can be -1. Therefore, $x^2 = 1$ is *not* a sufficient condition for $x = 1$. Instead, $x^2 = 1$ is only a necessary condition for $x = 1$. ▲

Hands-On Exercise 2.3.6 Write these statements:

(a) For $x^2 > 1$, it is sufficient that $x > 1$.

(b) For $x^2 > 1$, it is necessary that $x > 1$.

in the form of $p \Rightarrow q$. Be sure to specify what p and q are.

△

Summary and Review

- An implication $p \Rightarrow q$ is false only when p is true and q is false.
- This is how we typically use an implication. Assume we want to show that q is true. We have to find or prove a theorem that says $p \Rightarrow q$. Next, we need to show that hypothesis p is met, hence it follows that q must be true.
- An implication can be described in several other ways. Can you name a few of them?
- Converse, inverse, and contrapositive are obtained from an implication by switching the hypothesis and the consequence, sometimes together with negation.
- In an implication $p \Rightarrow q$, the component p is called the sufficient condition, and the component q is called the necessary condition.

Exercises 2.3

1. Let p, q, and r represent the following statements:

   p: Sam had pizza last night.
   q: Chris finished her homework.
   r: Pat watched the news this morning.

   Give a formula (using appropriate symbols) for each of these statements:

   (a) If Sam had pizza last night then Chris finished her homework.
   (b) Pat watched the news this morning only if Sam had pizza last night.
   (c) Chris finished her homework if Sam did not have pizza last night.
   (d) It is not the case that if Sam had pizza last night, then Pat watched the news this morning.
   (e) Sam did not have pizza last night and Chris finished her homework implies that Pat watched the news this morning.

2. Define the propositional variables as in Problem 1. Express in words the statements represented by the following formulas.

   (a) $q \Rightarrow r$
   (b) $p \Rightarrow (q \wedge r)$
   (c) $\overline{p} \Rightarrow (q \vee r)$
   (d) $r \Rightarrow (p \vee q)$

3. Consider the following statements:

   p: Niagara Falls is in New York.
   q: New York City is the state capital of New York.
   r: New York City will have more than 40 inches of snow in 2525.

   The statement p is true, and the statement q is false. Represent each of the following statements by a formula. What is their truth value if r is true? What if r is false?

   (a) If Niagara Falls is in New York, then New York City is the state capital of New York.
   (b) Niagara Falls is in New York only if New York City will have more than 40 inches of snow in 2525.
   (c) Niagara Falls is in New York or New York City is the state capital of New York implies that New York City will have more than 40 inches of snow in 2525.
   (d) For New York City to be the state capital of New York, it is necessary that New York City will have more than 40 inches of snow in 2525.
   (e) For Niagara Falls to be in New York, it is sufficient that New York City will have more than 40 inches of snow in 2525.

4. Express each of the following compound statements symbolically:

   (a) If the triangle ABC is equilateral, then it is isosceles.
   (b) If $\sqrt{47089}$ is greater than 200 and $\sqrt{47089}$ is an integer, then $\sqrt{47089}$ is prime.
   (c) If $\sqrt{47089}$ is greater than 200, then, if $\sqrt{47089}$ is prime, it is greater than 210.
   (d) The line L_1 is perpendicular to the line L_2 and the line L_2 is parallel to the line L_3 implies that L_1 is perpendicular to L_3.
   (e) If $x^3 - 3x^2 + x - 3 = 0$, then either x is positive or x is negative or $x = 0$.

   Recall that $\mathbb{Z}$ means the set of all integers.

5. Express each of the following compound statements in symbols.

   (a) $x^3 - 3x^2 + x - 3 = 0$ only if $x = 3$.
   (b) A necessary condition for $x^3 - 3x^2 + x - 3 = 0$ is $x = 3$.

(c) A sufficient condition for $x^3 - 3x^2 + x - 3 = 0$ is $x = 3$.
(d) If e^π is a real number, then e^π is either rational or irrational.
(e) All NFL players are huge.

6. Find the converse, inverse, and contrapositive of the following implication:

   If the quadrilateral $ABCD$ is a rectangle, then $ABCD$ is a parallelogram.

7. Construct the truth tables for the following expressions:

   (a) $(p \wedge q) \vee r$ (b) $(p \vee q) \Rightarrow (p \wedge r)$

   Hint: To help you get started, fill in the blanks.

(a)

p	q	r	$p \wedge q$	$(p \wedge q) \vee r$
T	T	T		
T	T	F		
T	F	T		
T	F	F		
F	T	T		
F	T	F		
F	F	T		
F	F	F		

(b)

p	q	r	$p \vee q$	$p \wedge r$	$(p \vee q) \Rightarrow (p \wedge r)$
T	T	T			
T	T	F			
T	F	T			
T	F	F			
F	T	T			
F	T	F			
F	F	T			
F	F	F			

8. Construct the truth tables for the following expressions:

   (a) $(p \Rightarrow q) \vee (\overline{p} \Rightarrow q)$ (b) $(p \Rightarrow q) \wedge (\overline{p} \Rightarrow q)$

9. Determine (you may use a truth table) the truth value of p if

   (a) $(p \wedge q) \Rightarrow (q \vee r)$ is false (b) $(q \wedge r) \Rightarrow (p \wedge q)$ is false

10. Assume $p \Rightarrow q$ is true.

    (a) If p is true, must q be true? Explain.
    (b) If p is false, must q be true? Explain.
    (c) If q is true, must p be false? Explain.
    (d) If q if false, must p be false? Explain.

2.4 Biconditional Statements

The ***biconditional statement*** "p if and only if q," denoted $p \Leftrightarrow q$, is true when both p and q carry the same truth value, and is false otherwise. It is sometimes abbreviated as "p iff q." Its truth table is depicted below.

p	q	$p \Leftrightarrow q$
T	T	T
T	F	F
F	T	F
F	F	T

Example 2.4.1 The following biconditional statements

- $2x - 5 = 0 \Leftrightarrow x = 5/2$,
- $x > y \Leftrightarrow x - y > 0$,

are true, because, in both examples, the two statements joined by $\Leftrightarrow$ are true or false simultaneously. ▲

A biconditional statement can also be defined as the compound statement

$$(p \Rightarrow q) \wedge (q \Rightarrow p).$$

This explains why we call it a biconditional statement. A biconditional statement is often used to define a new concept.

Example 2.4.2 A number is even if and only if it is a multiple of 2. Mathematically, this means

$$n \text{ is even} \Leftrightarrow n = 2q \text{ for some integer } q.$$

It follows that for any integer m,

$$mn = m \cdot 2q = 2(mq).$$

Since mq is an integer (because it is a product of two integers), by definition, mn is even. This shows that the product of any integer with an even integer is always even. ▲

Hands-On Exercise 2.4.1 Complete the following statement:

$$n \text{ is odd} \Leftrightarrow \qquad\qquad\qquad .$$

Use this to prove that if n is odd, then n^2 is also odd.

△

Example 2.4.3 The operation "exclusive or" can be defined as

$$p \veebar q \Leftrightarrow (p \vee q) \wedge \overline{(p \wedge q)}.$$

See Problem 10 in Exercises 2.2. ▲

When we have a complex statement involving more than one logical operation, care must be taken to determine which operation should be carried out first. The ***precedence*** or ***priority*** is listed below.

Connectives	Priority
$\neg$	Highest
$\wedge$	
$\vee$	$\vdots$
$\Rightarrow$	
$\Leftrightarrow$	Lowest

This is the order in which the operations should be carried out if the logical expression is read from left to right. To override the precedence, use parentheses.

Example 2.4.4 The precedence of logical operations can be compared to those of arithmetic operations.

Operations	Priority
$-$ (Negative)	Highest
Exponentiation	$\vdots$
Multiplication/Division	$\vdots$
Addition/Subtraction	Lowest

For example, $yz^{-3} \neq (yz)^{-3}$. To evaluate yz^{-3}, we have to perform exponentiation first. Hence, $yz^{-3} = y \cdot z^{-3} = \frac{y}{z^3}$.

Another example: the notation x^{2^3} means x raised to the power of 2^3, hence $x^{2^3} = x^8$; it should *not* be interpreted as $(x^2)^3$, because $(x^2)^3 = x^6$. ▲

Example 2.4.5 It is not true that $p \Leftrightarrow q$ can be written as "$p \Rightarrow q \wedge q \Rightarrow p$," because it would mean, technically,

$$p \Rightarrow (q \wedge q) \Rightarrow p.$$

The correct notation is $(p \Rightarrow q) \wedge (q \Rightarrow p)$. ▲

Hands-On Exercise 2.4.2 Insert parentheses in the following formula

$$p \Rightarrow q \wedge r$$

to identify the proper procedure for evaluating its truth value. Construct its truth table.

△

Hands-On Exercise 2.4.3 Insert parentheses in the following formula

$$p \wedge q \Leftrightarrow \overline{p} \vee \overline{q}.$$

to identify the proper procedure for evaluating its truth value. Construct its truth table.

△

We close this section with a justification of our choice in the truth value of $p \Rightarrow q$ when p is false. The truth value of $p \Rightarrow q$ is obvious when p is true.

p	q	$p \Rightarrow q$
T	T	T
T	F	F
F	T	?
F	F	?

We want to decide what are the best choices for the two missing values so that they are consistent with the other logical connectives. Observe that if $p \Rightarrow q$ is true, and q is false, then p must be false as well, because if p were true, with q being false, then the implication $p \Rightarrow q$ would have been false. For instance, if we promise

"If tomorrow is sunny, we will go to the beach"

but we do not go to the beach tomorrow, then we know tomorrow must not be sunny. This means the two statements $p \Rightarrow q$ and $\overline{q} \Rightarrow \overline{p}$ should share the same truth value.

When both p and q are false, then both $\overline{p}$ and $\overline{q}$ are true. Hence $\overline{q} \Rightarrow \overline{p}$ should be true, consequently so is $p \Rightarrow q$. Thus far, we have the following partially completed truth table:

p	q	$p \Rightarrow q$
T	T	T
T	F	F
F	T	?
F	F	T

If the last missing entry is F, the resulting truth table would be identical to that of $p \Leftrightarrow q$. To distinguish $p \Leftrightarrow q$ from $p \Rightarrow q$, we have to define $p \Rightarrow q$ to be true in this case.

Summary and Review

- A biconditional statement $p \Leftrightarrow q$ is the combination of the two implications $p \Rightarrow q$ and $q \Rightarrow p$.
- The biconditional statement $p \Leftrightarrow q$ is true when both p and q have the same truth value, and is false otherwise.
- A biconditional statement is often used in defining a notation or a mathematical concept.

Exercises 2.4

1. Let p, q, and r represent the following statements:

   p: Sam had pizza last night.
   q: Chris finished her homework.
   r: Pat watched the news this morning.

   Give a formula (using appropriate symbols) for each of these statements.

   (a) Sam had pizza last night if and only if Chris finished her homework.
   (b) Pat watched the news this morning iff Sam did not have pizza last night.
   (c) Pat watched the news this morning if and only if Chris finished her homework and Sam did not have pizza last night.
   (d) In order for Pat to watch the news this morning, it is necessary and sufficient that Sam had pizza last night and Chris finished her homework.

2. Define the propositional variables as in Problem 1. Express in words the statements represented by the following formulas:

   (a) $q \Leftrightarrow r$
   (b) $p \Leftrightarrow (q \wedge r)$
   (c) $\overline{p} \Leftrightarrow (q \vee r)$
   (d) $r \Leftrightarrow (p \vee q)$

3. Consider the following statements:

   p: Niagara Falls is in New York.
   q: New York City is the state capital of New York.
   r: New York City will have more than 40 inches of snow in 2525.

   The statement p is true, and the statement q is false. Represent each of the following statements by a formula. What is their truth value if r is true? What if r is false?

(a) Niagara Falls is in New York if and only if New York City is the state capital of New York.
(b) Niagara Falls is in New York iff New York City will have more than 40 inches of snow in 2525.
(c) Niagara Falls is in New York or New York City is the state capital of New York if and only if New York City will have more than 40 inches of snow in 2525.

4. Express each of the following compound statements symbolically:

   (a) The product $xy = 0$ if and only if either $x = 0$ or $y = 0$.
   (b) The integer $n = 4$ if and only if $7n - 5 = 23$.
   (c) A necessary condition for $x = 2$ is $x^4 - x^2 - 12 = 0$.
   (d) A sufficient condition for $x = 2$ is $x^4 - x^2 - 12 = 0$.
   (e) For $x^4 - x^2 - 12 = 0$, it is both sufficient and necessary to have $x = 2$.
   (f) The sum of squares $x^2 + y^2 > 1$ iff both x and y are greater than 1.

5. Determine the truth values of the following statements (assuming that x and y are real numbers):

   (a) The product $xy = 0$ if and only if either $x = 0$ or $y = 0$.
   (b) The sum of squares $x^2 + y^2 > 1$ iff both x and y are greater than 1.
   (c) $x^2 - 4x + 3 - 0 \Leftrightarrow x = 3$.
   (d) $x^2 > y^2 \Leftrightarrow x > y$.

6. Determine the truth values of the following statements (assuming that x and y are real numbers):

   (a) u is a vowel if and only if b is a consonant.
   (b) $x^2 + y^2 = 0$ if and only if $x = 0$ and $y = 0$.
   (c) $x^2 - 4x + 4 = 0$ if and only if $x = 2$.
   (d) $xy \neq 0$ if and only if x and y are both positive.

7. We have seen that a number n is even if and only if $n = 2q$ for some integer q. Accordingly, what can you say about an odd number?

8. We also say that an integer n is even if it is divisible by 2, hence it can be written as $n = 2q$ for some integer q, where q represents the quotient when n is divided by 2. Thus, n is even if it is a multiple of 2. What if the integer n is a multiple of 3? What form must it take? What if n is not a multiple of 3?

2.5 Logical Equivalences

A ***tautology*** is a proposition that is always true, regardless of the truth values of the propositional variables it contains. A proposition that is always false is called a ***contradiction***. A proposition that is neither a tautology nor a contradiction is called a ***contingency***.

Example 2.5.1 From the following truth table

p	$\overline{p}$	$p \vee \overline{p}$	$p \wedge \overline{p}$
T	F	T	F
F	T	T	F

we gather that $p \vee \overline{p}$ is a tautology, and $p \wedge \overline{p}$ is a contradiction.

In words, $p \vee \overline{p}$ says that either the statement p is true, or the statement $\overline{p}$ is true (that is, p is false). This claim is always true.

The compound statement $p \wedge \overline{p}$ claims that p is true, and at the same time, $\overline{p}$ is also true (which means p is false). This is clearly impossible. Hence, $p \wedge \overline{p}$ must be false. ▲

Example 2.5.2 Show that $(p \Rightarrow q) \Leftrightarrow (\overline{q} \Rightarrow \overline{p})$ is a tautology.

Solution: We can use a truth table to verify the claim.

p	q	$p \Rightarrow q$	$\overline{q}$	$\overline{p}$	$\overline{q} \Rightarrow \overline{p}$	$(p \Rightarrow q) \Leftrightarrow (\overline{q} \Rightarrow \overline{p})$
T	T	T	T	F	F	T
F	T	T	F	T	T	T
F	F	T	T	T	T	T

Note how we work on each component of the compound statement separately before putting them together to obtain the final answer. ▲

Example 2.5.3 Show that the argument

"If p and q, then r. Therefore, if not r, then not p or not q."

is valid. In other words, show that the logic used in the argument is correct.

Solution: Symbolically, the argument says

$$[(p \wedge q) \Rightarrow r] \Rightarrow [\overline{r} \Rightarrow (\overline{p} \vee \overline{q})]. \tag{2.1}$$

We want to show that it is a tautology. It is easy to verify with a truth table. We can also argue that this compound statement is always true by showing that it can never be false.

Suppose, on the contrary, that (2.1) is false for some choices of p, q, and r. Then

$$(p \wedge q) \Rightarrow r \quad \text{must be true,} \qquad \text{and} \qquad \overline{r} \Rightarrow (\overline{p} \vee \overline{q}) \quad \text{must be false.}$$

For the second implication to be false, we need

$$\overline{r} \quad \text{to be true,} \qquad \text{and} \qquad \overline{p} \vee \overline{q} \quad \text{to be false.}$$

They in turn imply that r is false, and both $\overline{p}$ and $\overline{q}$ are false; hence both p and q are true. This would make $(p \wedge q) \Rightarrow r$ false, contradicting the assumption that it is true. Thus, (2.1) cannot be false, it must be a tautology. ▲

Hands-On Exercise 2.5.1 Use a truth table to show that

$$[(p \wedge q) \Rightarrow r] \Rightarrow [\overline{r} \Rightarrow (\overline{p} \vee \overline{q})]$$

is a tautology.

Solution: We need eight combinations of truth values in p, q, and r. We list the truth values according to the following convention. In the first column for the truth values of p, fill the upper half with T and the lower half with F. In the next column for the truth values of q, repeat the same pattern, separately, with the upper half and the lower half. So we split the upper half of the second column into two halves, fill the top half with T and the lower half with F. Likewise, split the lower half of the second column into two halves, fill the top half with T and the lower half with F. Repeat the same pattern with the third column for the truth values of r, and so on if we have more propositional variables.

Complete the following table:

p	q	r	$p \wedge q$	$(p \wedge q) \Rightarrow r$	$\overline{r}$	$\overline{p}$	$\overline{q}$	$\overline{p} \vee \overline{q}$	$\overline{r} \Rightarrow (\overline{p} \vee \overline{q})$	$[(p \wedge q) \Rightarrow r] \Rightarrow [\overline{r} \Rightarrow (\overline{p} \vee \overline{q})]$
T	T	T								
T	T	F								
T	F	T								
T	F	F								
F	T	T								
F	T	F								
F	F	T								
F	F	F								

Question: If there are four propositional variables in a proposition, how many rows are there in the truth table? △

Definition. Two logical formulas p and q are said to be ***logically equivalent***, denoted

$$p \equiv q,$$

if $p \Leftrightarrow q$ is a tautology. ♢

Do not write $p = q$; instead, write $p \equiv q$.

We are *not* saying that p is equal to q. Since p and q represent two different statements, they cannot be the same. What we are saying is, they always produce the same truth value, regardless of the truth values of the underlying propositional variables. That is why we write $p \equiv q$ instead of $p = q$.

Example 2.5.4 We have learned that

$$p \Leftrightarrow q \equiv (p \Rightarrow q) \wedge (q \Rightarrow p),$$

which is the reason why we call $p \Leftrightarrow q$ a biconditional statement. ▲

Example 2.5.5 Use truth tables to verify the following equivalent statements.

(a) $p \Rightarrow q \equiv \overline{p} \vee q$.
(b) $p \wedge (q \vee r) \equiv (p \wedge q) \vee (p \wedge r)$.

Solution: The truth tables for (a) and (b) are depicted below.

p	q	$p \Rightarrow q$	$\overline{p}$	$\overline{p} \vee q$
T	T	T	F	T
T	F	F	F	F
F	T	T	T	T
F	F	T	T	T

p	q	r	$q \vee r$	$p \wedge (q \vee r)$	$p \wedge q$	$q \wedge r$	$(p \wedge q) \vee (p \wedge r)$
T	T	T	T	T	T	T	T
T	T	F	T	T	T	F	T
T	F	T	T	T	F	T	T
T	F	F	F	F	F	F	F
F	T	T	T	F	F	F	F
F	T	F	T	F	F	F	F
F	F	T	T	F	F	F	F
F	F	F	T	F	F	F	F

Example (a) is an important result. It says that $p \Rightarrow q$ is true when one of these two things happen: (i) when p is false, (ii) otherwise (when p is true) q must be true. ▲

Hands-On Exercise 2.5.2 Use truth tables to establish these logical equivalences.

(a) $p \Rightarrow q \equiv \overline{q} \Rightarrow \overline{p}$
(b) $p \vee p \equiv p$
(c) $p \wedge q \equiv \overline{\overline{p} \vee \overline{q}}$
(d) $p \Leftrightarrow q \equiv (p \Rightarrow q) \wedge (q \Rightarrow p)$

Solution: We have set up the table for (a), and leave the rest to you.

p	q	$p \Rightarrow q$	$\overline{q}$	$\overline{p}$	$\overline{q} \Rightarrow \overline{p}$
T	T				
T	F				
F	T				
F	F				

△

Hands-On Exercise 2.5.3 The logical connective exclusive or, denoted $p \veebar q$, means either p or q but not both. Consequently,

$$p \veebar q \equiv (p \vee q) \wedge \overline{(p \wedge q)} \equiv (p \wedge \overline{q}) \vee (\overline{p} \wedge q).$$

Construct a truth table to verify this claim.

△

Properties of Logical Equivalence. Denote by T and F a tautology and a contradiction, respectively. We have the following properties for any propositional variables p, q, and r.

1. ***Commutative properties***: $p \vee q \equiv q \vee p$,
   $p \wedge q \equiv q \wedge p$.
2. ***Associative properties***: $(p \vee q) \vee r \equiv p \vee (q \vee r)$,
   $(p \wedge q) \wedge r \equiv p \wedge (q \wedge r)$.
3. ***Distributive laws***: $p \vee (q \wedge r) \equiv (p \vee q) \wedge (p \vee r)$,
   $p \wedge (q \vee r) \equiv (p \wedge q) \vee (p \wedge r)$.
4. ***Idempotent laws***: $p \vee p \equiv p$,
   $p \wedge p \equiv p$.
5. ***De Morgan's laws***: $\overline{p \vee q} \equiv \overline{p} \wedge \overline{q}$,
   $\overline{p \wedge q} \equiv \overline{p} \vee \overline{q}$.
6. ***Laws of the excluded middle***, or ***inverse laws***: $p \vee \overline{p} \equiv T$,
   $p \wedge \overline{p} \equiv F$.
7. ***Identity laws***: $p \vee F \equiv p$,
   $p \wedge T \equiv p$.
8. ***Domination laws***: $p \vee T \equiv T$,
   $p \wedge F \equiv F$.

Two VERY important logical equivalences. Memorize them!

9. Equivalence of an implication and its contrapositive: $p \Rightarrow q \equiv \overline{q} \Rightarrow \overline{p}$.
10. Writing an implication as a disjunction: $p \Rightarrow q \equiv \overline{p} \vee q$.

Be sure you memorize the last two equivalences, because we will use them frequently in the rest of the course.

Remark. These properties are not easy to recall. Instead of focusing on the symbolic formulas, try to understand their meanings. Let us explain them in words, and compare them to similar operations on the real numbers,

1. ***Commutative properties***: In short, they say that "the order of operation does not matter." It does not matter which of the two logical statements comes first, the result from conjunction and disjunction always produces the same truth value. Compare this to addition of real numbers: $x+y = y+x$. Subtraction is not commutative, because it is not always true that $x-y = y-x$. This explains why we have to make sure that an operation is commutative.

2. ***Associative properties***: Roughly speaking, these properties also say that "the order of operation does not matter." However, there is a key difference between them and the commutative properties.

   - Commutative properties apply to operations on *two* logical statements, but associative properties involves *three* logical statements. Since $\wedge$ and $\vee$ are *binary* operations, we can only work on a pair of statements at a time. Given the three statements p, q, and r, appearing in that order, which pair of statements should we operate on first? The answer is: it does not matter. It is the order of *grouping* (hence the term associative) that does not matter in associative properties.
   - The important consequence of the associative property is: since it does not matter on which pair of statements we should carry out the operation first, we can eliminate the parentheses and write, for example,

     $$p \vee q \vee r$$

     without worrying about any confusion.
   - Not all operations are associative. Subtraction is not associative. Given three numbers 5, 7, and 4, in that order, how should we carry out two subtractions? Which interpretation should we use:

     $$(5-7)-4, \qquad \text{or} \qquad 5-(7-4)?$$

     Since they lead to different results, we have to be careful where to place the parentheses.

3. ***Distributive laws***: When we mix two *different* operations on three logical statements, one of them has to work on a pair of statements first, forming an "inner" operation. This is followed by the "outer" operation to complete the compound statement. Distributive laws say that we can distribute the "outer" operation over the inner one.

4. ***Idempotent laws***: When an operation is applied to a pair of identical logical statements, the result is the same logical statement. Compare this to the equation $x^2 = x$, where x is a real number. It is true only when $x = 0$ or $x = 1$. But the logical equivalences $p \vee p \equiv p$ and $p \wedge p \equiv p$ are true for all p.

5. ***De Morgan's laws***: When we negate a disjunction (respectively, a conjunction), we have to negate the two logical statements, and change the operation from disjunction to conjunction (respectively, from conjunction to a disjunction).

6. ***Laws of the excluded middle***, or ***inverse laws***: Any statement is either true or false, hence $p \vee \overline{p}$ is always true. Likewise, a statement cannot be both true and false at the same time, hence $p \wedge \overline{p}$ is always false.

7. ***Identity laws***: Compare them to the equation $x \cdot 1 = x$: the value of x is unchanged after multiplying by 1. We call the number 1 the multiplicative identity. For logical operations, the identity for disjunction is F, and the identity for conjunction is T.

8. ***Domination laws***: Compare them to the equation $x \cdot 0 = 0$ for real numbers: the result is always 0, regardless of the value x. The "zero" for disjunction is T, and the "zero" for conjunction is F. ♢

Example 2.5.6 What is the negation of $2 \leq x \leq 3$? Give a logical explanation as well as a graphical explanation.

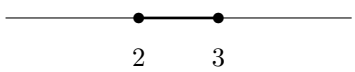
Solution: The inequality $2 \leq x \leq 3$ means

$$(2 \leq x) \wedge (x \leq 3).$$

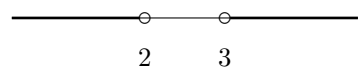
Its negation, according to De Morgan's laws, is

$$(2 > x) \vee (x > 3).$$

The inequality $2 \leq x \leq 3$ yields a closed interval. Its negation yields two open intervals. Their graphical representations on the real number line are depicted below.



$(2 \leq x) \wedge (x \leq 3)$



$(2 > x) \vee (x > 3)$

Take note of the two endpoints 2 and 3. They change from inclusion to exclusion when we take negation. ▲

Hands-On Exercise 2.5.4 Since $0 \leq x \leq 1$ means "$0 \leq x$ and $x \leq 1$," its negation should be "$0 > x$ or $x > 1$," which is often written as "$x < 0$ or $x > 1$." Explain why it is inappropriate, and indeed incorrect, to write "$0 > x > 1$." *See Example 2.2.5.*

△

Example 2.5.7 Expand $(p \wedge q) \vee (r \wedge s)$.

Solution: Compare this problem to the expansion of $(x + y)(u + v)$. We use the distributive law twice to obtain

$$\begin{aligned}(x+y)(u+v) &= x(u+v) + y(u+v) \\ &= xu + xv + yu + yv.\end{aligned}$$

Let us follow the same procedure to expand $(p \wedge q) \vee (r \wedge s)$. We need to apply the distributive law twice. The first time, regard $(r \wedge s)$ as a single statement, and distribute it over $p \wedge q$. In the second round, distribute p and q, separately, over $r \wedge s$. The complete solution is shown below.

$$\begin{aligned}(p \wedge q) \vee (r \wedge s) &\equiv [p \vee (r \wedge s)] \wedge [q \vee (r \wedge s)] \\ &\equiv (p \vee r) \wedge (p \vee s) \wedge (q \vee r) \wedge (q \vee s).\end{aligned}$$

We can also proceed as follows:

$$\begin{aligned}(p \wedge q) \vee (r \wedge s) &\equiv [(p \wedge q) \vee r] \wedge [(p \wedge q) \vee s] \\ &\equiv (p \vee r) \wedge (q \vee r) \wedge (p \vee s) \wedge (q \vee s).\end{aligned}$$

The two results are identical because $\wedge$ is commutative. ▲

Hands-On Exercise 2.5.5 Expand $(p \vee q) \wedge (r \vee s)$.

△

Example 2.5.8 We have used a truth table to verify that

$$[(p \wedge q) \Rightarrow r] \Rightarrow [\overline{r} \Rightarrow (\overline{p} \vee \overline{q})]$$

is a tautology. We can use the properties of logical equivalence to show that this compound statement is logically equivalent to T. This kind of proof is usually more difficult to follow, so it is a good idea to supply the explanation in each step. Here is a complete proof:

$$\begin{aligned}[(p \wedge q) \Rightarrow r] \Rightarrow [\overline{r} \Rightarrow (\overline{p} \vee \overline{q})] &\equiv \overline{(p \wedge q) \Rightarrow r} \vee [\overline{r} \Rightarrow (\overline{p} \vee \overline{q})] && \text{(implication as disjunction)} \\ &\equiv \overline{(p \wedge q) \Rightarrow r} \vee [\overline{\overline{p} \vee \overline{q}} \Rightarrow r] && \text{(implication as disjunction)} \\ &\equiv \overline{(p \wedge q) \Rightarrow r} \vee [(p \wedge q) \Rightarrow r] && \text{(De Morgan's law)} \\ &\equiv T && \text{(inverse law)}\end{aligned}$$

This is precisely what we called the left-to-right method for proving an identity (in this case, a logical equivalence). ▲

Example 2.5.9 Write $\overline{p \Rightarrow q}$ as a conjunction.

Wrong ways to negate an implication.

Solution: It is important to remember that

$$\overline{p \Rightarrow q} \not\equiv q \Rightarrow p,$$

and

$$\overline{p \Rightarrow q} \not\equiv \overline{p} \Rightarrow \overline{q}$$

either. Instead, since $p \Rightarrow q \equiv \overline{p} \vee q$, it follows from De Morgan's law that

$$\overline{p \Rightarrow q} \equiv \overline{\overline{p} \vee q} \equiv p \wedge \overline{q}.$$

How to negate an implication?

Alternatively, we can argue as follows. Interpret $\overline{p \Rightarrow q}$ as saying $p \Rightarrow q$ is false. This requires p to be true and q to be false, which translates into $p \wedge \overline{q}$. Thus, $\overline{p \Rightarrow q} \equiv p \wedge \overline{q}$. ▲

Summary and Review

- Two logical statements are logically equivalent if they always produce the same truth value.
- Consequently, $p \equiv q$ is same as saying $p \Leftrightarrow q$ is a tautology.
- Beside distributive and De Morgan's laws, remember these two equivalences as well; they are very helpful when dealing with implications.

$$p \Rightarrow q \equiv \overline{q} \Rightarrow \overline{p} \qquad \text{and} \qquad p \Rightarrow q \equiv \overline{p} \vee q.$$

Exercises 2.5

1. Use a truth table to verify the De Morgan's law $\overline{p \vee q} \equiv \overline{p} \wedge \overline{q}$.

2. Use truth tables to verify the two associative properties.

3. Construct a truth table for each formula below. Which ones are tautologies?

   (a) $(\overline{p} \vee q) \Rightarrow p$
   (b) $(p \Rightarrow q) \vee (p \Rightarrow \overline{q})$
   (c) $(p \Rightarrow q) \Rightarrow r$

4. Use truth tables to verify these logical equivalences.

   (a) $(p \wedge q) \Leftrightarrow p \equiv p \Rightarrow q$
   (b) $(p \wedge q) \Rightarrow r \equiv p \Rightarrow (\overline{q} \vee r)$
   (c) $(p \Rightarrow \overline{q}) \wedge (p \Rightarrow \overline{r}) \equiv p \wedge \overline{(q \vee r)}$

5. Use only the properties of logical equivalences to verify (b) and (c) in Problem 4.

6. Determine whether formulas u and v are logically equivalent (you may use truth tables or properties of logical equivalences).

   (a) $u:\ (p \Rightarrow q) \wedge (p \Rightarrow \overline{q})$ $\quad v:\ \overline{p}$
   (b) $u:\ p \Rightarrow q$ $\quad v:\ q \Rightarrow p$
   (c) $u:\ p \Leftrightarrow q$ $\quad v:\ q \Leftrightarrow p$
   (d) $u:\ (p \Rightarrow q) \Rightarrow r$ $\quad v:\ p \Rightarrow (q \Rightarrow r)$

7. Find the converse, inverse, and contrapositive of these implications.

   (a) If triangle ABC is isosceles and contains an angle of 45 degrees, then ABC is a right triangle.
   (b) If quadrilateral $ABCD$ is a square, then it is both a rectangle and a rhombus.
   (c) If quadrilateral $ABCD$ has two sides of equal length, then it is either a rectangle or a rhombus.

8. Negate the following implications:

   (a) $x^2 > 0 \Rightarrow x > 0$.
   (b) If $PQRS$ is a square, then $PQRS$ is a parallelogram.
   (c) If $n > 1$ is prime, then $n + 1$ is composite.
   (d) If x and y are integers such that $xy \geq 1$, then either $x \geq 1$ or $y \geq 1$.

9. Determine whether the following formulas are true or false:

   (a) $\overline{p \Leftrightarrow q} \equiv \overline{p} \Leftrightarrow \overline{q}$
   (b) $(p \Rightarrow q) \vee (p \Rightarrow \overline{q}) \equiv \overline{p}$
   (c) $p \Rightarrow q \equiv q \Rightarrow p$

10. Determine whether the following formulas are true or false:

    (a) $(p \Rightarrow q) \Rightarrow r \equiv p \Rightarrow (q \Rightarrow r)$
    (b) $p \Rightarrow (q \vee r) \equiv (p \Rightarrow q) \vee (p \Rightarrow r)$
    (c) $p \Rightarrow (q \wedge r) \equiv (p \Rightarrow q) \wedge (p \Rightarrow r)$

11. Which of the following statements are equivalent to the statement "if $x^2 > 0$, then $x > 0$"?

    (a) If $x > 0$, then $x^2 > 0$.
    (b) If $x \leq 0$, then $x^2 \leq 0$.
    (c) If $x^2 \leq 0$, then $x \leq 0$.
    (d) If $x^2 \not> 0$, then $x \not> 0$.

12. Determine whether the following formulas are tautologies, contradictions, or neither:

    (a) $(p \Rightarrow q) \wedge \overline{p}$
    (b) $(p \Rightarrow \overline{q}) \wedge (p \wedge q)$
    (c) $(p \Rightarrow \overline{q}) \wedge q$

13. Simplify the following formulas:

    (a) $p \wedge (p \wedge q)$ (b) $\overline{\overline{p} \vee q}$ (c) $\overline{p \Rightarrow \overline{q}}$

14. Simplify the following formulas:

    (a) $(p \Rightarrow \overline{q}) \wedge (\overline{q} \Rightarrow p)$ (b) $\overline{p \wedge \overline{q}}$ (c) $p \wedge (\overline{p} \vee q)$

2.6 Logical Quantifiers

The expression

$$x > 5$$

is neither true nor false. In fact, we cannot even determine its truth value unless we know the value of x. This is an example of a ***propositional function***, because it behaves like a function of x, it becomes a proposition when a specific value is assigned to x. Propositional functions are also called ***predicates***.

Example 2.6.1 Denote the propositional function "$x > 5$" by $p(x)$. We often write

$$p(x): \quad x > 5.$$

It is not a proposition because its truth value is undecidable, but $p(6)$, $p(3)$ and $p(-1)$ are propositions. ▲

Example 2.6.2 Define

$$q(x, y): \quad x + y = 1.$$

Which of the following are propositions; which are not?

(a) $q(x, y)$
(b) $q(x, 3)$
(c) $q(1, 1)$
(d) $q(5, -4)$

For those that are, determine their truth values.

Solution: Both (a) and (b) are not propositions, because they contain at least one variable. Both (c) and (d) are propositions; $q(1, 1)$ is false, and $q(5, -4)$ is true. ▲

Hands-On Exercise 2.6.1 Determine the truth values of these statements, where $q(x, y)$ is defined in Example 2.6.2.

(a) $q(5, -7)$

(b) $q(-6, 7)$

(c) $q(x + 1, -x)$

△

Although a propositional function is not a proposition, we can form a proposition by means of ***quantification***. The idea is to specify whether the propositional function is true for all or for some values that the underlying variables can take on.

Definition. The ***universal quantification*** of $p(x)$ is the proposition in any of the following forms:

- $p(x)$ is true for all values of x.
- For all x, $p(x)$.
- For each x, $p(x)$.
- For every x, $p(x)$.
- Given any x, $p(x)$.

All of them are symbolically denoted by

Write $\forall$, not A.

$$\forall x\, p(x),$$

which is pronounced as

"for all x, $p(x)$".

The symbol $\forall$ is called the ***universal quantifier***, and can be extended to several variables. $\diamondsuit$

Example 2.6.3 The statement

"For any real number x, we always have $x^2 \geq 0$"

is true. Symbolically, we can write

$$\forall x \in \mathbb{R}\,(x^2 \geq 0), \qquad \text{or} \qquad \forall x\,(x \in \mathbb{R} \Rightarrow x^2 \geq 0).$$

The second form is a bit wordy, but could be useful in some situations. ▲

Example 2.6.4 The statement

Counterexamples can be used to disprove a claim.

$$\forall x \in \mathbb{R}\,(x > 5)$$

is false because x is not always greater than 5. To disprove a claim, it suffices to provide only one counterexample. We can use $x = 4$ as a counterexample.

However, examples cannot be used to prove a universally quantified statement. Consider the statement

$$\forall x \in \mathbb{R}\,(x^2 \geq 0).$$

By direct calculations, one may demonstrate that $x^2 \geq 0$ is true for many x-values. But it does not *prove* that it is true for *every* x, because there may be a counterexample that we have not found yet. We have to use mathematical and logical argument to prove a statement of the form "$\forall x\, p(x)$." ▲

Examples alone do not prove a statement of the form "$\forall x\, p(x)$".

Example 2.6.5 The statement

"Every Discrete Mathematics student has taken Calculus I and Calculus II"

is clearly a universally quantified proposition. To express it in a logical formula, we can use an implication:

$$\forall x\,(x \text{ is a Discrete Mathematics student} \Rightarrow x \text{ has taken Calculus I and Calculus II})$$

An alternative is to say

See Example 2.6.3.

$$\forall x \in S\,(x \text{ has taken Calculus I and Calculus II})$$

where S represents the set of all Discrete Mathematics students. Although the second form looks simpler, we must define what S stands for. ▲

Definition. The ***existential quantification*** of $p(x)$ takes one of these forms:

- There exists an x such that $p(x)$.
- For some x, $p(x)$.
- There is some x such that $p(x)$.

Write $\exists$*, not* E.

We write, in symbol,

$$\exists x\, p(x),$$

which is pronounced as

"There exists x such that $p(x)$."

The symbol $\exists$ is called the ***existential quantifier***. It can be extended to several variables. $\diamondsuit$

One example suffices to prove a statement of the form "$\exists x\, p(x)$."

Example 2.6.6 To prove that a statement of the form "$\exists x\, p(x)$" is true, it suffices to find an example of x such that $p(x)$ is true. Using this guideline, can you determine whether these two propositions

(a) $\exists x \in \mathbb{R}\,(x > 5)$
(b) $\exists x \in \mathbb{R}\,(\sqrt{x} = 0)$

are true?

Solution: (a) True. For example: $x = 6$.

(b) True. For example: $x = 0$. ▲

Example 2.6.7 The proposition

"There exists a prime number x such that $x + 2$ is also prime"

is true. We call such a pair of primes ***twin primes***. ▲

Hands-On Exercise 2.6.2 Name a few more examples of twin primes.

△

Example 2.6.8 The proposition

"There exists a real number x such that $x > 5$"

can be expressed, symbolically, as

$$\exists x \in \mathbb{R}\,(x > 5), \qquad \text{or} \qquad \exists x\,(x \in \mathbb{R} \wedge x > 5).$$

Notice that in an existential quantification, we use $\wedge$ instead of $\Rightarrow$ to specify that x is a real number. ▲

Hands-On Exercise 2.6.3 Determine the truth value of each of the following propositions:

(a) For any prime number x, the number $x + 1$ is composite.

(b) For any prime number $x > 2$, the number $x + 1$ is composite.

(c) There exists an integer k such that $2k + 1$ is even.

(d) For all integers k, the integer $2k$ is even.

(e) For any real number x, if x^2 is an integer, then x is also an integer.

△

Hands-On Exercise 2.6.4 The proposition

"The square of any real number is positive"

is a universal quantification

"For any real number x, $x^2 > 0$."

Is it true or false?

△

Example 2.6.9 When multiple quantifiers are present, the order in which they appear is important. Determine whether these two statements are true or false.

(a) $\forall x \in \mathbb{Z}\, \exists y \in \mathbb{R}^*\, (xy < 1)$

(b) $\exists y \in \mathbb{R}^*\, \forall x \in \mathbb{Z}\, (xy < 1)$

Here, $\mathbb{R}^*$ denotes the set of all nonzero real numbers.

Solution: (a) To prove that the statement is true, we need to show that no matter what integer x we start with, we can always find a nonzero real number y such that $xy < 1$. For $x \leq 0$, we can pick $y = 1$, which makes $xy = x \leq 0 < 1$. For $x > 0$, let $y = \frac{1}{x+1}$, then $xy = \frac{x}{x+1} < 1$. This concludes the proof that the first statement is true.

(b) Let $y = 1$. Can we find an integer x such that $xy \not< 1$? Definitely! For example, we can set $x = 2$. This counterexample shows that the second statement is false. ▲

Hands-On Exercise 2.6.5 True or false: $\exists y \in \mathbb{R}\, \forall x \in \mathbb{Z}\, (xy < 1)$?

△

Example 2.6.10 Many theorems in mathematics can be expressed as quantified statements. Consider

"If x is rational and y is irrational, then $x + y$ is irrational."

This is same as saying

"Whenever x is rational and y is irrational, then $x + y$ is irrational."

The keyword "whenever" suggests that we should use a universal quantifier.

$$\forall x, y\,(x \text{ is rational} \wedge y \text{ is irrational} \Rightarrow x + y \text{ is irrational}).$$

It can also be written as

$$\forall x \in \mathbb{Q}\, \forall y \notin \mathbb{Q}\,(x + y \text{ is irrational}).$$

Although this form looks complicated and seems difficult to understand (primarily because it is quite symbolic, hence appears to be abstract and incomprehensible to many students), it provides an easy form for negation. See the discussion below.

The fact that an implication can be expressed as a universally quantified statement sounds familiar. See Example 2.3.7. ▲

We shall learn several basic proof techniques in Chapter 3. Some of them require negating a logical statement. Since many mathematical results are stated as quantified statements, it is necessary for us to learn how to negate a quantification. The rule is rather simple. Interchange $\forall$ and $\exists$, and negate the statement that is being quantified. In other words,

$$\overline{\forall x\, p(x)} \equiv \exists x\, \overline{p(x)}, \qquad \text{and} \qquad \overline{\exists x\, p(x)} \equiv \forall x\, \overline{p(x)}.$$

If we have $\forall x \in \mathbb{Z}$, we only change it to $\exists x \in \mathbb{Z}$ when we take negation. It should *not* be negated as $\exists x \notin \mathbb{Z}$. The reason is: we are only negating the quantification, not the membership of x. In symbols, we write

$$\overline{\forall x \in \mathbb{Z}\, p(x)} \equiv \exists x \in \mathbb{Z}\, \overline{p(x)}.$$

The negation of "$\exists x \in \mathbb{Z}\, p(x)$" is obtained in a similar manner.

Example 2.6.11 We find

$$\overline{\forall x \in \mathbb{Z}\, \exists y \in \mathbb{R}^*\,(xy < 1)} \equiv \exists x \in \mathbb{Z}\, \forall y \in \mathbb{R}^*\,(xy \geq 1),$$

and

$$\overline{\exists y \in \mathbb{R}^*\, \forall x \in \mathbb{Z}\,(xy < 1)} \equiv \forall y \in \mathbb{R}^*\, \exists x \in \mathbb{Z}\,(xy \geq 1).$$

Remember that we do not change the membership of x and y. ▲

Recall how to negate an implication.

Hands-On Exercise 2.6.6 Negate the propositions in Hands-On Exercise 2.6.3.

△

Example 2.6.12 The statement

"All real numbers x satisfy $x^2 \geq 0$"

can be written as, symbolically, $\forall x \in \mathbb{R}\,(x^2 \geq 0)$. Its negation is $\exists x \in \mathbb{R}\,(x^2 < 0)$. In words, it says "There exists a real number x that satisfies $x^2 < 0$." ▲

Hands-On Exercise 2.6.7 Negate the statement

"Every Discrete Mathematics student has taken Calculus I and Calculus II."

△

Summary and Review

- There are two ways to quantify a propositional function: universal quantification and existential quantification.
- They are written in the form of "$\forall x\, p(x)$" and "$\exists x\, p(x)$" respectively.
- To negate a quantified statement, change $\forall$ to $\exists$, and $\exists$ to $\forall$, and then negate the statement.

Exercises 2.6

1. Consider these propositional functions:

$$\begin{array}{ll} p(n): & n \text{ is prime} \\ q(n): & n \text{ is even} \\ r(n): & n > 2 \end{array}$$

   Express these formulas in words:

   (a) $\exists n \in \mathbb{Z}\, (p(n) \wedge q(n))$
   (b) $\forall n \in \mathbb{Z}\, [r(n) \Rightarrow p(n) \vee q(n)]$
   (c) $\exists n \in \mathbb{Z}\, [p(n) \wedge (q(n) \vee r(n)]$
   (d) $\forall n \in \mathbb{Z}\, [(p(n) \wedge r(n)) \Rightarrow \overline{q(n)}]$

2. Give a formula for each of the following statements:

   (a) For every even integer n there exists an integer k such that $n = 2k$.
   (b) There exists a right triangle T that is an isosceles triangle.
   (c) Given any quadrilateral Q, if Q is a parallelogram and Q has two adjacent sides that are perpendicular, then Q is a rectangle.

   Recall that the set of all even integers can be written as $2\mathbb{Z}$.

3. Determine whether these statements are true or false:

   (a) There exists an even prime integer.
   (b) There exist integers s and t such that $1 < s < t < 187$ and $st = 187$.
   (c) There is an integer m such that both $m/2$ is an integer and, for every integer k, $m/(2k)$ is not an integer.
   (d) Given any real numbers x and y, $x^2 - 2xy + y^2 > 0$.
   (e) For every integer n, there exists an integer m such that $m > n^2$.

4. Determine whether these statements are true or false:

   (a) There is a rational number x such that $x^2 \leq 0$.
   (b) There exists a number x such that for every real number y, $xy = 0$.
   (c) For all $x \in \mathbb{Z}$, either x is even, or x is odd.
   (d) There exists a unique number x such that $x^2 = 1$.

5. Find the negation (in simplest form) of each formula.

   (a) $\forall x < 0\, \forall y, z \in \mathbb{R}\, (y < z \Rightarrow xy > xz)$
   (b) $\forall x \in \mathbb{Z}\, [p(x) \vee q(x)]$
   (c) $\forall x, y \in \mathbb{R}\, [p(x, y) \Rightarrow q(x, y)]$

6. Negate the following statements:

   (a) For all real numbers x, there exists an integer y such that $p(x, y)$ implies $q(x, y)$.
   (b) There exists a rational number x such that for all integers y, either $p(x, y)$ or $r(x, y)$ is true.
   (c) For all integers x, there exists an integer y such that if $p(x, y)$ is true, then there exists an integer z so that $q(x, y, z)$ is true.

7. For each statement, (i) represent it as a formula, (ii) find the negation (in simplest form) of this formula, and (iii) express the negation in words.

   (a) For all real numbers x and y, $x + y = y + x$.
   (b) For every positive real number x there exists a real number y such that $y^2 = x$.
   (c) There exists a real number y such that, for every integer x, $2x^2 + 1 > x^2 y$.

8. For each statement, (i) represent it as a formula, (ii) find the negation (in simplest form) of this formula, and (iii) express the negation in words.

   (a) There exist rational numbers x_1 and x_2 such that $x_1 < x_2$ and $x_1^3 - x_1 > x_2^3 - x_2$.
   (b) For all real numbers x and y there exists an integer z such that $2z = x + y$.
   (c) For all real numbers x_1 and x_2, if $x_1^3 + x_1 - 2 = x_2^3 + x_2 - 2$, then $x_1 = x_2$.

9. The easiest way to negate the proposition

   "A square must be a parallelogram"

   is to say

   "It is not true that a square must be a parallelogram."

   Yet, it is not the same as saying

   "A square must not be a parallelogram."

   Can you explain why? What are other ways to express its negation in words?

10. Negate these statements:

    (a) All squared numbers are positive.
    (b) All basketball players are over 6 feet tall.
    (c) No quarterback is under 6 feet tall.

Chapter 2

SWDM 3: Proof Techniques

Chapter 3

Proof Techniques

3.1 An Introduction to Proof Techniques

A proof is a logical argument that verifies the validity of a statement. A good proof must be correct, but it also needs to be clear enough for others to understand. In the following sections, we want to show you how to write mathematical arguments. It takes practice to learn how to write mathematical proofs; you have to keep trying! We would like to start with some suggestions.

1. **Write at the level of your peers**. A common question asked by many students is: how much detail should I include in a proof? One simple guideline is to write at the level that your peers can understand. Although you can skip the detailed computation, be sure to include the major steps in an argument.

2. **Use symbols and notations appropriately**. Do not use mathematical symbols as abbreviations. For example, do not write "x is a number > 4." Use "x is a number greater than 4" instead. Do not use symbols excessively either. It is often clearer if we express our idea in words. Finally, do not start a sentence with a symbol, as in "Suppose $xy > 0$. x and y have the same signs." It would look better if we combine the two sentences, and write "Suppose $xy > 0$, then x and y have the same signs."

3. **Display long and important equations separately**. Make the key mathematical results stand out by displaying them separately on their own. Be sure to center these expressions. Number them if you need to refer to them later. See Examples 1.3.1 and 1.3.2.

4. **Write in complete sentences, with proper usage of grammar and punctuation**. A proof is, after all, a piece of writing. It should conform to the usual writing rules. Use complete sentences, and do not forget to check the grammar and punctuation.

5. **Start with a draft**. Prepare a draft. When you feel it is correct, start revising it: check the accuracy, remove redundancy, and simplify the sentence structure. Organize the argument into short paragraphs to enhance the readability of a proof. Go over the proof and refine it further.

Some proofs only require direct computation.

Example 3.1.1 Let a and b be two rational numbers such that $a < b$. Show that the weighted average $\frac{1}{3}a + \frac{2}{3}b$ is a rational number between a and b.

Solution: Since a and b are rational numbers, we can write $a = \frac{m}{n}$ and $b = \frac{p}{q}$ for some integers m, n, p, and q, where $n, q \neq 0$. Then

$$\frac{1}{3}a + \frac{2}{3}b = \frac{1}{3} \cdot \frac{m}{n} + \frac{2}{3} \cdot \frac{p}{q} = \frac{mq + 2np}{3nq}$$

is clearly a rational number because $mq + 2np$ and $3np$ are integers, and $3nq \neq 0$. Since $a < b$, we know $b - a > 0$. It follows that

$$\left(\frac{1}{3}\,a + \frac{2}{3}\,b\right) - a = \frac{2}{3}\,(b - a) > 0,$$

which means $\frac{1}{3}\,a + \frac{2}{3}\,b > a$. In a similar fashion, we also find $\frac{1}{3}\,a + \frac{2}{3}\,b < b$. Thus, $\frac{1}{3}\,a + \frac{2}{3}\,b$ is a rational number between a and b. ▲

Hands-On Exercise 3.1.1 Show that $\frac{1}{3}\,a + \frac{2}{3}\,b$ is closer to b than to a.

Hint: Compute the distance between a and $\frac{1}{3}\,a + \frac{2}{3}\,b$, and compare it to the distance between $\frac{1}{3}\,a + \frac{2}{3}\,b$ and b.

△

Sometimes, we can use a ***constructive proof*** when a proposition claims that certain values or quantities exist.

Example 3.1.2 Prove that every positive integer can be written in the form of $2^e t$ for some nonnegative integer e and some odd integer t.

Remark. The problem statement only says "every positive integer." It often helps if we assign a name to the integer; it will make it easier to go through the discussion. Consequently, we customarily start a proof with the phrase "Let n be ... " ♢

Solution: Let n be a positive integer. Keep dividing n by 2 until an odd number t remains. Let e be the number of times we factor out a copy of 2. It is clear that e is nonnegative, and we have found $n = 2^e t$. ▲

Hands-On Exercise 3.1.2 Express 6, 40, 32, and 15 in the form stated in Example 3.1.2.

△

Example 3.1.3 Given any positive integer n, show that there exist n consecutive composite positive integers.

Solution: For each positive integer n, we claim that the n integers

$$(n+1)! + 2, \quad (n+1)! + 3, \quad \ldots \quad (n+1)! + n, \quad (n+1)! + (n+1)$$

are composite. Here is the reason. For each i, where $2 \leq i \leq n+1$, the integer

$$\begin{aligned} (n+1)! + i &= 1 \cdot 2 \cdot 3 \cdots (i-1) i (i+1) \cdots (n+1) + i \\ &= i\,[\,1 \cdot 2 \cdot 3 \cdots (i-1)(i+1) \cdots (n+1) + 1\,] \end{aligned}$$

is divisible by i and greater than i, and hence is composite. ▲

Hands-On Exercise 3.1.3 Construct five consecutive positive integers that are composite. Verify their compositeness by means of factorization.

△

Example 3.1.4 Let m and n be positive integers. Show that, if mn is even, then an $m \times n$ chessboard can be fully covered by non-overlapping dominoes.

Remark. This time, the names m and n have already been assigned to the two positive integers. Thus, we can refer to them in the proof without an introduction.

Solution: Since mn is even, one of the two integers m and n must be even. Without loss of generality (since the other case is similar), we may assume m, the number of rows, is even. Then $m = 2t$ for some integer t. Each column can be filled with $m/2 = t$ non-overlapping dominoes placed vertically. As a result, the entire chessboard can be covered with nt non-overlapping vertical dominoes. ▲

Hands-On Exercise 3.1.4 Show that, between any two rational numbers a and b, where $a < b$, there exists another rational number.

Hint: Try the midpoint of the interval $[a, b]$.

△

Hands-On Exercise 3.1.5 Show that, between any two rational numbers a and b, where $a < b$, there exists another rational number closer to b than to a.

Hint: Use a weighted average of a and b.

△

Sometimes a non-constructive proof can be used to show the existence of a certain quantity that satisfies some conditions. We have learned two such existence theorems from calculus.

Theorem 3.1.1 (Mean Value Theorem) *Let f be a differentiable function defined over a closed interval $[a, b]$. Then there exists a number c strictly inside the open interval (a, b) such that $f'(c) = \frac{f(b)-f(a)}{b-a}$.*

Theorem 3.1.2 (Intermediate Value Theorem) *Let f be a function that is continuous over a closed interval $[a, b]$. Then f assumes all values between $f(a)$ and $f(b)$. In other words, for any value t between $f(a)$ and $f(b)$, there exists a number c inside $[a, b]$ such that $f(c) = t$.*

Both results only guarantee the existence of a number c with some specific property; they do not tell us how to find this number c. Nevertheless, the Mean-Value Theorem plays a very important role in analysis; many of its applications are beyond the scope of this course. We could, however, demonstrate an application of the Intermediate Value Theorem.

Corollary 3.1.3 *Let f be a continuous function defined over a closed interval $[a, b]$. If $f(a)$ and $f(b)$ have opposite signs, then the equation $f(x) = 0$ has a solution between a and b.*

Proof: According to the Intermediate Value Theorem, $f(x)$ can take on any value between $f(a)$ and $f(b)$. Since they have opposite signs, 0 is a number between them. Hence, $f(c) = 0$ for some number c between a and b. ■

Example 3.1.5 The function $f(x) = 5x^3 - 2x - 1$ is a polynomial function, which is known to be continuous over the real numbers. Since $f(0) = -1$ and $f(1) = 2$, Corollary 3.1.3 implies that there exists a number between 0 and 1 such that $5x^3 - 2x - 1 = 0$. ▲

Hands-On Exercise 3.1.6 Show that the equation $1+x\cos x = 0$ has at least one real solution between 0 and $\frac{\pi}{2}$.

Hint: No function is mentioned here, so you need to define a function, say $g(x)$. Next, you need to make sure that $g(x)$ is continuous. What else do you need to do before you can apply Corollary 3.1.3?

△

Summary and Review

- Sometimes we can prove a statement by showing how the result can be obtained through a construction, and we can describe the construction in an algorithm.
- Sometimes all we need to do is apply an existence theorem to verify the existence of a certain quantity.

Exercises 3.1

1. Show that a chessboard with 7 rows and 12 columns can be covered by non-overlapping dominoes.

2. Show that there is a rational number between 1 and 5 whose distance from 5 is seven times as long as its distance from 1.

3. Show that the equation $x^3 - 12x + 2 = 0$ has at least three real solutions.

4. Show that if the equation $(x^2+4)(x-2)(3x+5) = 0$ has a real solution, the solution must be either $x = 2$ or $x = -\frac{5}{3}$.

5. Show that given any rational number x, there exists an integer y such that x^2y is an integer.

   Hint: Since x is rational, we can write $x = \frac{m}{n}$ for some integers m and n, where $n \neq 0$. All you need to do is to describe y in terms of m and n.

6. Show that given any rational number x, and any positive integer k, there exists an integer y such that x^ky is an integer.

7. Show that there exists an integer n such that n, $n+2$ and $n+4$ are all primes.

8. Find a counterexample to the following claim: For any positive integer n, if n is prime, then $n^2 + 4$ is also prime.

3.2 Direct Proofs

To show that a statement q is true, follow these steps:

1. Either find a result that states $p \Rightarrow q$, or prove that $p \Rightarrow q$ is true.
2. Show or verify that p is true.
3. Conclude that q must be true.

The logic is valid because if $p \Rightarrow q$ is true and p is true, then q must be true. Symbolically, we are saying that the logical formula

$$[(p \Rightarrow q) \wedge p] \Rightarrow q$$

is a tautology (we can easily verify this with a truth table). Symbolically, we present the argument as

$$\begin{array}{ll} & p \Rightarrow q \\ & p \\ \hline \therefore & q \end{array}$$

Such an argument is called ***modus ponens*** or the ***law of detachment***.

Example 3.2.1 The argument

$$\begin{array}{ll} & b^2 > 4ac \Rightarrow ax^2 + bx + c = 0 \text{ has two real solutions.} \\ & x^2 - 5x + 6 \text{ satisfies } b^2 > 4ac. \\ \hline \therefore & x^2 - 5x + 6 = 0 \text{ has two real solutions.} \end{array}$$

is an example of modus ponens. ▲

It is clear that implications play an important role in mathematical proofs. If we have a sequence of implications, we could join them "head to tail" to form another implication:

$$\begin{array}{ll} & p \Rightarrow q \\ & q \Rightarrow r \\ \hline \therefore & p \Rightarrow r \end{array}$$

This is called the ***law of syllogism***.

Example 3.2.2 The argument

$$\begin{array}{ll} & \text{German shepherds are dogs.} \\ & \text{Dogs are mammals.} \\ & \text{Mammals are vertebrates.} \\ \hline \therefore & \text{German shepherds are vertebrates.} \end{array}$$

is valid because of the law of syllogism. ▲

The big question is, how can we prove an implication? The most basic approach is the ***direct proof***:

1. Assume p is true.
2. Deduce from p that q is true.

Proving $p \Rightarrow q$ via a direct proof.

The important thing to remember is: use the information derived from p to show that q is true. This is how a typical direct proof may look:

Proof: Assume p is true. Then ...

Because of p, we find ...

... Therefore q is true. ■

Example 3.2.3 Prove that if an $m \times n$ chessboard can be fully covered by non-overlapping dominoes, then mn must be even.

Solution: Assume the chessboard can be covered by non-overlapping dominoes, and let t be the number of dominoes that cover the chessboard. Then the chessboard must contain $2t$ squares. Hence $mn = 2t$, which means mn must be an even number. ▲

Before we continue with more examples, we would like to introduce the formal definition of even and odd integers.

Definition. An integer is ***even*** if it can be written as $2q$ for some integer q, and ***odd*** if it can be written as $2q + 1$ for some integer q. ♢

We do not have to use q to denote the integer that, when multiplied by 2, produces an even integer. Any letter will work, provided that we mention it is an integer. For example, if n is an even integer, then we can write $n = 2t$ for some integer t. The notion of even integers can be further generalized.

Definition. Let m be a nonzero integer. An integer is said to be a ***multiple*** of m if it can be written as mq for some integer q. ♢

We are now ready to study more examples.

Example 3.2.4 Show that the square of an odd integer is odd.

Solution: Let n be an odd integer. Then $n = 2t + 1$ for some integer t, and

$$n^2 = (2t + 1)^2 = 4t^2 + 4t + 1 = 2(2t^2 + 2t) + 1,$$

where $2t^2 + 2t$ is an integer. Hence, n^2 is odd. ▲

Hands-On Exercise 3.2.1 Let n be an integer. Show that if n is odd, then n^3 is odd.

△

Example 3.2.5 Show that the product of two odd integers is odd.

Describe your goal.

Solution: Let x and y be two odd integers. We want to prove that xy is odd. Then $x = 2s + 1$ and $y = 2t + 1$ for some integers s and t, and

$$xy = (2s + 1)(2t + 1) = 4st + 2s + 2t + 1 = 2(2st + s + t) + 1,$$

where $2st + s + t$ is an integer. Therefore, xy is odd. ▲

In this proof, we need to use two different quantities s and t to describe x and y because they need not be the same. If we write $x = 2s + 1$ and $y = 2s + 1$, we are in effect saying that $x = y$. We have to stress that s and t are integers, because just saying $x = 2s + 1$ and $y = 2t + 1$ does not guarantee x and y are odd. For instance, the even number 4 can be written as $2 \cdot \frac{3}{2} + 1$, which is of the form $2s + 1$. It is obvious that 4 is not odd. Even though we can write a number in the form $2s + 1$, it does not necessarily mean the number must be odd, *unless* we know with certainty that s is an integer. This example illustrates the importance of paying attention to the details in our writing.

Pay attention to the details!

Example 3.2.6 Show that if $x^3 - 7x^2 + x - 7 = 0$, then $x = 7$.

Solution: Assume $x^3 - 7x^2 + x - 7 = 0$. Since

$$x^3 - 7x^2 + x - 7 = x^2(x - 7) + (x - 7) = (x^2 + 1)(x - 7),$$

if it is equal to zero, we need either $x^2 + 1 = 0$, or $x - 7 = 0$. Since $x^2 + 1$ can never be zero, we must have $x - 7 = 0$; thus $x = 7$. ▲

Hands-On Exercise 3.2.2 Show that if $x^3 + 6x^2 + 12x + 8 = 0$, then $x = -2$.

△

The last example demonstrates a technique called ***proof by cases***. There are two possibilities, namely, either (i) $x^2 + 1 = 0$, or (ii) $x - 7 = 0$. The final conclusion is drawn after we study these two cases separately.

Example 3.2.7 Show that if an integer n is not divisible by 3, then $n^2 - 1$ must be a multiple of 3.

Remark. The letter n has been used to identify the integer of interest to us, and it appears in the hypothesis of the implication that we want to prove. Nonetheless, many authors would start their proofs with the familiar phrase "Let n be . . . ." ♢

Solution: Let n be an integer that is not divisible by 3. When it is divided by 3, the remainder is 1 or 2. Hence, $n = 3q + 1$ or $n = 3q + 2$ for some integer q.

- Case 1: If $n = 3q + 1$ for some integer q, then

  $$n^2 - 1 = 9q^2 + 6q = 3(3q^2 + 2q),$$

  where $3q^2 + 2q$ is an integer.

- Case 2: If $n = 3q + 2$ for some integer q, then

  $$n^2 - 1 = 9q^2 + 12q + 3 = 3(3q^2 + 4q + 1),$$

  where $3q^2 + 4q + 1$ is an integer.

In both cases, we have shown that $n^2 - 1$ is a multiple 3. ▲

Hands-On Exercise 3.2.3 Show that $n^3 + n$ is even for all $n \in \mathbb{N}$.

△

Hands-On Exercise 3.2.4 Show that $n(n + 1)(2n + 1)$ is divisible by 6 for all $n \in \mathbb{N}$.

Hint: One of the two integers n and $n+1$ must be even, so we already know that the product $n(n+1)(2n+1)$ is a multiple of 2. Hence, it remains to show that it is also a multiple of 3. Consider three cases: $n = 3q$, $n = 3q+1$, or $n = 3q+2$, where q is an integer.

△

We close our discussion with two common fallacies (logical errors). The first one is the ***fallacy of the inverse*** or the ***denial of the antecedent***:

$$\begin{array}{ll} & p \Rightarrow q \\ & \overline{p} \\ \hline \therefore & \overline{q} \end{array}$$

This in effect proves the inverse $\overline{p} \Rightarrow \overline{q}$, which we know is *not* logically equivalent to the original implication. Hence, this is an incorrect method for proving an implication.

Example 3.2.8 Is the following argument

	Dictionaries are valuable.
	This book is not a dictionary.
$\therefore$	This book is not valuable.

valid? Why? ▲

Another common mistake is known as the ***fallacy of the converse*** or the ***affirmation of the consequence***:

$$\begin{array}{ll} & p \Rightarrow q \\ & q \\ \hline \therefore & p \end{array}$$

This only proves the converse $q \Rightarrow p$. Since the converse is *not* logically equivalent to the original implication, this is an incorrect way to prove an implication.

Example 3.2.9 Is this argument

	No medicine tastes good.
	This drink tastes bad.
$\therefore$	This must be medicine.

a valid argument? Why? ▲

Summary and Review

- To prove an implication $p \Rightarrow q$, start by assuming that p is true. Use the information from this assumption, together with any other known results, to show that q must also be true.
- If necessary, you may break p into several cases $p_1, p_2, \ldots,$ and prove each implication $p_i \Rightarrow q$ (separately, one at a time) as indicated above.
- Be sure to write the mathematical expressions clearly. Use different variables if the quantities involved may not be the same.

- To get started, write down the given information, the assumption, and what you want to prove.
- In the next step, use the definition if necessary, and rewrite the information in mathematical notations. The point is, try to obtain some mathematical equations or logical statements that we can manipulate.

Exercises 3.2

1. Prove or disprove: $2^n + 1$ is prime for all nonnegative integer n.

2. Show that for any integer $n \geq 5$, the integers n, $n+2$ and $n+4$ cannot be all primes.

   Hint: If n is a multiple of 3, then n itself is composite, and the proof will be complete. So we may assume n is not divisible by 3. Then what would n look like, and, what can you say about $n+2$ and $n+4$?

3. Let n be an integer.

   (a) Show that if n is odd, then n^2 is also odd.
   (b) Show that if n is odd, then n^4 is also odd.
   (c) A ***corollary*** is a result that can be derived easily from another result. Derive (b) as a corollary of (a).
   (d) Show that if m and n are odd, then so is mn.
   (e) Show that if m is even, and n is odd, then mn is even.

4. Prove that, for any odd integer n, the number $2n^2 + 5n + 4$ must be odd.

5. Let n be an integer.

   (a) Prove that if n is a multiple of 3, then n^2 is also a multiple of 3.
   (b) Prove that if n is a multiple of 7, then n^3 is also a multiple of 7.

6. Prove that if n is not a multiple of 3, then n^2 is also not a multiple of 3.

   Hint: If n is not a multiple of 3, then $n = 3q + 1$ or $n = 3q + 2$ for *some* integer q.

7. Use the facts that

   (i) $\sqrt{2}$ is irrational, and
   (ii) if x is irrational, then $\sqrt{x}$ is also irrational,

   to prove that $\sqrt[8]{2}$ is irrational.

8. Recall that we can use a counterexample to disprove an implication. Show that the following claims are false:

   (a) If x and y are integers such that $x^2 > y^2$, then $x > y$.
   (b) If n is a positive integer, then $n^2 + n + 41$ is prime.

9. Explain why the following arguments are invalid:

   (a) Let n be an integer. If n^2 is odd, then n is odd. Therefore, n must be odd.
   (b) Let n be an integer. If n is even, then n^2 is also even. As an integer, n^2 could be odd. Hence, n cannot be even. Therefore, n must be odd.

10. Analyze the following reasoning:

    (a) Let S be a set of real numbers. If x is in S, then x^2 is in S. But x is not in S, hence x^2 is not in S.
    (b) Let S be a set of real numbers. If x is in S, then x^2 is in S. Therefore, if x^2 is in S, then x is in S.

3.3 Indirect Proofs

Instead of proving $p \Rightarrow q$ directly, it is sometimes easier to prove it indirectly. There are two kinds of ***indirect proofs***: the proof by contrapositive, and the proof by contradiction.

The ***proof by contrapositive*** is based on the fact that an implication is equivalent to its contrapositive. Therefore, instead of proving $p \Rightarrow q$, we may prove its contrapositive $\overline{q} \Rightarrow \overline{p}$. Since it is an implication, we could use a direct proof:

Proof by contrapositive.

1. Assume $\overline{q}$ is true (hence, assume q is false).
2. Show that $\overline{p}$ is true (that is, show that p is false).

The proof may proceed as follow:

> *Proof:* We want to prove the contrapositive of the stated result. Assume q is false, ...
>
> $\vdots$
>
> ... Therefore p is false. ■

Example 3.3.1 Let n be an integer. Show that if n^2 is even, then n is also even.

Remember to describe your goal.

Solution: Proof by contrapositive: We want to prove that if n is odd, then n^2 is odd. If n is odd, then $n = 2t + 1$ for some integer t. Hence,

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

is odd. This completes the proof. ▲

Example 3.3.2 Show that if n is a positive integer such that the sum of its positive divisors is $n + 1$, then n is prime.

Solution: We shall prove the contrapositive of the given statement. We want to prove that if n is composite, then the sum of its positive divisors is not $n+1$. Let n be a composite number. Then its divisors include 1, n, and at least one other positive divisor x different from 1 and n. So the sum of its positive divisors is at least $1 + n + x$. Since x is positive, we gather that

$$1 + n + x > 1 + n.$$

We deduce that the sum of the divisors cannot be $n + 1$. Therefore, if the sum of the divisors of n is precisely $n + 1$, then n must be prime. ▲

Example 3.3.3 Let x be a real number. Prove that if $x^3 - 7x^2 + x - 7 = 0$, then $x = 7$.

Solution: Assume $x \neq 7$, then

$$x^3 - 7x^2 + x - 7 = x^2(x - 7) + (x - 7) = (x^2 + 1)(x - 7) \neq 0.$$

Thus, if $x^3 - 7x^2 + x - 7 = 0$, then $x = 7$. ▲

Hands-On Exercise 3.3.1 Let x be a real number. Prove that if $(2x^2 + 3)(x + 5)(x - 7) = 0$, then either $x = -5$, or $x = 7$.

△

Hands-On Exercise 3.3.2 Let x and y be two real numbers. Prove that if $x \neq 0$ and $y \neq 0$, then $xy \neq 0$.

△

Another indirect proof is the ***proof by contradiction***. To prove that $p \Rightarrow q$, we proceed as follows:

Proof by contradiction.

1. Suppose $p \Rightarrow q$ is false; that is, assume that p is true and q is false.
2. Argue until we obtain a contradiction, which could be any result that we know is false.

How does this prove that $p \Rightarrow q$? Assuming that the logic used in every step in the argument is correct, yet we still end up with a contradiction, then the only possible flaw must come from the supposition that $p \Rightarrow q$ is false. Consequently, $p \Rightarrow q$ must be true.

This is what a typical proof by contradiction may look like:

> *Proof:* Suppose $p \Rightarrow q$ is false. Then p is true and q is false. Then ...
>
> $\vdots$
>
> ..., which is a contradiction. Therefore, $p \Rightarrow q$ must be true. ■

There is a more general form for proving a statement r, which needs not be an implication. To prove the proposition r by contradiction, we follow these steps:

Another form of proof by contradiction.

1. Suppose r is false.
2. Argue until we obtain a contradiction.

> *Proof:* Suppose r is false. Then ...
>
> $\vdots$
>
> ..., which is a contradiction. Therefore, r must be true. ■

Example 3.3.4 Show that if $x^3 - 7x^2 + x - 7 = 0$, then $x = 7$.

Solution: Assume $x^3 - 7x^2 + x - 7 = 0$, we want to show that $x = 7$. Suppose $x \neq 7$, then $x - 7 \neq 0$, and

$$0 = x^3 - 7x^2 + x - 7 = x^2(x-7) + (x-7) = (x^2+1)(x-7)$$

would have implied that $x^2 + 1 = 0$, which is impossible. Therefore, we must have $x = 7$. ▲

Example 3.3.5 Show that if P is a point not on a line L, then there exists exactly one perpendicular line from P onto L.

Solution: Suppose we can find more than one perpendicular line from P onto L. Pick any two of them, and denote their intersections with L as Q and R. Then we have a triangle PQR, where the angles PQR and PRQ are both $90°$. This implies that the sum of the interior angles

of the triangle PQR exceeds $180°$, which is impossible. Hence, there is only one perpendicular line from P onto L. ▲

Example 3.3.6 Show that if $x^2 < 5$, then $|x| < \sqrt{5}$.

Solution: Assume $x^2 < 5$, we want to show that $|x| < \sqrt{5}$. Suppose, on the contrary, we have $|x| \geq \sqrt{5}$. Then either $x \geq \sqrt{5}$, or $x \leq -\sqrt{5}$. If $x \geq \sqrt{5}$, then $x^2 \geq 5$. If $x \leq -\sqrt{5}$, we again have $x^2 \geq 5$. In either case, we have a contradiction. Hence $|x| < \sqrt{5}$. ▲

Hands-On Exercise 3.3.3 Prove that if $x^2 \geq 49$, then $|x| \geq 7$.

△

Example 3.3.7 Prove that the logical formula

$$[(p \Rightarrow q) \wedge p] \Rightarrow q$$

is a tautology.

Solution: Suppose $[(p \Rightarrow q) \wedge p] \Rightarrow q$ is false for some statements p and q. Then we find

- $(p \Rightarrow q) \wedge p$ is true, and
- q is false.

For the conjunction $(p \Rightarrow q) \wedge p$ to be true, we need

- $p \Rightarrow q$ to be true, and
- p to be true.

Having p true and q false would make $p \Rightarrow q$ false. This directly contradicts what we have found. Therefore, the logical formula $[(p \Rightarrow q) \wedge p] \Rightarrow q$ is always true, hence it is a tautology. ▲

Example 3.3.8 Prove, by contradiction, that if x is rational and y is irrational, then $x + y$ is irrational.

Solution: Let x be a rational number and y an irrational number. We want to show that $x + y$ is irrational. Suppose, on the contrary, that $x + y$ is rational. Then

$$x + y = \frac{m}{n}$$

for some integers m and n, where $n \neq 0$. Since x is rational, we also have

$$x = \frac{p}{q}$$

for some integers p and q, where $q \neq 0$. It follows that

$$\frac{m}{n} = x + y = \frac{p}{q} + y.$$

Hence,

$$y = \frac{m}{n} - \frac{p}{q} = \frac{mq - np}{nq},$$

where $mq - np$ and nq are both integers, with $nq \neq 0$. This makes y rational, which contradicts the assumption that y is irrational. Thus, $x + y$ cannot be rational, it must be irrational. ▲

Hands-On Exercise 3.3.4 Prove that

$$\sqrt{x+y} \neq \sqrt{x} + \sqrt{y}$$

for any positive real numbers x and y.

Hint. The words "for any" suggest this is a universal quantification. Be sure you negate the problem statement properly.

△

Example 3.3.9 Prove that $\sqrt{2}$ is irrational.

Solution: Suppose, on the contrary, $\sqrt{2}$ is rational. Then we can write

$$\sqrt{2} = \frac{m}{n}$$

for some positive integers m and n such that m and n do not share any common divisor except 1 (hence $\frac{m}{n}$ is in its simplest term). Squaring both sides and cross-multiplying yields

$$2n^2 = m^2.$$

Thus, 2 divides m^2. Consequently, 2 must also divide m. Then we can write $m = 2s$ for some integer s. The equation above becomes

$$2n^2 = m^2 = (2s)^2 = 4s^2.$$

Hence,

$$n^2 = 2s^2,$$

which implies that 2 divides n^2; thus, 2 also divides n. We have proved that both m and n are divisible by 2. This contradicts the assumption that m and n do not share any common divisor. Therefore, $\sqrt{2}$ must be irrational. ▲

Hands-On Exercise 3.3.5 Prove that $\sqrt{3}$ is irrational.

△

Very often, a proof by contradiction can be rephrased into a proof by contrapositive or even a direct proof, both of which are easier to follow. If this is the case, rewrite the proof.

Example 3.3.10 Show that $x^2 + 4x + 6 = 0$ has no real solution. In symbols, show that $\nexists x \in \mathbb{R}\,(x^2 + 4x + 6 = 0)$.

Solution: Consider the following proof by contradiction:

> Suppose there exists a real number x such that $x^2 + 4x + 6 = 0$. Using calculus, it can be shown that the function $f(x) = x^2 + 4x + 6$ has an absolute minimum at $x = -2$. Thus, $f(x) \geq f(-2) = 2$ for any x. This contradicts the assumption that there exists an x such that $x^2 + 4x + 6 = 0$. Thus, $x^2 + 4x + 6 = 0$ has no real solution.

A close inspection reveals that we do not really need a proof by contradiction. The crux of the proof is the fact that $x^2 + 4x + 6 \geq 2$ for all x. This already shows that $x^2 + 4x + 6$ could never be zero. It is easier to use a direct proof, as follows.

> Using calculus, we find that the function $f(x) = x^2 + 4x + 6$ has an absolute minimum at $x = -2$. Therefore, for any x, we always have $f(x) \geq f(-2) = 2$. Hence, there does not exist any x such that $x^2 + 4x + 6 = 0$.

Do you agree that the second proof (the direct proof) is more elegant? ▲

Recall that a biconditional statement $p \Leftrightarrow q$ consists of two implications $p \Rightarrow q$ and $q \Rightarrow p$. Hence, to prove $p \Leftrightarrow q$, we need to establish these two "directions" separately.

Example 3.3.11 Let n be an integer. Prove that n^2 is even if and only if n is even.

Solution: ($\Rightarrow$) We first prove that if n^2 is even, then n must be even. We shall prove its contrapositive: if n is odd, then n^2 is odd. If n is odd, then we can write $n = 2t + 1$ for some integer t. Then

$$n^2 = (2t + 1) = 4t^2 + 4t + 1 = 2(2t^2 + 2t) + 1,$$

where $2t^2 + 2t$ is an integer. Thus, n^2 is odd.

($\Leftarrow$) Next, we prove that if n is even, then n^2 is even. If n is even, we can write $n = 2t$ for some integer t. Then

$$n^2 = (2t)^2 = 4t^2 = 2 \cdot 2t^2,$$

where $2t^2$ is an integer. Hence, n^2 is even, which completes the proof. ▲

Hands-On Exercise 3.3.6 Let n be an integer. Prove that n is odd if and only if n^2 is odd.

△

Summary and Review

- We can use indirect proofs to prove an implication.
- There are two kinds of indirect proofs: proof by contrapositive and proof by contradiction.
- In a proof by contrapositive, we actually use a direct proof to prove the contrapositive of the original implication.
- In a proof by contradiction, we start with the supposition that the implication is false, and use this assumption to derive a contradiction. This would prove that the implication must be true.
- A proof by contradiction can also be used to prove a statement that is not of the form of an implication. We start with the supposition that the statement is false, and use this assumption to derive a contradiction. This would prove that the statement must be true.
- Sometimes a proof by contradiction can be rewritten as a proof by contrapositive or even a direct proof. If this is true, rewrite the proof.

Exercises 3.3

1. Let n be an integer. Prove that if n^2 is even, then n must be even. Use

   (a) A proof by contrapositive.

   (b) A proof by contradiction.

   Remark: The two proofs are very similar, but the wording is slightly different, so be sure you present your proofs clearly.

2. Let n be an integer. Show that if n^2 is a multiple of 3, then n must also be a multiple of 3. Use

   (a) A proof by contrapositive.

   (b) A proof by contradiction.

3. Let n be an integer. Prove that if n is even, then $n^2 = 4s$ for some integer s.

4. Let m and n be integers. Show that $mn = 1$ implies that $m = 1$ or $m = -1$.

5. Let x be a real number. Prove by contrapositive: if x is irrational, then $\sqrt{x}$ is irrational. Apply this result to show that $\sqrt[4]{2}$ is irrational, using the assumption that $\sqrt{2}$ is irrational.

6. Let x and y be real numbers such that $x \neq 0$. Prove that if x is rational, and y is irrational, then xy is irrational.

7. Prove that $\sqrt{5}$ is irrational.

8. Prove that $\sqrt[3]{2}$ is irrational.

9. Let a and b be real numbers. Show that if $a \neq b$, then $a^2 + b^2 \neq 2ab$.

10. Use contradiction to prove that, for all integers $k \geq 1$,

$$2\sqrt{k+1} + \frac{1}{\sqrt{k+1}} \geq 2\sqrt{k+2}.$$

11. Let m and n be integers. Show that mn is even if and only if m is even or n is even.

12. Let x and y be real numbers. Show that $x^2 + y^2 = 0$ if and only if $x = 0$ and $y = 0$.

13. Prove that, if x is a real number such that $0 < x < 1$, then $x(1-x) \leq \frac{1}{4}$.

14. Let m and n be positive integers such that 3 divides mn. Show that 3 divides m, or 3 divides n.

15. Prove that the logical formula
$$(p \Rightarrow q) \vee (p \Rightarrow \overline{q})$$
is a tautology.

16. Prove that the logical formula
$$[(p \Rightarrow q) \wedge (p \Rightarrow \overline{q})] \Rightarrow \overline{p}$$
is a tautology.

3.4 Mathematical Induction: An Introduction

Mathematical induction can be used to prove that an identity is valid for all integers $n \geq 1$. Here is a typical example of such an identity:

$$1 + 2 + 3 + \cdots + n = \frac{n(n+1)}{2}.$$

More generally, we can use mathematical induction to prove that a propositional function $P(n)$ is true for all integers $n \geq 1$.

Mathematical Induction. To show that a propositional function $P(n)$ is true for all integers $n \geq 1$, follow these steps:

1. ***Basis Step***: Verify that $P(1)$ is true.
2. ***Inductive Step***: Show that if $P(k)$ is true for some integer $k \geq 1$, then $P(k+1)$ is also true.

The basis step is also called the ***anchor step*** or the ***initial step***. This proof technique is valid because of the next theorem.

Theorem 3.4.1 (Principle of Mathematical Induction) *If $S \subseteq \mathbb{N}$ such that*

(i) $1 \in S$, and
(ii) $k \in S \Rightarrow k + 1 \in S$,

then $S = \mathbb{N}$.

Remark. Here is a sketch of the proof. From (i), we know that $1 \in S$. It then follows from (ii) that $2 \in S$. Applying (ii) again, we find that $3 \in S$. Likewise, $4 \in S$, then $5 \in S$, and so on. Since this argument can go on indefinitely, we find that $S = \mathbb{N}$.

There is a subtle problem with this argument. It is unclear why "and so on" will work. After all, what does "and so on" or "continue in this manner" really mean? Can it really continue indefinitely? The trouble is, we do not have a formal definition of the natural numbers. It turns out that we cannot completely prove the principle of mathematical induction with just the usual properties for addition and multiplication. Consequently, we will take the theorem as an axiom without giving any formal proof. ♢

Although we cannot provide a satisfactory proof of the principle of mathematical induction, we can use it to justify the validity of the mathematical induction. Let S be the set of integers n for which a propositional function $P(n)$ is true. The basis step of mathematical induction verifies that $1 \in S$. The inductive step shows that $k \in S$ implies $k + 1 \in S$. Therefore, the principle of mathematical induction proves that $S = \mathbb{N}$. It follows that $P(n)$ is true for all integers $n \geq 1$.

The basis step and the inductive step, together, prove that

$$P(1) \Rightarrow P(2) \Rightarrow P(3) \Rightarrow \cdots .$$

Therefore, $P(n)$ is true for all integers $n \geq 1$. Compare induction to falling dominoes. When the first domino falls, it knocks down the next domino. The second domino in turn knocks down the third domino. Eventually, all the dominoes will be knocked down. But it will not happen unless these conditions are met:

Compare induction to the domino effect.

- The first domino must fall to start the motion. If it does not fall, no chain reaction will occur. This is the basis step.
- The distance between adjacent dominoes must be set up correctly. Otherwise, a certain domino may fall down without knocking over the next. Then the chain reaction will stop, and will never be completed. Maintaining the right inter-domino distance ensures that $P(k) \Rightarrow P(k+1)$ for each integer $k \geq 1$.

To prove the implication

$$P(k) \Rightarrow P(k+1)$$

in the inductive step, we need to carry out two steps: assuming that $P(k)$ is true, then using it to prove $P(k+1)$ is also true. So we can refine an induction proof into a 3-step procedure:

1. Verify that $P(1)$ is true.
2. Assume that $P(k)$ is true for some integer $k \geq 1$.
3. Show that $P(k+1)$ is also true.

The second step, the assumption that $P(k)$ is true, is sometimes referred to as the ***inductive hypothesis*** or ***induction hypothesis***. This is how a mathematical induction proof may look:

Proof: We proceed by induction on n. When $n = 1$, the left-hand side of the identity reduces to $\ldots$, and the right-hand side becomes $\ldots$. Hence, the identity holds when $n = 1$. Assume the identity holds when $n = k$ for some integer $k \geq 1$; that is, assume

$$\ldots \tag{3.1}$$

for some integer $k \geq 1$. We want to show that it also holds when $n = k + 1$; that is, we want to show that

$$\ldots .$$

Using the inductive hypothesis (3.1), we find

$$\ldots .$$

Therefore, the identity also holds when $n = k + 1$. This completes the induction. ■

The idea behind mathematical induction is rather simple. However, it must be delivered with precision.

(i) Be sure to say "Assume the identity holds for *some* integer $k \geq 1$." Do not say "Assume it holds for *all* integers $k \geq 1$." If we already know the result holds for all $k \geq 1$, then there is no need to prove anything at all.

(ii) Be sure to specify the requirement $k \geq 1$. This ensures that the chain reaction of the falling dominoes starts with the first one.

(iii) Do not say "let $n = k$" or "let $n = k + 1$." The point is, you are not assigning the value of k and $k + 1$ to n. Rather, you are *assuming* that the statement is true *when* n equals k, and using it to show that the statement also holds *when* n equals $k + 1$.

Example 3.4.1 Use mathematical induction to show that

$$1 + 2 + 3 + \cdots + n = \frac{n(n+1)}{2}$$

for all integers $n \geq 1$.

Discussion. In the basis step, it would be easier to check the two sides of the equation separately. The inductive step is the key step in any induction proof, and the last part, the part that proves $P(k+1)$ is true, is the most difficult part of the entire proof. In this regard, it is helpful to write out exactly what the inductive hypothesis proclaims, and what we really want to prove. In this problem, the inductive hypothesis claims that

$$1 + 2 + 3 + \cdots + k = \frac{k(k+1)}{2}.$$

We want to prove that $P(k+1)$ is also true. What does $P(k+1)$ really mean? It says

$$1 + 2 + 3 + \cdots + (k+1) = \frac{(k+1)(k+2)}{2}.$$

Compare the left-hand sides of these two equations. The first one is the sum of k quantities, and the second is the sum of $k+1$ quantities, and the extra quantity is the last number $k+1$. The sum of the first k terms is precisely what we have on the left-hand side of the inductive hypothesis. Hence, by writing

$$1 + 2 + 3 + \cdots + (k+1) = 1 + 2 + \cdots + k + (k+1),$$

we can regroup the right-hand side as

$$1 + 2 + 3 + \cdots + (k+1) = [1 + 2 + \cdots + k] + (k+1),$$

so that $1 + 2 + \cdots + k$ can be replaced by $\frac{k(k+1)}{2}$, according to the inductive hypothesis. With additional algebraic manipulation, we try to show that the sum does equal to $\frac{(k+1)(k+2)}{2}$. ♢

Solution: We proceed by induction on n. When $n = 1$, the left-hand side of the identity reduces to 1, and the right-hand side becomes $\frac{1 \cdot 2}{2} = 1$; hence, the identity holds when $n = 1$. Assume it holds when $n = k$ for some integer $k \geq 1$; that is, assume that

$$1 + 2 + 3 + \cdots + k = \frac{k(k+1)}{2}$$

for some integer $k \geq 1$. We want to show that it also holds when $n = k+1$. In other words, we want to show that

$$1 + 2 + 3 + \cdots + (k+1) = \frac{(k+1)(k+2)}{2}.$$

Using the inductive hypothesis, we find

$$\begin{aligned} 1 + 2 + 3 + \cdots + (k+1) &= 1 + 2 + 3 + \cdots + k + (k+1) \\ &= \frac{k(k+1)}{2} + (k+1) \\ &= (k+1)\left(\frac{k}{2} + 1\right) \\ &= (k+1) \cdot \frac{k+2}{2}. \end{aligned}$$

Therefore, the identity also holds when $n = k+1$. This completes the induction. ▲

We can use the ***summation notation*** (also called the ***sigma notation***) to abbreviate a sum. For example, the sum in the last example can be written as

$$\sum_{i=1}^{n} i.$$

The letter i is the ***index of summation***. By putting $i = 1$ under $\sum$ and n above, we declare that the sum starts with $i = 1$, and ranges through $i = 2$, $i = 3$, and so on, until $i = n$. The quantity that follows $\sum$ describes the pattern of the terms that we are adding in the summation. Accordingly,

$$\sum_{i=1}^{10} i^2 = 1^2 + 2^2 + 3^2 + \cdots + 10^2.$$

In general, the sum of the first n terms in a sequence $\{a_1, a_2, a_3, \dots\}$ is denoted $\sum_{i=1}^{n} a_i$. Observe that

$$\sum_{i=1}^{k+1} a_i = \left(\sum_{i=1}^{k} a_i\right) + a_{k+1},$$

which provides the link between $P(k+1)$ and $P(k)$ in an induction proof.

Example 3.4.2 Use mathematical induction to show that, for all integers $n \geq 1$,

$$\sum_{i=1}^{n} i^2 = 1^2 + 2^2 + 3^2 + \cdots + n^2 = \frac{n(n+1)(2n+1)}{6}.$$

Solution: We proceed by induction on n. When $n = 1$, the left-hand side reduces to $1^2 = 1$, and the right-hand side becomes $\frac{1 \cdot 2 \cdot 3}{6} = 1$; hence, the identity holds when $n = 1$. Assume it holds when $n = k$ for some integer $k \geq 1$; that is, assume that

$$\sum_{i=1}^{k} i^2 = \frac{k(k+1)(2k+1)}{6}$$

for some integer $k \geq 1$. We want to show that it still holds when $n = k + 1$. In other words, we want to show that

$$\sum_{i=1}^{k+1} i^2 = \frac{(k+1)(k+2)[2(k+1)+1]}{6} = \frac{(k+1)(k+2)(2k+3)}{6}.$$

From the inductive hypothesis, we find

$$\begin{aligned}
\sum_{i=1}^{k+1} i^2 &= \left(\sum_{i=1}^{k} i^2\right) + (k+1)^2 \\
&= \frac{k(k+1)(2k+1)}{6} + (k+1)^2 \\
&= \tfrac{1}{6}(k+1)[k(2k+1) + 6(k+1)] \\
&= \tfrac{1}{6}(k+1)(2k^2 + 7k + 6) \\
&= \tfrac{1}{6}(k+1)(k+2)(2k+3).
\end{aligned}$$

Therefore, the identity also holds when $n = k + 1$. This completes the induction. ▲

Example 3.4.3 Use mathematical induction to show that

$$3 + \sum_{i=1}^{n}(3+5i) = \frac{(n+1)(5n+6)}{2}$$

for all integers $n \geq 1$.

Solution: Proceed by induction on n. When $n = 1$, the left-hand side reduces to $3+(3+5) = 11$, and the right-hand side becomes $\frac{2\cdot 11}{2} = 11$; hence, the identity holds when $n = 1$. Assume it holds when $n = k$ for some integer $k \geq 1$; that is, assume that

$$3 + \sum_{i=1}^{k}(3+5i) = \frac{(k+1)(5k+6)}{2}$$

for some integer $k \geq 1$. We want to show that it still holds when $n = k+1$. In other words, we want to show that

$$3 + \sum_{i=1}^{k+1}(3+5i) = \frac{[(k+1)+1]\,[5(k+1)+6]}{2} = \frac{(k+2)(5k+11)}{2}.$$

From the inductive hypothesis, we find

$$\begin{aligned}
3 + \sum_{i=1}^{k+1}(3+5i) &= \left(3 + \sum_{i=1}^{k}(3+5i)\right) + [3+5(k+1)] \\
&= \frac{(k+1)(5k+6)}{2} + 5k + 8 \\
&= \tfrac{1}{2}\left[(k+1)(5k+6) + 2(5k+8)\right] \\
&= \tfrac{1}{2}\left(5k^2 + 21k + 22\right) \\
&= \tfrac{1}{2}\,(k+2)(5k+11).
\end{aligned}$$

This completes the induction. ▲

Hands-On Exercise 3.4.1 It is time for you to write your own induction proof. Prove that

$$1\cdot 2 + 2\cdot 3 + 3\cdot 4 + \cdots + n(n+1) = \frac{n(n+1)(n+2)}{3}$$

for all integers $n \geq 1$.

Remark. We give you a hand on this one, after which, you will be on your own. We lay out the template, all you need to do is fill in the blanks. ♢

Solution: Proceed by induction on n. When $n = 1$, the left-hand side reduces to . . . ,
and the right-hand side becomes . . . . Hence, ... Assume the identity holds when $n = k$ for . . . ; that is, assume

for some integer $k \geq 1$. We want to show that it also holds when $n = k+1$; that is, we want to show that

It follows from the inductive hypothesis that

$$= \qquad\qquad +$$

$$= \qquad\qquad +$$

$$= \qquad\qquad .$$

This completes the induction. ▲

Hands-On Exercise 3.4.2 Use induction to prove that, for all positive integers n,

$$1 \cdot 2 \cdot 3 + 2 \cdot 3 \cdot 4 + \cdots + n(n+1)(n+2) = \frac{n(n+1)(n+2)(n+3)}{4}.$$

△

Hands-On Exercise 3.4.3 Use induction to prove that, for all positive integers n,

$$1 + 4 + 4^2 + \cdots + 4^n = \frac{1}{3}\,(4^{n+1} - 1).$$

△

All three steps in an induction proof must be completed; otherwise, the proof may not be correct.

Example 3.4.4 *Never attempt to prove $P(k) \Rightarrow P(k+1)$ by examples alone.* Consider

$$P(n): \qquad n^2 + n + 11 \text{ is prime.}$$

In the inductive step, we want to prove that

$$P(k) \Rightarrow P(k+1) \qquad \text{for } \textit{any } k \geq 1.$$

The following table verifies that it is true for $1 \leq k \leq 8$:

n	1	2	3	4	5	6	7	8	9
$n^2 + n + 11$	13	17	23	31	41	53	67	83	101

Nonetheless, when $n = 10$, $n^2 + n + 11 = 121$ is composite. So $P(9) \not\Rightarrow P(10)$. The inductive step breaks down when $k = 9$. ▲

Example 3.4.5 *The basis step is equally important.* Consider proving

$$P(n): \qquad 3n + 2 = 3q \text{ for some integer } q$$

for all $n \in \mathbb{N}$. Assume $P(k)$ is true for some integer $k \geq 1$; that is, assume $3k + 2 = 3q$ for some integer q. Then

$$3(k+1) + 2 = 3k + 3 + 2 = 3 + 3q = 3(1+q).$$

Therefore, $3(k+1) + 2$ can be written in the same form. This proves that $P(k+1)$ is also true. Does it follow that $P(n)$ is true for all integers $n \geq 1$? We know that $3n + 2$ cannot be written as a multiple of 3. What is the problem?

Solution: The problem is: we need $P(k)$ to be true for at least one value of k so as to start the sequence of implications

$$P(1) \Rightarrow P(2), \qquad P(2) \Rightarrow P(3), \qquad P(3) \Rightarrow P(4), \qquad \ldots$$

The induction fails because we have not established the basis step. In fact, $P(1)$ is false. Since the first domino does not fall, we cannot even start the chain reaction. ▲

Remark. Thus far, we have learned how to use mathematical induction to prove identities. In general, we can use mathematical induction to prove a statement about n. This statement can take the form of an identity, an inequality, or simply a verbal statement about n. We shall learn more about mathematical induction in the next few sections. ♢

Summary and Review

- Mathematical induction can be used to prove that a statement about n is true for all integers $n \geq 1$.
- We have to complete three steps.
- In the basis step, verify the statement for $n = 1$.
- In the inductive hypothesis, assume that the statement holds when $n = k$ for some integer $k \geq 1$.
- In the inductive step, use the information gathered from the inductive hypothesis to prove that the statement also holds when $n = k + 1$.
- Be sure to complete all three steps.
- Pay attention to the wording. At the beginning, follow the template closely. When you feel comfortable with the whole process, you can start venturing out on your own.

Exercises 3.4

1. Use induction to prove that

$$1^3 + 2^3 + 3^3 + \cdots + n^3 = \frac{n^2(n+1)^2}{4}$$

for all integers $n \geq 1$.

2. Use induction to prove that the following identity holds for all integers $n \geq 1$:

$$1 + 3 + 5 + \cdots + (2n - 1) = n^2.$$

3. Use induction to show that

$$1 + \frac{1}{3} + \frac{1}{3^2} + \cdots + \frac{1}{3^n} = \frac{3}{2}\left(1 - \frac{1}{3^{n+1}}\right)$$

for all positive integers n.

4. Use induction to establish the following identity for any integer $n \geq 1$:

$$1 - 3 + 9 - \cdots + (-3)^n = \frac{1 - (-3)^{n+1}}{4}.$$

5. Use induction to show that, for any integer $n \geq 1$:

$$\sum_{i=1}^{n} i \cdot i! = (n+1)! - 1.$$

6. Use induction to prove the following identity for integers $n \geq 1$:

$$\sum_{i=1}^{n} \frac{1}{(2i-1)(2i+1)} = \frac{n}{2n+1}.$$

7. Evaluate $\displaystyle\sum_{i=1}^{n} \frac{1}{i(i+1)}$ for a few values of n. What do you think the result should be? Use induction to prove your conjecture.

8. Use induction to prove that

$$\sum_{i=1}^{n} (2i-1)^3 = n^2(2n^2 - 1)$$

whenever n is a positive integer.

9. Use induction to show that, for any integer $n \geq 1$:

$$1^2 - 2^2 + 3^2 - \cdots + (-1)^{n-1}n^2 = (-1)^{n-1}\,\frac{n(n+1)}{2}.$$

10. Use mathematical induction to show that

$$\sum_{i=1}^{n} \frac{i+4}{i(i+1)(i+2)} = \frac{n(3n+7)}{2(n+1)(n+2)}$$

for all integers $n \geq 1$.

3.5 More on Mathematical Induction

Besides identities, we can also use mathematical induction to prove a statement about a positive integer n.

Example 3.5.1 Prove that $n(n+1)(2n+1)$ is a multiple of 6 for all integers $n \geq 1$.

See Hands-On Exercise 3.2.4.

Remark. We have already seen how to prove this claim using a proof by cases, which is actually an easier way to prove that $n(n+1)(2n+1)$ is divisible by 6. Nonetheless, we shall demonstrate below how to use induction to prove the claim. ♢

Discussion. In the inductive hypothesis, it is clear that we are assuming $k(k+1)(2k+1)$ is a multiple of 6. In the inductive step, we want to prove that

$$(k+1)(k+2)[2(k+1)+1] = (k+1)(k+2)(2k+3)$$

is also a multiple of 6. A multiple of 6 can be written as $6q$ for some integer q. Since we have two multiples of 6, we need to write

$$k(k+1)(2k+1) = 6q$$

and

$$(k+1)(k+2)(2k+3) = 6Q$$

to distinguish them. By using the lowercase and uppercase of the same letter, we indicate that they are different values. Yet, because they come from the same letter, they both share some common attribute, in this case, being the quotients when the respective values are divided by 6.

Now, in the inductive step, we need to make use of the equation $k(k+1)(2k+1) = 6q$ from the inductive hypothesis. This calls for connecting the product $(k+1)(k+2)(2k+3)$ to the expression $k(k+1)(2k+1)$. Since they share the common factor $k+1$, what remains to do is write $(k+2)(2k+3)$ in terms of $k(2k+1)$.

We are asked to prove that $n(n+1)(2n+1)$ is a multiple of 6. This is not an identity. Therefore, do not say "assume/show that the *identity* holds when . . . ." Instead, say "assume/show that the *claim* is true when . . . ." ♢

Solution: Proceed by induction on n. When $n = 1$, we have $n(n+1)(2n+1) = 1 \cdot 2 \cdot 3 = 6$, which is clearly a multiple of 6. Hence, the claim is true when $n = 1$. Assume the claim is true when $n = k$ for some integer $k \geq 1$; that is, assume that we can write

$$k(k+1)(2k+1) = 6q$$

for some integer q. We want to show that the claim is still true when $n = k+1$; that is, we want to show that

$$(k+1)(k+2)[2(k+1)+1] = (k+1)(k+2)(2k+3) = 6Q$$

for some integer Q. Using the inductive hypothesis, we find

$$\begin{aligned}(k+1)(k+2)(2k+3) &= (k+1)(2k^2+7k+6)\\ &= (k+1)[(2k^2+k)+(6k+6)]\\ &= (k+1)[k(2k+1)+6(k+1)]\\ &= k(k+1)(2k+1)+6(k+1)^2\\ &= 6q+6(k+1)^2\\ &= 6\,[q+(k+1)^2],\end{aligned}$$

where $q+(k+1)^2$ is clearly an integer. This completes the induction. ▲

Hands-On Exercise 3.5.1 Prove that n^2+3n+2 is even for all integers $n \geq 1$.

△

Induction can also be used to prove inequalities, which often require more work to finish.

Example 3.5.2 Prove that

$$1+\frac{1}{4}+\cdots+\frac{1}{n^2} \leq 2-\frac{1}{n}$$

for all positive integers n.

Draft. In the inductive hypothesis, we assume that the inequality holds when $n=k$ for some integer $k \geq 1$. This means we assume

$$\sum_{i=1}^{k} \frac{1}{i^2} \leq 2-\frac{1}{k}.$$

In the inductive step, we want to show that it also holds when $n=k+1$. In other words, we want to prove that

$$\sum_{i=1}^{k+1} \frac{1}{i^2} \leq 2-\frac{1}{k+1}.$$

In order to use the inductive hypothesis, we have to find a connection between these two inequalities. Obviously, we have

$$\sum_{i=1}^{k+1} \frac{1}{i^2} = \left(\sum_{i=1}^{k} \frac{1}{i^2}\right) + \frac{1}{(k+1)^2}.$$

Hence, it follows from the inductive hypothesis that

$$\sum_{i=1}^{k+1} \frac{1}{i^2} = \left(\sum_{i=1}^{k} \frac{1}{i^2}\right) + \frac{1}{(k+1)^2} \leq 2 - \frac{1}{k} + \frac{1}{(k+1)^2}.$$

The proof would be complete if we could show that

$$2 - \frac{1}{k} + \frac{1}{(k+1)^2} \leq 2 - \frac{1}{k+1}.$$

There is no guarantee that this idea will work, but this should be the first thing we try.

After rearrangement, the inequality becomes

$$\frac{1}{k+1} + \frac{1}{(k+1)^2} \leq \frac{1}{k},$$

which is equivalent to $\frac{k+2}{(k+1)^2} \leq \frac{1}{k}$. Cross-multiplication yields

$$k(k+2) \leq (k+1)^2.$$

Since

$$k(k+2) = k^2 + 2k, \qquad \text{and} \qquad (k+1)^2 = k^2 + 2k + 1,$$

it is clear that what we want to prove is indeed true. ♢

Polish It Up! Next, we rearrange the argument to make it read more smoothly. Essentially all we need is to run the argument *backward*. To improve the flow of the argument, we can prove a separate result on the side before we return to the main argument. ♢

Proof 1: Proceed by induction on n. When $n = 1$, the left-hand side becomes 1, and so does the right-hand side; thus, the inequality holds. Assume it holds when $n = k$ for some integer $k \geq 1$:

$$\sum_{i=1}^{k} \frac{1}{i^2} \leq 2 - \frac{1}{k}.$$

We want to show that it also holds when $n = k + 1$:

$$\sum_{i=1}^{k+1} \frac{1}{i^2} \leq 2 - \frac{1}{k+1}.$$

To finish the proof, we need to derive an inequality. Notice that

$$k(k+2) = k^2 + 2k < k^2 + 2k + 1 = (k+1)^2.$$

Hence, after dividing both sides by $k(k+1)^2$, we obtain

$$\frac{k+2}{(k+1)^2} < \frac{1}{k}.$$

This leads to

$$\frac{1}{k+1} + \frac{1}{(k+1)^2} = \frac{(k+1)+1}{(k+1)^2} = \frac{k+2}{(k+1)^2} < \frac{1}{k},$$

which is equivalent to

$$-\frac{1}{k} + \frac{1}{(k+1)^2} < -\frac{1}{k+1}. \tag{3.2}$$

We now return to our original problem. It follows from the inductive hypothesis and (3.2) that

$$\begin{aligned}\sum_{i=1}^{k+1}\frac{1}{i^2} &= \left(\sum_{i=1}^{k}\frac{1}{i^2}\right)+\frac{1}{(k+1)^2}\\ &\le 2-\frac{1}{k}+\frac{1}{(k+1)^2}\\ &< 2-\frac{1}{k+1}.\end{aligned}$$

Therefore, the inequality still holds when $n = k + 1$, which completes the induction. ▲

Remark. The key step in the proof is to establish (3.2), which can be done by means of contradiction. ♢

Proof 2: Proceed by induction on n. When $n = 1$, the left-hand side becomes 1, and so does the right-hand side; thus, the inequality holds. Assume it holds when $n = k$ for some integer $k \ge 1$:

$$\sum_{i=1}^{k}\frac{1}{i^2} \le 2-\frac{1}{k}.$$

We want to show that it also holds when $n = k + 1$:

$$\sum_{i=1}^{k+1}\frac{1}{i^2} \le 2-\frac{1}{k+1}.$$

To finish the proof, we need the following inequality. We claim that

$$-\frac{1}{k}+\frac{1}{(k+1)^2} < -\frac{1}{k+1}. \tag{3.3}$$

Suppose, on the contrary, that

$$-\frac{1}{k}+\frac{1}{(k+1)^2} \ge -\frac{1}{k+1}.$$

Clear the denominators by multiplying $k(k+1)^2$ to both sides of the inequality. We find

$$-(k+1)^2+k \ge -k(k+1),$$

or equivalently,

$$-k^2-k-1 \ge -k^2-k,$$

which is the same as saying $-1 \ge 0$. This contradiction proves that (3.3) must be true.

We now return to our original problem. It follows from the inductive hypothesis and (3.3) that

$$\begin{aligned}\sum_{i=1}^{k+1}\frac{1}{i^2} &= \left(\sum_{i=1}^{k}\frac{1}{i^2}\right)+\frac{1}{(k+1)^2}\\ &\le 2-\frac{1}{k}+\frac{1}{(k+1)^2}\\ &< 2-\frac{1}{k+1}.\end{aligned}$$

Therefore, the inequality still holds when $n = k + 1$, which completes the induction. ▲

Hands-On Exercise 3.5.2 Show that $n < 2^n$ for all integers $n \geq 1$.

△

We do not have to start with $n = 1$ in the basis step. We can start with any integer n_0.

Generalization. To show that $P(n)$ is true for all integers $n \geq n_0$, follow these steps:

1. Verify that $P(n_0)$ is true.
2. Assume that $P(k)$ is true for some integer $k \geq n_0$.
3. Show that $P(k+1)$ is also true.

The major difference is in the basis step: we need to verify that $P(n_0)$ is true. In addition, in the inductive hypothesis, we need to stress that $k \geq n_0$.

Example 3.5.3 Use mathematical induction to show that

$$\sum_{i=0}^{n} 4^i = \frac{1}{3}(4^{n+1} - 1)$$

for all integers $n \geq 0$.

Solution: Proceed by induction on n. When $n = 0$, the left-hand side reduces to $\sum_{i=0}^{0} 4^i = 4^0 = 1$, and the right-hand side becomes $\frac{1}{3}(4^1 - 1) = \frac{1}{3} \cdot 3 = 1$. Hence, the formula holds when $n = 0$. Assume it holds when $n = k$ for some integer $k \geq 0$; that is, assume

$$\sum_{i=0}^{k} 4^i = \frac{1}{3}(4^{k+1} - 1).$$

We want to show that it also holds when $n = k + 1$; that is,

$$\sum_{i=0}^{k+1} 4^i = \frac{1}{3}(4^{k+2} - 1).$$

Using the inductive hypothesis, we find

$$\begin{aligned}
\sum_{i=0}^{k+1} 4^i &= \left(\sum_{i=0}^{k} 4^i\right) + 4^{k+1} \\
&= \tfrac{1}{3}\left(4^{k+1} - 1\right) + 4^{k+1} \\
&= \tfrac{1}{3}\left(4^{k+1} - 1 + 3 \cdot 4^{k+1}\right) \\
&= \tfrac{1}{3}\left(4 \cdot 4^{k+1} - 1\right) \\
&= \tfrac{1}{3}\left(4^{k+2} - 1\right),
\end{aligned}$$

which is what we want to prove, thereby completing the induction. ▲

Hands-On Exercise 3.5.3 Prove that, for any integer $n \geq 0$,

$$1 + \frac{2}{3} + \frac{4}{9} + \cdots + \left(\frac{2}{3}\right)^n = 3\left[1 - \left(\frac{2}{3}\right)^{n+1}\right].$$

△

Example 3.5.4 Use mathematical induction to show that

$$n^n \geq 2^n$$

for all integers $n \geq 2$.

Solution: Proceed by induction on n. When $n = 2$, the inequality becomes $2^2 \geq 2^2$, which is obviously true. Assume it holds when $n = k$ for some integer $k \geq 2$:

$$k^k \geq 2^k.$$

We want to show that it still holds when $n = k + 1$:

$$(k+1)^{k+1} \geq 2^{k+1}.$$

Since $k \geq 2$, it follows from the inductive hypothesis that

$$(k+1)^{k+1} \geq k^{k+1} = k \cdot k^k \geq 2 \cdot 2^k = 2^{k+1}.$$

Therefore, the inequality still holds when $n = k+1$. This completes the induction. ▲

Summary and Review

- We can use induction to prove a general statement involving an integer n.
- The statement can be an identity, an inequality, or a claim about the property of an expression involving n.
- An induction proof need not start with $n = 1$.
- If we want to prove that a statement is true for all integers $n \geq n_0$, we have to verify the statement for $n = n_0$ in the basis step.
- In addition, we need to assume that $k \geq n_0$ in the inductive hypothesis.

Exercises 3.5

1. Use induction to prove that $n(n+1)(n+2)$ is a multiple of 3 for all integers $n \geq 1$.

2. Use induction to show that $n^3 + 5n$ is a multiple of 6 for any nonnegative integer n.

3. Use induction to prove that

$$2 + \left(1 + \frac{1}{\sqrt{2}} + \frac{1}{\sqrt{3}} + \cdots + \frac{1}{\sqrt{n}}\right) > 2\sqrt{n+1}$$

for all integers $n \geq 1$.

4. Use induction to prove that

$$2\left(1 + \frac{1}{8} + \frac{1}{27} + \cdots + \frac{1}{n^3}\right) \leq 3 - \frac{1}{n^2}$$

for all integers $n \geq 1$.

5. Use induction to prove that

$$a + ar + ar^2 + \cdots + ar^n = \frac{a(r^{n+1} - 1)}{r - 1}$$

for all nonnegative integers n, where a and r are real numbers with $r \neq 1$.

6. Use induction to prove that, for any integer $n \geq 2$,

$$6\sum_{i=2}^{n} i(i+2) = 2n^3 + 9n^2 + 7n - 18.$$

7. Use induction to prove that, for any integer $n \geq 0$,

$$1 - \frac{2}{5} + \frac{4}{25} + \cdots + \left(-\frac{2}{5}\right)^n = \frac{5}{7}\left[1 - \left(-\frac{2}{5}\right)^{n+1}\right].$$

8. Use induction to show that $n! > 2^n$ for all integers $n \geq 4$.

9. Use induction to prove that $n^2 > 4n + 1$ for all integers $n \geq 5$.

10. Prove that $2n + 1 < 2^n$ for all integers $n \geq 3$.

11. Define
$$S_n = \frac{1}{2!} + \frac{2}{3!} + \frac{3}{4!} + \cdots + \frac{n}{(n+1)!}.$$
   (a) Evaluate S_n for $n = 1, 2, 3, 4, 5$.
   (b) Propose a simple formula for S_n.
   (c) Use induction to prove your conjecture for all integers $n \geq 1$.

12. Define $T_n = \displaystyle\sum_{i=0}^{n} \frac{1}{(2i+1)(2i+3)}$.
   (a) Evaluate T_n for $n = 0, 1, 2, 3, 4$.
   (b) Propose a simple formula for T_n.
   (c) Use induction to prove your conjecture for all integers $n \geq 0$.

3.6 Mathematical Induction: The Strong Form

You may have heard of ***Fibonacci numbers***. They occur frequently in mathematics and life sciences. They have even been applied to study the stock market! Fibonacci numbers form a sequence every term of which, except the first two, is the sum of the previous two numbers. Mathematically, if we denote the nth Fibonacci number F_n, then

$$F_n = F_{n-1} + F_{n-2}. \tag{3.4}$$

This is called the ***recurrence relation*** for F_n.

Some students have trouble using (3.4): we are *not* adding $n-1$ and $n-2$. The subscripts only indicate the *locations* within the Fibonacci sequence. Hence, F_1 means the first Fibonacci number, F_2 the second Fibonacci number, and so forth. Compare this to dropping ten numbers into ten boxes, and each box is labeled with the numbers 1 through 10. Let us use a_i to denote the value in the ith box. When we say a_7, we do not mean the number 7. Instead, we mean the number stored in Box 7. Expressed in words, the recurrence relation (3.4) tells us that the nth Fibonacci number is the sum of the $(n-1)$th and the $(n-2)$th Fibonacci numbers. This is easy to remember: we add the last two Fibonacci numbers to get the next Fibonacci number.

The recurrence relation implies that we need to start with two initial values. We often start with $F_0 = 0$ (image F_0 as the zeroth Fibonacci number, the number stored in Box 0) and $F_1 = 1$. We combine the recurrence relation for F_n and its initial values together in one definition:

$$\boxed{F_0 = 0, \quad F_1 = 1, \qquad F_n = F_{n-1} + F_{n-2}, \quad \text{for } n \geq 2.}$$

We have to specify that the recurrence relation is valid only when $n \geq 2$, because this is the smallest value of n for which we can use the recurrence relation. What happens if you want to find F_1 using this formula? You will get $F_1 = F_0 + F_{-1}$, but F_{-1} is undefined!

The sum of the zeroth and the first Fibonacci numbers give us the second Fibonacci number:

$$F_2 = F_1 + F_0 = 1 + 0 = 1.$$

Continuing in this fashion, we find

$$\begin{array}{rclclcl}
F_3 & = & F_2 + F_1 & = & 1+1 & = & 2, \\
F_4 & = & F_3 + F_2 & = & 2+1 & = & 3, \\
F_5 & = & F_4 + F_3 & = & 3+2 & = & 5, \\
F_6 & = & F_5 + F_4 & = & 5+3 & = & 8, \\
\vdots & & \vdots & & \vdots & & \vdots
\end{array}$$

Following this pattern, what are the values of F_7 and F_8?

Fibonacci numbers enjoy many interesting properties, and there are numerous results concerning Fibonacci numbers. As a starter, consider the property

$$F_n < 2^n, \qquad n \geq 1.$$

How would we prove it by induction?

Since we want to prove that the inequality holds for all $n \geq 1$, we should check the case of $n = 1$ in the basis step. When $n = 1$, we have $F_1 = 1$ which is, of course, less than $2^1 = 2$. In the inductive hypothesis, we assume that the inequality holds when $n = k$ for some integer $k \geq 1$; that is, we assume

$$F_k < 2^k$$

for some integer $k \geq 1$. Next, we want to prove that the inequality still holds when $n = k + 1$. So we need to prove that

$$F_{k+1} < 2^{k+1}.$$

To make use of the inductive hypothesis, we need to apply the recurrence relation of Fibonacci numbers. It tells us that F_{k+1} is the sum of the previous two Fibonacci numbers; that is,

$$F_{k+1} = F_k + F_{k-1}.$$

The only thing we know from the inductive hypothesis is $F_k < 2^k$. So, as it stands, it does not tell us much about F_{k+1}.

A remedy is to assume in the inductive hypothesis that the inequality also holds when $n = k - 1$; that is, we also assume that

$$F_{k-1} < 2^{k-1}.$$

Therefore, unlike all the problems we have seen thus far, the inductive step in this problem relies on the last two n-values instead of just one. In terms of dominoes, imagine they are so heavy that we need the combined weight of two dominoes to knock down the next. Then

$$F_{k+1} = F_k + F_{k-1} < 2^k + 2^{k-1} = 2^{k-1}(2 + 1) < 2^{k-1} \cdot 2^2 = 2^{k+1},$$

which will complete the induction. This modified induction is known as the strong form of mathematical induction. In contrast, we call the ordinary mathematical induction the ***weak form*** of induction.

The proof still has a minor glitch! To be able to use the inductive hypothesis in the recurrence relation

$$F_{k+1} = F_k + F_{k-1},$$

both subscripts k and $k - 1$ must be at least 1, because the statement claims that $F_n < 2^n$ for all $n \geq 1$. This means we need $k \geq 2$. Consequently, in the basis step, we have to assume the inequality holds for *at least the first two values of* n.

In terms of the domino effect, the chain reaction of the falling dominoes starts at $k = 2$. We have to make sure that the first two dominoes will fall, so that their combined weight will knock down the third domino. Then the combined weight of the second and the third dominoes will knock over the fourth domino. The chain reaction will carry on indefinitely.

Symbolically, the ordinary mathematical induction relies on the implication $P(k) \Rightarrow P(k+1)$. Sometimes, $P(k)$ alone is not enough to prove $P(k + 1)$. In the case of proving $F_n < 2^n$, we actually use

$$[P(k - 1) \wedge P(k)] \Rightarrow P(k + 1).$$

We need to assume in the inductive hypothesis that the result is true when $n = k-1$ and $n = k$.

More generally, in the strong form of mathematical induction, we can use as many previous cases as we like to prove $P(k + 1)$.

Strong Form of Mathematical Induction. To show that $P(n)$ is true for all $n \geq n_0$, follow these steps:

1. Verify that $P(n)$ is true for some small values of $n \geq n_0$.
2. Assume that $P(n)$ is true for $n = n_0, n_0 + 1, \ldots, k$ for some integer $k \geq n^*$.
3. Show that $P(k+1)$ is also true.

The idea behind the inductive step is to show that

$$[\,P(n_0) \wedge P(n_0+1) \wedge \cdots \wedge P(k-1) \wedge P(k)\,] \Rightarrow P(k+1).$$

We could use more than one previous case, say, $n = k$, $k-1$, $k-2$, ..., to establish the next case $n = k+1$.

We may not need to use all of $P(n_0), P(n_0+1), \ldots, P(k-1), P(k)$. In fact, we may only need the last few of them, for example, $P(k-3), P(k-2), P(k-1)$ and $P(k)$. The number of previous cases required to establish $P(k+1)$ tells us how many initial cases we have to verify in the basis step. We do not know how many we need until the inductive step. For this reason, it is wise to start with a draft.

Example 3.6.1 Show that $F_n < 2^n$ for all $n \geq 1$.

Remark. We have already worked on the draft in the discussion above. We know that we need to verify the first two n-values in the basis step, and to assume that the inequality holds for at least two cases. ♢

Solution: Proceed by induction on n. When $n = 1$ and $n = 2$, we find

$$F_1 = 1 < 2 = 2^1,$$
$$F_2 = 1 < 4 = 2^2.$$

Therefore, the inequality holds when $n = 1, 2$. Assume it holds for $n = 1, 2, \ldots, k$, where $k \geq 2$. In particular, we have

$$F_k < 2^k, \qquad \text{and} \qquad F_{k-1} < 2^{k-1},$$

where $k \geq 2$. Then

$$F_{k+1} = F_k + F_{k-1} < 2^k + 2^{k-1} = 2^{k-1}(2+1) < 2^{k-1} \cdot 2^2 = 2^{k+1}.$$

Hence, the inequality still holds when $n = k+1$, which completes the induction. ▲

Recurrence relation can be used to define a sequence. For example, if the sequence $\{a_n\}_{n=1}^{\infty}$ is defined recursively by

$$a_n = 3a_{n-1} - 2 \qquad \text{for } n \geq 2,$$

with $a_1 = 4$, then

$$a_2 = 3a_1 - 2 = 3 \cdot 4 - 2 = 10,$$
$$a_3 = 3a_2 - 2 = 3 \cdot 10 - 2 = 28.$$

Identity involving such sequences can often be proved by means of induction.

Example 3.6.2 The sequence $\{b_n\}_{n=1}^{\infty}$ is defined as

$$b_1 = 5, \quad b_2 = 13, \qquad b_n = 5b_{n-1} - 6b_{n-2} \quad \text{for } n \geq 3.$$

Prove that $b_n = 2^n + 3^n$ for all $n \geq 1$.

Solution: Proceed by induction on n. When $n = 1$, the proposed formula for b_n says $b_1 = 2 + 3 = 5$, which agrees with the initial value $b_1 = 5$. When $n = 2$, the proposed formula claims $b_2 = 4 + 9 = 13$, which again agrees with the definition $b_2 = 13$. Assume the formula is valid for $n = 1, 2, \ldots, k$ for some integer $k \geq 2$. In particular, assume

$$b_k = 2^k + 3^k, \qquad \text{and} \qquad b_{k-1} = 2^{k-1} + 3^{k-1}.$$

We want to show that the formula still works when $n = k+1$. In other words, we want to show that

$$b_{k+1} = 2^{k+1} + 3^{k+1}.$$

Using the recurrence relation and the inductive hypothesis, we find

$$\begin{aligned}
b_{k+1} &= 5b_k - 6b_{k-1} \\
&= 5(2^k + 3^k) - 6(2^{k-1} + 3^{k-1}) \\
&= 5 \cdot 2^k + 5 \cdot 3^k - 6 \cdot 2^{k-1} - 6 \cdot 3^{k-1} \\
&= 5 \cdot 2^k + 5 \cdot 3^k - 2 \cdot 3 \cdot 2^{k-1} - 2 \cdot 3 \cdot 3^{k-1} \\
&= 5 \cdot 2^k + 5 \cdot 3^k - 3 \cdot 2^k - 2 \cdot 3^k \\
&= 2 \cdot 2^k + 3 \cdot 3^k \\
&= 2^{k+1} + 3^{k+1}
\end{aligned}$$

which is what we want to establish. This completes the induction, and hence, the claim that $b_n = 2^n + 3^n$. ▲

Hands-On Exercise 3.6.1 The sequence $\{c_n\}_{n=1}^{\infty}$ is defined as

$$c_1 = 7, \quad b_2 = 29, \qquad c_n = 5b_{n-1} - 6b_{n-2} \quad \text{for } n \geq 3.$$

Prove that $c_n = 5 \cdot 3^n - 4 \cdot 2^n$ for all integers $n \geq 1$.

△

Example 3.6.3 Show that all integers $n \geq 24$ can be expressed as $4x + 9y$ for some integers $x, y \geq 0$.

Definition. The expression $4x + 9y$ is called a ***linear combination*** of 4 and 9, and x and y are called the ***coefficients*** of the linear combination. ♢

Remark. We want to prove that any sufficiently large integer n can be written as a linear combination of 4 and 9 with nonnegative coefficients. This problem is called the ***postage stamp problem*** for the obvious reason: can we use only 4-cent and 9-cent stamps to obtain an n-cent postage for all integers $n \geq 24$? Not too surprisingly, it is also called the ***money changing problem*** (imagine replacing stamps with coins). ♢

Remark. The spirit behind mathematical induction (both weak and strong forms) is making use of what we know about a smaller size problem. In the weak form, we use the result from $n = k$ to establish the result for $n = k + 1$. In the strong form, we use some of the results from $n = k, k - 1, k - 2, \ldots$ to establish the result for $n = k + 1$. ♢

Discussion. Let us first look at the inductive step, in which we want to show that we can write $k + 1$ as a linear combination of 4 and 9. The key step of any induction proof is to relate the case of $n = k + 1$ to a problem with a smaller size (hence, with a smaller value in n).

Imagine you want to send a letter that requires a $(k+1)$-cent postage, and you can use only 4-cent and 9-cent stamps. You could first put down a 4-cent stamp. Then you still need to come up with the remaining postage of $(k + 1) - 4 = k - 3$ cents. If you could use 4-cent and 9-cent stamps to make up the remaining $(k - 3)$-cent postage, the problem is solved. Therefore, in the inductive hypothesis, we need to assume that it can be done when $n = k - 3$.

For the whole argument to work, $k - 3$ has to be within the range of the n-values that we consider. So we need $k - 3 \geq 24$; that is, we want $k \geq 27$. Consequently, we have to verify the claim for $n = 24, 25, 26, 27$ in the basis step. ♢

Solution: Proceed by induction on n. We find

$$\begin{aligned} 24 &= 4 \cdot 6 + 9 \cdot 0, \\ 25 &= 4 \cdot 4 + 9 \cdot 1, \\ 26 &= 4 \cdot 2 + 9 \cdot 2, \\ 27 &= 4 \cdot 0 + 9 \cdot 3. \end{aligned}$$

Hence, the claim is true when $n = 24, 25, 26, 27$. Assume it is true when $n = 24, 25, \ldots, k$ for some integer $k \geq 27$. In particular, since $k - 3 \geq 24$, this assumption assures that

$$k - 3 = 4x + 9y$$

for some nonnegative integers x and y. It follows that

$$\begin{aligned} k + 1 &= 4 + (k - 3) \\ &= 4 + 4x + 9y \\ &= 4(1 + x) + 9y, \end{aligned}$$

where $1+x$ and y are nonnegative integers. This shows that the claim is still true when $n = k+1$, thereby completing the induction. ▲

Hands-On Exercise 3.6.2 Show that all integers $n \geq 2$ can be expressed as $2x + 3y$ for some nonnegative integers x and y.

△

Summary and Review

- If, in the inductive step, we need to use more than one previous instance of the statement that we are proving, we may use the strong form of the induction.
- In such an event, we have to modify the inductive hypothesis to include more cases in the assumption.
- We also need to verify more cases in the basis step.
- Finally, we need to rewrite the whole proof to make it coherent.

Exercises 3.6

1. Use mathematical induction to prove the identity

$$F_1^2 + F_2^2 + F_3^2 + \cdots + F_n^2 = F_n F_{n+1}$$

for any integer $n \geq 1$.

2. Use induction to prove the following identity for all integers $n \geq 1$:

$$F_1 + F_3 + F_5 + \cdots + F_{2n-1} = F_{2n}.$$

3. Use induction to prove that

$$\frac{F_1}{F_2F_3} + \frac{F_2}{F_3F_4} + \frac{F_3}{F_4F_5} + \cdots + \frac{F_{n-2}}{F_{n-1}F_n} = 1 - \frac{1}{F_n}$$

for all integers $n \geq 3$.

4. Use induction to prove that any integer $n \geq 8$ can be written as a linear combination of 3 and 5 with nonnegative coefficients.

5. A football team may score a field goal for 3 points or[1] a touchdown (with conversion) for 7 points. Prove that, for any integer $n \geq 12$, it is possible for a football team to score n points with field goals and touchdowns.

6. An island country only issues 1-cent, 5-cent and 9-cent coins. Due to shortage in copper, all 1-cent coins were recalled. Prove that, using just 5-cent and 9-cent coins, one can pay an n-cent purchase for any $n \geq 32$.

[1] Although it is possible for a team to score 2 points for a safety or 8 points for a touchdown with a two-point conversion, we would not consider these possibilities in this simplified version of a real football game.

7. The sequence $\{b_n\}_{n=1}^{\infty}$ is defined recursively by

$$b_n = 3b_{n-1} - 2 \qquad \text{for } n \geq 2,$$

with $b_1 = 4$. Use induction to prove that $b_n = 3^n + 1$ for all $n \geq 1$.

8. The sequence $\{c_n\}_{n=1}^{\infty}$ is defined recursively as

$$c_1 = 3, \quad c_2 = -9, \qquad c_n = 7c_{n-1} - 10c_{n-2}, \quad \text{for } n \geq 3.$$

Use induction to show that $c_n = 4 \cdot 2^n - 5^n$ for all integers $n \geq 1$.

9. The sequence $\{d_n\}_{n=1}^{\infty}$ is defined recursively as

$$d_1 = 2, \quad d_2 = 56, \qquad d_n = d_{n-1} + 6d_{n-2}, \quad \text{for } n \geq 3.$$

Use induction to show that $d_n = 5(-2)^n + 4 \cdot 3^n$ for all integers $n \geq 1$.

10. The sequence $\{a_n\}_{n=1}^{\infty}$ is defined recursively as

$$a_1 = 2, \quad a_2 = 4, \qquad a_n = 2a_{n-1} + 3a_{n-2}, \quad \text{for } n \geq 3.$$

Use induction to show that $a_n > \left(\frac{5}{2}\right)^n$ for any integer $n \geq 4$.

Chapter 3

SWDM 4: Sets

Chapter 4

Sets

4.1 An Introduction

A ***set*** is a collection of objects. The objects in a set are called its ***elements*** or ***members***. The elements in a set can be any types of objects, including sets! The members of a set do not even have to be of the same type. For example, although it may not have any meaningful application, a set can consist of numbers and names.

We usually use capital letters such as A, B, C, S, and T to represent sets, and denote their generic elements by their corresponding lowercase letters a, b, c, s, and t, respectively. To indicate that b is an element of the set B, we adopt the notation $b \in B$, which means "b belongs to B" or "b is an element of B." We also write $B \ni b$, and say "B contains b (as an element)."

We designate these notations for some special sets of numbers:

$$\begin{aligned}
\mathbb{R} &= \text{the set of real numbers,} \\
\mathbb{Q} &= \text{the set of rational numbers,} \\
\mathbb{Z} &= \text{the set of integers,} \\
\mathbb{N} &= \text{the set of natural numbers (positive integers).}
\end{aligned}$$

All these are infinite sets, because they all contain infinitely many elements. In contrast, finite sets contain finitely many elements.

We can use the ***roster method*** to describe a set if it has only a small number of elements. We list all its elements explicitly, as in

$$A = \text{the set of natural numbers not exceeding } 7 = \{1, 2, 3, 4, 5, 6, 7\}.$$

For sets with more elements, show the first few entries to display a pattern, and use an ellipsis to indicate "and so on." For example,

$$\{1, 2, 3, \ldots, 20\}$$

represents the set of the first 20 positive integers. The repeating pattern can be extended indefinitely, as in

$$\begin{aligned}
\mathbb{N} &= \{1, 2, 3, \ldots\} \\
\mathbb{Z} &= \{\ldots, -2, -1, 0, 1, 2, \ldots\}
\end{aligned}$$

There are three kinds of integers: positive, negative, and the signless integer zero. In regards to parity, an integer is either even or odd. An integer is even if and only if it is divisible by two. Therefore, the set of even integers can be described as $\{\ldots, -4, -2, 0, 2, 4, \ldots\}$.

We can use a ***set-builder notation*** to describe a set. For example, the set of natural numbers is defined as

$$\mathbb{N} = \{x \in \mathbb{Z} \mid x > 0\}.$$

Here, the vertical bar | is read as "such that" or "for which." Hence, the right-hand side of the equation is pronounced as "the set of x belonging to the set of integers such that $x > 0$," or simply "the set of integers x such that $x > 0$." In general, this descriptive method appears in the format

$$\{\,\text{membership} \mid \text{properties}\,\}.$$

The notation | means "such that" or "for which" only when it is used in the set notation. It may mean something else in a different context. Therefore, *do not* write "let x be a real number | $x^2 > 3$" if you want to say " let x be a real number such that $x^2 > 3$." It is considered improper to use a mathematical notation as an abbreviation.

Example 4.1.1 Write these two sets

$$\{x \in \mathbb{Z} \mid x^2 \leq 1\} \quad \text{and} \quad \{x \in \mathbb{N} \mid x^2 \leq 1\}$$

by listing their elements explicitly.

Solution: The first set has three elements, and equals $\{-1, 0, 1\}$. The second set is a singleton set; it is equal to $\{1\}$. ▲

Hands-On Exercise 4.1.1 Use the roster method to describe the sets $\{x \in \mathbb{Z} \mid x^2 \leq 20\}$ and $\{x \in \mathbb{N} \mid x^2 \leq 20\}$.

△

Hands-On Exercise 4.1.2 Use the roster method to describe the set

$$\{x \in \mathbb{N} \mid x \leq 20 \text{ and } x = n^2 \text{ for some integer } n\}.$$

△

There is a slightly different format for the set-builder notation. Before the vertical bar, we describe the form the elements assume, and after the vertical bar, we indicate from where we are going to pick these elements:

$$\{\,\text{pattern} \mid \text{membership}\,\}.$$

Here the vertical bar | means "where." For example,

$$\{x^2 \mid x \in \mathbb{Z}\}$$

is the set of x^2 where $x \in \mathbb{Z}$. It represents the set of squares: $\{0, 1, 4, 9, 16, 25, \ldots\}$.

Example 4.1.2 The set

$$\{2n \mid n \in \mathbb{Z}\}$$

describes the set of even numbers. We can also write the set as $2\mathbb{Z}$. ▲

Hands-On Exercise 4.1.3 Describe the set $\{2n+1 \mid n \in \mathbb{Z}\}$ with the roster method.

△

Hands-On Exercise 4.1.4 Use the roster method to describe the set $\{3n \mid n \in \mathbb{Z}\}$.

△

An interval is a set of real numbers, all of which lie between two real numbers. Should the endpoints be included or excluded depends on whether the interval is ***open***, ***closed***, or ***half-open***. We adopt the following ***interval notation*** to describe them:

$$\begin{aligned}(a,b) &= \{x \in \mathbb{R} \mid a < x < b\},\\ [a,b] &= \{x \in \mathbb{R} \mid a \le x \le b\},\\ [a,b) &= \{x \in \mathbb{R} \mid a \le x < b\},\\ (a,b] &= \{x \in \mathbb{R} \mid a < x \le b\}.\end{aligned}$$

It is understood that a must be less than or equal to b. Hence, the notation $(5,3)$ does not make much sense. How about $[3,3]$? Is it a legitimate notation?

An interval contains not just integers, but all the numbers between the two endpoints. By numbers, we mean whole numbers *and* decimal numbers. For instance, $(1,5) \neq \{2,3,4\}$ because the interval $(1,5)$ also includes decimal numbers such at 1.276, $\sqrt{2}$, and π.

The interval (a,b) includes decimal numbers.

We can use $\pm\infty$ in the interval notation:

$$\begin{aligned}(a,\infty) &= \{x \in \mathbb{R} \mid a < x\},\\ (-\infty,a) &= \{x \in \mathbb{R} \mid x < a\}.\end{aligned}$$

However, we cannot write $(a,\infty]$ or $[-\infty,a)$, because $\pm\infty$ are *not* numbers. It is nonsense to say $x \le \infty$ or $-\infty \le x$. For the same reason, we can write $[a,\infty)$ and $(-\infty,a]$, but *not* $[a,\infty]$ or $[-\infty,a]$.

∞ is not a number, neither is $-\infty$

Example 4.1.3 Write the intervals $(2,3)$, $[2,3]$, and $(2,3]$ in the descriptive form.

Solution: According to the definition of an interval, we find

$$\begin{aligned}(2,3) &= \{x \in \mathbb{R} \mid 2 < x < 3\},\\ [2,3] &= \{x \in \mathbb{R} \mid 2 \le x \le 3\},\\ (2,3] &= \{x \in \mathbb{R} \mid 2 < x \le 3\}.\end{aligned}$$

What would you say about $[2,3)$? ▲

Example 4.1.4 Write these sets

$$\{x \in \mathbb{R} \mid -2 \le x < 5\} \quad \text{and} \quad \{x \in \mathbb{R} \mid x^2 \le 1\}$$

in the interval form.

Solution: The answers are $[-2,5)$ and $[-1,1]$, respectively. The membership of x affects the answers. If we change the second set to $\{x \in \mathbb{Z} \mid x^2 \le 1\}$, the answer would have been $\{-1,0,1\}$. Can you explain why $\{-1,0,1\} \neq [-1,1]$? ▲

Example 4.1.5 Be sure you are using the right types of numbers. Compare these two sets

$$\begin{aligned} S &= \{x \in \mathbb{Z} \mid x^2 \leq 5\}, \\ T &= \{x \in \mathbb{R} \mid x^2 \leq 5\}. \end{aligned}$$

One consists of integers only, while the other contains real numbers. Thus, $S = \{-2, -1, 0, 1, 2\}$, and $T = \left[-\sqrt{5}, \sqrt{5}\right]$. ▲

Hands-On Exercise 4.1.5 Which of the following sets

$$\{x \in \mathbb{Z} \mid 1 < x < 7\} \quad \text{and} \quad \{x \in \mathbb{R} \mid 1 < x < 7\}$$

can be represented by the interval notation $(1, 7)$? Explain.

△

Hands-On Exercise 4.1.6 Explain why $[2, 7] \neq \{2, 3, 4, 5, 6, 7\}$.

△

Hands-On Exercise 4.1.7 True or false: $(-2, 3) = \{-1, 0, 1, 2\}$? Explain.

△

Let S be a set of numbers; we define

$$\begin{aligned} S^+ &= \{x \in S \mid x > 0\}, \\ S^- &= \{x \in S \mid x < 0\}, \\ S^* &= \{x \in S \mid x \neq 0\}. \end{aligned}$$

In plain English, S^+ is the subset of S containing only those elements that are positive, S^- contains only the negative elements of S, and S^* contains only the nonzero elements of S.

Example 4.1.6 It should be obvious that $\mathbb{N} = \mathbb{Z}^+$. ▲

Hands-On Exercise 4.1.8 What is the notation for the set of negative integers?

△

See Example 2.1.8.

Some mathematicians also adopt these notations:

$$\begin{aligned} bS &= \{bx \mid x \in S\}, \\ a + bS &= \{a + bx \mid x \in S\}. \end{aligned}$$

Accordingly, we can write the set of even integers as $2\mathbb{Z}$, and the set of odd integers can be represented by $1 + 2\mathbb{Z}$.

An ***empty set*** is a set that does not contain any element. Both

$$\{x \in \mathbb{R} \mid x > 0 \text{ and } x < 0\} \qquad \text{and} \qquad \{x \in \mathbb{R} \mid x^2 < 0\}$$

are examples of empty sets. The second example illustrates a typical application of an empty set. It provides a convenient way of declaring that a problem has no solution: we say that the solution set is an empty set. We denote an empty set with the notation $\emptyset$ or $\{\ \}$. For example, can you explain why $(3, 3) = \emptyset$?

Hands-On Exercise 4.1.9 What does the notation $[7, 7]$ mean? How would you describe the sets $(7, 7)$, $(7, 7]$ and $[7, 7)$?

△

Example 4.1.7 Determine which of these statements are true.

$$\begin{aligned}
\{x \in \mathbb{R} \mid (x^2 + 2)(x^2 + 3) = 0\} &= \emptyset, \\
\{x \in \mathbb{Z} \mid (x^2 - 2)(x^2 + 3) = 0\} &= \emptyset, \\
\{x \in \mathbb{R} \mid (x^2 - 2)(x^2 + 3) = 0\} &= \emptyset, \\
\{x \in \mathbb{R} \mid (x^2 - 2)(x^2 + 3) \geq 0\} &= \emptyset.
\end{aligned}$$

Solution: The answers are: true, true, false, and false, respectively. ▲

Example 4.1.8 When we write $3, 4, 5, \ldots, n$, we are referring to a list of integers between 3 and n, inclusive. It is understood that $n \geq 3$. Consequently, the set

$$\{3, 4, 5, \ldots, n\}$$

is empty when $n = 2$. ▲

Two sets A and B are said to be ***equal*** if they contain the same collection of elements. More rigorously, we define

$$A = B \Leftrightarrow \forall x\, (x \in A \Leftrightarrow x \in B).$$

Since the elements of a set can themselves be sets, exercise caution and use proper notation when you compare the contents of two sets.

Example 4.1.9 Explain why $\big\{0, \{1\}\big\} \neq \{0, 1\}$.

A set may contain another set as an element.

Solution: The set $\big\{0, \{1\}\big\}$ consists of two elements: the integer 0 and the set $\{1\}$. The set $\{0, 1\}$ also consists of two elements, both of them integers; namely, 0 and 1.

You may find the following analogy helpful. Imagine a set being a box. You open a box to look at its contents. The box itself can be compared to the curly braces $\{$ and $\}$. What it holds is exactly what we call the elements of the set it represents. The contents of the two sets $\big\{0, \{1\}\big\}$ and $\{0, 1\}$ are depicted in the boxes shown in Figure 4.1.

When you open the first box, you find two items. One of them is the number 0; the other is another box that contains the number 1. The second box also contains two items that are both numbers. What you find in these two boxes is not the same. Hence, the sets they represent are different. ▲

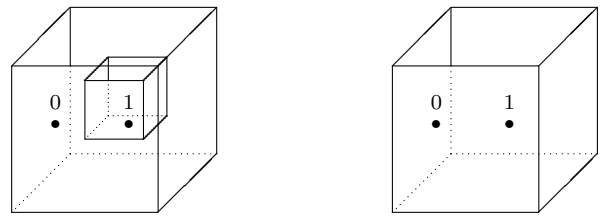


Figure 4.1: The two sets $\big\{0, \{1\}\big\}$ and $\{0, 1\}$.

Hands-On Exercise 4.1.10 Name some differences between the sets $\big\{0, \{1\}\big\}$ and $\big\{\{0\}, \{1\}\big\}$.

△

Example 4.1.10 True or false: $\mathbb{Z} = \big\{\{\ldots, -3, -2, -1\}, 0, \{1, 2, 3, \ldots\}\big\}$?

Solution: The set on the left is $\mathbb{Z}$, and

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

Pay attention to the details.

It is an infinite set. The set on the right consists of only three elements:

(i) the set $\{\ldots, -3, -2, -1\}$, which is the set of negative integers,
(ii) the integer 0, and
(iii) the set $\{1, 2, 3, \ldots\}$, which is the set of positive integers.

Hence, they are not equal. Notice that

$$\mathbb{Z} \neq \big\{\{\ldots, -3, -2, -1\}, \{0\}, \{1, 2, 3, \ldots\}\big\}$$

either, because the set on the right is a set of three sets, while the set on the left is a set of integers. One has three elements; the other has infinitely many elements. ▲

To reduce confusion, instead of saying a set of sets, we could say a ***collection of sets*** or a ***family of sets***. For example,

$$\big\{\{1, 3, 5, \ldots, \}, \{2, 4, 6, \ldots\}\big\}$$

is a family of two sets, one of which is the set of positive odd integers; the other is the set of positive even integers.

$|A|$ is nonnegative.

Definition. A set is said to be ***finite*** if it has a finite number of elements. The number of elements in a finite set A is called its ***cardinality***, and is denoted by $|A|$. Hence, $|A|$ is always nonnegative. If A is an infinite set, some authors would write $|A| = \infty$. ♢

Example 4.1.11 While it is trivial that $|\{1, 4, 7, 8\}| = 4$, and $|\{0, 1\}| = 2$, it may not be obvious that

$$\big|\{0, \{1\}\}\big| = 2,$$

and

$$\big|\big\{\{\ldots, -3, -2, -1\}, 0, \{1, 2, 3, \ldots\}\big\}\big| = 3.$$

What matters is the number of entries in a set, which can be compared to how many items you can find when you open a box. Here is another example:

$$|\{x \in \mathbb{R} \mid x^2 = 9\}| = 2$$

because the equation $x^2 = 9$ has two real solutions. What is $|\{x \in \mathbb{N} \mid x^2 = 9\}|$? ▲

Hands-On Exercise 4.1.11 Determine these cardinalities:

(a) $|\{x \in \mathbb{Z} \mid x^2 - 7x - 6 = 0\}|$

(b) $|\{x \in \mathbb{R} \mid x^2 - x - 12 < 0\}|$

(c) $|\{x \in \mathbb{Z} \mid x \text{ is prime and } x \text{ is even}\}|$

Recall that your answers should be nonnegative. △

Hands-On Exercise 4.1.12 Explain why it is incorrect to say $|\emptyset| = \emptyset$. In fact, it is nonsense to say $|\emptyset| = \emptyset$. Explain. What should be the value of $|\emptyset|$?

$|\emptyset|$ is an integer.

△

We close this section with an important remark about sets. It follows from the definition of equality of sets that we do not count repeated elements as separate elements. For example, suppose a small student club has three officers:

chair:	Mary,
vice chair:	John,
secretary:	John;

and let A represent the set of its officers, and B the set of positions in its executive board, then $|A| = 2$ and $|B| = 3$, because

$$A = \{\text{Mary}, \text{John}\},$$

and

$$B = \{\text{chair}, \text{vice chair}, \text{secretary}\}.$$

Example 4.1.12 Find the errors in the following statement:

$$|\{-2, 2\}| = \{\,|-2|, |2|\} = \{2\} = 2,$$

and correct them.

Solution: This statement contains several errors. The first mistake is assuming that we can distribute the "absolute value" symbols $|\quad|$ over the contents of a set:

$$|\{-2, 2\}| \neq \{\,|-2|, |2|\}.$$

After all, the two vertical bars do not mean absolute value in this case. Instead, it means the cardinality of the set $\{-2, 2\}$. Hence, $|\{-2, 2\}| = 2$.

The second equality $\{\,|-2|, |2|\} = \{2\}$ is correct. After taking absolute values, both entries become 2. However, we do not write $\{|-2|, |2|\} = \{2, 2\}$, because a set should not contain repetition. Therefore, it is correct to say $\{\,|-2|, |2|\} = \{2\}$.

The last equality $\{2\} = 2$ is wrong. We cannot compare a set to a number. Imagine the set $\{2\}$ as a box containing only one object, and that object is the number 2. In contrast, 2 on the right-hand side is left in the open air without any containment. It is clear that $\{2\} \neq 2$.

The entire statement contains multiple mistakes; some of them are syntactical errors while some are conceptual. Nevertheless, we do have $|\{-2, 2\}| = 2$. Although the final answer is correct, the argument used to obtain it is not. ▲

In some situations, we do want to count repeated elements as separate elements, as in $S = \{1, 2, 2, 2, 3, 3, 4, 4\}$. We call such a collection a ***multiset*** instead of an ordinary set. In this case, $|S| = 8$.

Summary and Review

- A set is a collection of objects (without repetitions).
- To describe a set, either list all its elements explicitly, or use a descriptive method.
- Intervals are sets of real numbers.
- The elements in a set can be any type of object, including sets.
- We can even have a set containing dissimilar elements. In particular, we can mix elements and sets inside a set.
- If a set A is finite, its cardinality $|A|$ is the number of elements it contains. Consequently, $|A|$ is always nonnegative.

Exercises 4.1

1. Write each of these sets by listing its elements explicitly (that is, using the roster method).

   (a) $\{n \in \mathbb{Z} \mid -6 < n < 4\}$ (b) $\{n \in \mathbb{N} \mid -6 < n < 4\}$
   (c) $\{x \in \mathbb{Q} \mid x^3 - x^2 - 6x = 0\}$ (d) $\{x \in \mathbb{Q} \mid x^4 - 11x^2 + 18 = 0\}$.

2. Use the roster method to describe these sets:

   (a) $\{x \in \mathbb{N} \mid x < 20 \text{ and } x \text{ is a multiple of } 3\}$
   (b) $\{x \in \mathbb{Z} \mid |x| < 20 \text{ and } x \text{ is a multiple of } 3 \text{ or a multiple of } 5\}$
   (c) $\{x \in \mathbb{Z} \mid |x| < 20 \text{ and } x \text{ is a multiple of } 3 \text{ and a multiple of } 5\}$
   (d) $\{x \in \mathbb{N} \mid x < 20 \text{ and } x \text{ is a multiple of } 3 \text{ but not a multiple of } 5\}$

3. Write each of the following sets in the form $\{n \in \mathbb{Z} \mid p(n)\}$ with a logical statement $p(n)$ describing the property of n.

   (a) $\{\ldots, -3, -2, -1\}$ (b) $\{\ldots, -27, -8, -1, 0, 1, 8, 27, \ldots\}$
   (c) $\{0, 1, 4, 9, 16, \ldots\}$ (d) $\{\ldots, -15, -10, -5, 0, 5, 10, 15, \ldots\}$
   (e) $\{0, 4, 8, 12, \ldots\}$ (f) $\{\ldots, -14, -8, -2, 4, 10, 16, \ldots\}$

4. Repeat Problem 3, but write the sets in the form $\{f(n) \mid n \in S\}$, where $f(n)$ is a formula that describes the pattern of the elements, and S is an appropriate set of numbers.

5. Whenever possible, express the sets in Problem 3 in the form S^+, S^-, bS, or $a + bS$ for some appropriate set S.

6. Determine whether the following sets are empty, finite sets, or infinite sets:

   (a) $\{2n + 1 \mid n \in \mathbb{N}\}$ (b) $\{x \in \mathbb{R} \mid x^2 < 0\}$
   (c) $\{x \in \mathbb{Q} \mid x \geq 0 \text{ and } x \leq 0\}$ (d) $\{x \in \mathbb{N} \mid x < 0 \text{ or } x > 0\}$

7. Write each of these sets in the interval notation.

   (a) $\{x \in \mathbb{R} \mid -4 < x < 7\}$
   (b) $\{x \in \mathbb{R} \mid -4 < x \leq 7\}$
   (c) $\{x \in \mathbb{R}^+ \mid -4 < x \leq 7\}$
   (d) $\{x \in \mathbb{R} \mid -4 < x\}$
   (e) $\{x \in \mathbb{R} \mid x \leq 6\}$
   (f) $\{x \in \mathbb{R}^- \mid 0 \leq x \leq 6\}$

8. Is $[-\infty, \infty]$ a legitimate or correct notation? Explain.

9. Evaluate the following expressions.

   (a) $|\{x \in \mathbb{Z} \mid -4 < x < 7\}|$
   (b) $|\{x \in \mathbb{Z} \mid -4 < x \leq 7\}|$
   (c) $|\{x \in \mathbb{N} \mid -4 < x \leq 7\}|$
   (d) $|\{x \in \mathbb{R} \mid x^4 - 2x^3 - 35x^2 = 0\}|$
   (e) $|\{-3, -2, 2, 3\}|$
   (f) $|\{x \in \mathbb{Q} \mid x^2 = 3\}|$

10. Determine which of the following statements are true, and which are false.

    (a) $a \in \{a\}$
    (b) $\{3, 5\} = \{5, 3\}$
    (c) $\emptyset \in \emptyset$
    (d) $\emptyset = \{\emptyset\}$
    (e) $\{\ \} = \emptyset$
    (f) $\emptyset \in \{\emptyset\}$

11. Determine which of the following statements are true, and which are false.

    (a) $2 \in (2, 7)$
    (b) $\sqrt{2} \in (1, 3)$
    (c) $\left(\sqrt{5}\right)^2 \in \mathbb{Q}$
    (d) $-5 \in \mathbb{N}$

12. Give examples of sets A, B and C such that:

    (a) $A \in B$ and $B \in C$, and $A \notin C$
    (b) $A \in B$ and $B \in C$, and $A \in C$

13. Determine whether the following statements are correct or incorrect *syntactically*. For those that are syntactically correct, determine their truth values; for those that are syntactically incorrect, suggest ways to fix them.

    (a) $(3, 7] = 3 < x \leq 7$.
    (b) $\{x \in \mathbb{R} \mid x^2 < 0\} \equiv \emptyset$.

14. Determine whether the following statements are correct or incorrect *syntactically*. For those that are syntactically correct, determine their truth values; for those that are syntactically incorrect, suggest ways to fix them.

    (a) $\frac{7}{4} \in [2, \sqrt{7})$.
    (b) There does not exist x such that $x \in \mathbb{R}^+$ and $\mathbb{R}^-$.
    (c) If $(0, \infty)$, then x is positive.

4.2 Subsets and Power Sets

We usually consider sets containing elements of similar types. The collection of all the objects under consideration is called the ***universal set***, and is denoted $\mathcal{U}$. For example, for numbers, the universal set is $\mathbb{R}$.

Example 4.2.1 Venn diagrams are useful in demonstrating set relationship. Let

$$\begin{aligned}
\mathcal{U} &= \text{set of geometric figures,}\\
S &= \text{set of squares,}\\
P &= \text{set of parallelogram,}\\
R &= \text{set of rhombuses,}\\
L &= \text{set of rectangles,}\\
C &= \text{set of circles.}
\end{aligned}$$

Their relationship is displayed in Figure 4.2. ▲

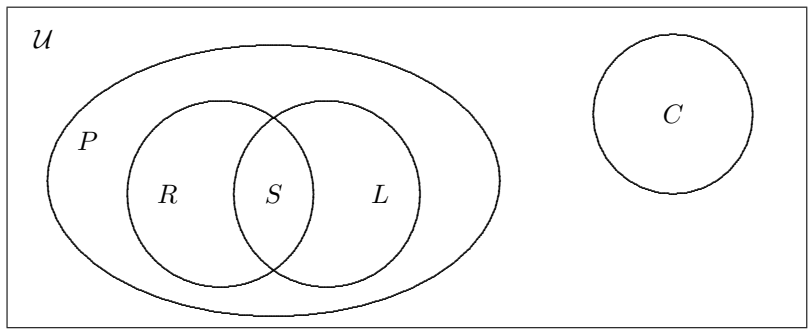


Figure 4.2: The relationship among various sets of geometric figures.

The pictorial representation in Figure 4.2 is called a ***Venn diagram***. We use a rectangle to represent the universal set, and circles or ovals to represent the sets inside the universal set. The relative positions of these circles and ovals indicate the relationship of the respective sets. For example, having R, S, and L inside P means that rhombuses, squares, and rectangles are parallelograms. In contrast, circles are incomparable to parallelograms.

A set A is a subset of another set B, denoted by $A \subseteq B$, if every element of A is also an element of B. See Figure 4.3. We also call B a ***superset*** of A, and write $B \supseteq A$, which is similar to $y \geq x$.

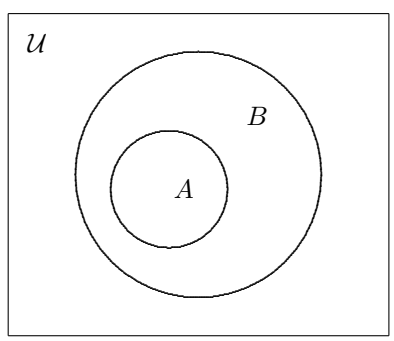


Figure 4.3: The Venn diagram for $A \subseteq B$.

Example 4.2.2 It is clear that $\mathbb{N} \subseteq \mathbb{Z}$ and $\mathbb{Z} \subseteq \mathbb{R}$. We can nest these two relationships into one, and write $\mathbb{N} \subseteq \mathbb{Z} \subseteq \mathbb{R}$. More generally, we have

$$\mathbb{N} \subseteq \mathbb{Z} \subseteq \mathbb{Q} \subseteq \mathbb{R}.$$

Compare this to $x \leq y \leq z \leq w$. We shall discover many similarities between $\subseteq$ and $\leq$. ▲

Example 4.2.3 It is obvious that

$$\{1, 2, 7\} \subseteq \{1, 2, 3, 6, 7, 9\}$$

because all three elements 1, 2, and 7 from the set on the left also appear as elements in the set on the right. Meanwhile,

$$\{1, 2, 7\} \nsubseteq \{1, 2, 3, 6, 8, 9\}$$

because 7 belongs to the first set but not the second. ▲

Example 4.2.4 The following statements are true:

(a) $\{1, 2, 3\} \subseteq \mathbb{N}$.
(b) $\{x \in \mathbb{R} \mid x^2 = 1\} \subseteq \mathbb{Z}$.

Be sure you can explain clearly why these subset relationships hold. ▲

Hands-On Exercise 4.2.1 Are these statements true or false?

(a) $\{-1, 2\} \nsubseteq \mathbb{N}$, and $\{-1, 2\} \subseteq \mathbb{Z}$.

(b) $\{x \in \mathbb{Z} \mid x^2 \leq 1\} \subseteq \mathbb{R}$. △

Example 4.2.5 Do not assume that if $A \nsubseteq B$ then we must have $B \subseteq A$. For instance, if $A = \{1, 5, 7\}$ and $B = \{3, 8\}$, then $A \nsubseteq B$; but we also have $B \nsubseteq A$. ▲

The last example demonstrates that $A \nsubseteq B$ is more complicated than just changing the subset notation like we do with inequalities. We need a more precise definition of the subset relationship:

The definition of $A \subseteq B$.

$$\boxed{A \subseteq B \Leftrightarrow \forall x \in \mathcal{U}\,(x \in A \Rightarrow x \in B).}$$

It follows that

$$A \nsubseteq B \Leftrightarrow \exists x \in \mathcal{U}\,(x \in A \wedge x \notin B).$$

Hence, to show that A is not a subset of B, we need to find an element x that belongs to A but not B. There are three possibilities; their Venn diagrams are depicted in Figure 4.4.

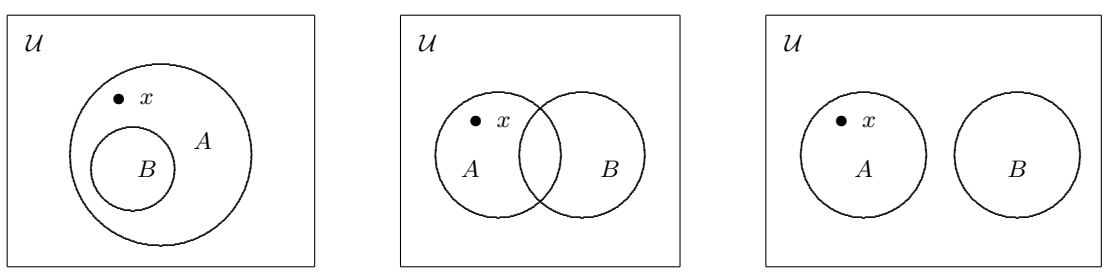


Figure 4.4: Three cases of $A \nsubseteq B$.

Example 4.2.6 We have $[3, 6] \subseteq [2, 7)$, and $[3, 6] \nsubseteq [4, 7)$. We also have $(3, 4) \subseteq [3, 4]$. ▲

Hands-On Exercise 4.2.2 True or false: $[3, 4) \subseteq (3, 4)$? Explain.

△

With the notion of universal set, we can now refine the definition for set equality:

$$\boxed{A = B \Leftrightarrow \forall x \in \mathcal{U}\,(x \in A \Leftrightarrow x \in B).}$$

Logically, $x \in A \Leftrightarrow x \in B$ is equivalent to

$$(x \in A \Rightarrow x \in B) \wedge (x \in B \Rightarrow x \in A).$$

Therefore, we can also define the equality of sets via subset relationship:

$$\boxed{A = B \Leftrightarrow (A \subseteq B) \wedge (B \subseteq A),}$$

which can be compared to

$$x = y \Leftrightarrow (x \leq y) \wedge (y \leq x)$$

for real numbers x and y.

This new definition of set equality suggests that in order to prove that $A = B$, we could use this two-step argument:

A way to prove $A = B$.

1. Show that $A \subseteq B$.
2. Show that $B \subseteq A$.

This technique is useful when it is impossible or impractical to list the elements of A and B for comparison. This is particularly true when A and B are defined abstractly. We will apply this technique in the coming sections.

The two relationship $\subseteq$ and $\leq$ share many common properties. The ***transitive property*** is another example.

Compare this to $(x \leq y) \wedge (y \leq z) \Rightarrow x \leq z$.

Theorem 4.2.1 *Let A, B, and C be sets. If $A \subseteq B$ and $B \subseteq C$, then $A \subseteq C$.*

Discussion. The theorem statement is in the form of an implication. To prove $p \Rightarrow q$, we start with the assumption p, and use it to show that q must also be true. In this case, these two steps become

(i) Assume that $A \subseteq B$ and $B \subseteq C$.
(ii) Show that $A \subseteq C$.

How can we prove that $A \subseteq C$? We know that $A \subseteq C$ means

$$\forall x \in \mathcal{U}\,(x \in A \Rightarrow x \in C).$$

So we have to start with $x \in A$, and attempt to show that $x \in C$ as well. How can we show that $x \in C$? We need to use the assumption $A \subseteq B$ and $B \subseteq C$. ♢

Proof: Assume $A \subseteq B$ and $B \subseteq C$. Let $x \in A$. Since $A \subseteq B$, we also have $x \in B$. Likewise, $B \subseteq C$ implies that $x \in C$. Since every element x in A is also an element of C, we conclude that $A \subseteq C$. ■

The proof relies on the definition of the subset relationship. Many proofs in mathematics are rather simple if you know the underlying definitions.

Example 4.2.7 Prove that $x \in A \Leftrightarrow \{x\} \subseteq A$, for any element $x \in \mathcal{U}$.

Discussion. We call $p \Leftrightarrow q$ a biconditional statement because it consists of two implications $p \Rightarrow q$ and $p \Leftarrow q$. Hence, we need to prove it in two steps:

1. Show that $p \Rightarrow q$.
2. Show that $q \Rightarrow p$.

We call these two implications the ***necessity*** and ***sufficiency*** of the biconditional statement, and denote them ($\Rightarrow$) and ($\Leftarrow$), respectively. In this problem,

- ($\Rightarrow$) means "$x \in A \Rightarrow \{x\} \subseteq A$".
- ($\Leftarrow$) means "$\{x\} \subseteq A \Rightarrow x \in A$".

This is how the proof may look:

($\Rightarrow$)	Assume $x \in A$.	...	Therefore $\{x\} \subseteq A$.
($\Leftarrow$)	Assume $\{x\} \subseteq A$.	...	Therefore $x \in A$.

We now proceed to finish the proof. ♢

Solution: ($\Rightarrow$) Assume $x \in A$. The set $\{x\}$ contains only one element x, which is also an element of A. Thus, every element of $\{x\}$ is also an element of A. By definition, $\{x\} \subseteq A$.

($\Leftarrow$) Assume $\{x\} \subseteq A$. The definition of the subset relationship asserts that every element of $\{x\}$ is also an element of A. In particular, x is an element of $\{x\}$, so it is also an element of A. Thus, $x \in A$. ▲

Definition. The set A is a ***proper subset*** of B, denoted $A \subsetneq B$ or $A \subset B$, if A is a subset of B, and $A \neq B$. Symbolically, $A \subset B \Leftrightarrow (A \subseteq B) \wedge (A \neq B)$. Equivalently,

$$A \subset B \Leftrightarrow (A \subseteq B) \wedge \exists x \in \mathcal{U}\,(x \in B \wedge x \notin A).$$

See the Venn diagram in Figure 4.5. ♢

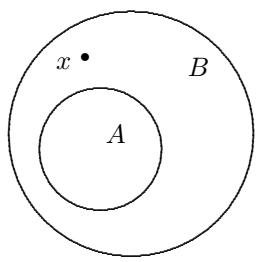


Figure 4.5: The definition of a proper subset.

Example 4.2.8 It is clear that $[0, 5] \subset \mathbb{R}$. We also have

$$\mathbb{N} \subset \mathbb{Z} \subset \mathbb{Q} \subset \mathbb{R}.$$

Note the similarities between $\subset$ and $<$. Compare the last expression to

$$x < y < z < w.$$

Here is another similarity between $\subset$ and $<$. For numbers, $x < y$ and $y < z$ together imply that $x < z$. We call this the transitive property. In a similar fashion, for sets, if $A \subset B$ and $B \subset C$, then $A \subset C$; see Theorem 4.2.1. ▲

Hands-On Exercise 4.2.3 True or false: $(3, 4) \subset [3, 4]$? How about $(3, 4) \subset (3, 4]$?

△

Theorem 4.2.2 *For any set A, we have $\emptyset \subseteq A$ and $A \subseteq A$. In particular, $\emptyset \subseteq \emptyset$.*

Proof: Since every element of A also appears in A, it follows immediately that $A \subseteq A$. To show that $\emptyset \subseteq A$, we need to verify the implication

$$x \in \emptyset \Rightarrow x \in A$$

An implication is always true if its hypothesis is false.

for any arbitrary $x \in \mathcal{U}$. Since $\emptyset$ is empty, $x \in \emptyset$ is always false; hence, the implication is always true. Consequently, $\emptyset \subseteq A$ for any set A. In particular, when $A = \emptyset$, we obtain $\emptyset \subseteq \emptyset$. ■

Example 4.2.9 Determine the truth values of these expressions.

(a) $\emptyset \in \emptyset$ (b) $1 \subseteq \{1\}$ (c) $\emptyset \in \{\emptyset\}$

Solution: (a) By definition, an empty set contains no element. Consequently, the statement $\emptyset \in \emptyset$ is false.

(b) A subset relation only exists between two sets. To the left of the symbol $\subseteq$, we have only a number, which is not a set. Hence, the statement is false. In fact, this expression is syntactically incorrect.

(c) The set $\{\emptyset\}$ contains one element, which happens to be an empty set. Compare this to an empty box inside another box. The outer box is described by the pair of set brackets $\{\ldots\}$, and the (empty) box inside is $\emptyset$. It follows that $\emptyset \in \{\emptyset\}$ is a true statement. ▲

Hands-On Exercise 4.2.4 Determine the truth values of these expressions.

(a) $\emptyset \subseteq \{\emptyset\}$ (b) $\{1\} \subseteq \big\{1, \{1, 2\}\big\}$ (c) $\{1\} \subseteq \big\{\{1\}, \{1, 2\}\big\}$

△

Definition. The set of all subsets of A is called the ***power set*** of A, denoted $\wp(A)$. ◇

Every element in $\wp(A)$ is a stand-alone set; more precisely, it is a subset of A.

Since a power set itself is a set, we need to use a pair of left and right curly braces (set brackets) to enclose all its elements. Its elements are themselves sets, each of which requires its own pair of left and right curly braces. Consequently, we need at least two levels of set brackets to describe a power set.

Example 4.2.10 Let $A = \{1, 2\}$ and $B = \{1\}$. The subsets of A are $\emptyset$, $\{1\}$, $\{2\}$ and $\{1, 2\}$. Therefore,

$$\wp(A) = \big\{\emptyset, \{1\}, \{2\}, \{1, 2\}\big\}.$$

In a similar manner, we find

$$\wp(B) = \big\{\emptyset, \{1\}\big\}.$$

We can write directly

$$\wp(\{1, 2\}) = \big\{\emptyset, \{1\}, \{2\}, \{1, 2\}\big\}, \qquad \text{and} \qquad \wp(\{1\}) = \big\{\emptyset, \{1\}\big\}$$

without introducing letters to represent the sets involved. ▲

Hands-On Exercise 4.2.5 Let us evaluate $\wp(\{1,2,3,4\})$. To ensure that no subset is missed, we list these subsets according to their sizes. Since $\emptyset$ is the subset of any set, $\emptyset$ is always an element in the power set. This is the subset of size 0. Next, list the singleton subsets (subsets with only one element). Then the doubleton subsets, and so forth. Complete the following table.

Since $\emptyset \subseteq A$ for any set A, we always have $\emptyset \in \wp(A)$.

size	subsets
0	$\emptyset$
1	$\{1\}, \{2\}, \ldots$
2	$\{1,2\}, \{1,3\}, \ldots$
3	$\{1,2,3\}, \ldots$
4	$\ldots$

Since $A \subseteq A$ for any set A, the power set $\wp(A)$ always contains A itself. As a result, the last subset in the list should be A itself.

We are now ready to put them together to form the power set. All you need is to put all the subsets inside a pair of bigger curly braces (a power set is itself a set; hence, it needs a pair of curly braces in its description). Put your final answer in the space below.

Check to make sure that the left and right braces match perfectly. △

Example 4.2.11 Since A is a subset of A, it belongs to $\wp(A)$. Nonetheless, it is improper to say $A \subseteq \wp(A)$. Can you explain why? What should be the correct notation?

Solution: The power set $\wp(A)$ is the collection of all the subsets of A. Thus, the elements in $\wp(A)$ are subsets of A. One of these subsets is the set A itself. Hence, A itself appears as an *element* in $\wp(A)$, and we write $A \in \wp(A)$ to describe this *membership*.

This is different from saying that $A \subseteq \wp(A)$. In order to have the *subset* relationship $A \subseteq \wp(A)$, every element in A must also appear as an element in $\wp(A)$. The elements of $\wp(A)$ are sets (they are subsets of A, and subsets are sets). An element of A is not the same as a subset of A. Therefore, although $A \subseteq \wp(A)$ is syntactically correct, its truth value is false. ▲

Hands-On Exercise 4.2.6 Explain the difference between $\emptyset$ and $\{\emptyset\}$. How many elements are there in $\emptyset$ and $\{\emptyset\}$? Is it true that $\wp(\emptyset) = \{\emptyset\}$?

△

Theorem 4.2.3 *If A is an n-element set, then $\wp(A)$ has 2^n elements. In other words, an n-element set has 2^n distinct subsets.*

Proof: How many subsets of A can we construct? To form a subset, we go through each of the n elements and ask ourselves if we want to include this particular element or not. Since there are two choices (yes or no) for each of the n elements in A, we have found $\underbrace{2 \cdot 2 \cdot \cdots \cdot 2}_{n \text{ times}} = 2^n$ subsets. ■

Hands-On Exercise 4.2.7 How many elements are there in $\wp(\{\alpha, \beta, \gamma\})$? What are they?

△

Hands-On Exercise 4.2.8 What is the cardinality of $\emptyset$? How about $\wp(\emptyset)$? Describe $\wp(\emptyset)$.

△

Hands-On Exercise 4.2.9 Is it correct to write $|\wp(A)| = 2^{|A|}$? How about $|\wp(A)| = 2^A$? Explain.

△

Example 4.2.12 When a set contains sets as elements, its power set could become rather complicated. Here are two examples.

$$\begin{aligned} \wp(\{\{a\},\{1\}\}) &= \Big\{\emptyset, \{\{a\}\}, \{\{1\}\}, \{\{a\},\{1\}\}\Big\}, \\ \wp(\{\emptyset,\{1\}\}) &= \Big\{\emptyset, \{\emptyset\}, \{\{1\}\}, \{\emptyset,\{1\}\}\Big\}. \end{aligned}$$

Be sure you understand the notations used in these examples. In particular, examine the number of levels of set brackets used in each example. ▲

Summary and Review

- A set S is a subset of another set T if and only if every element in S can be found in T.
- In symbols, $S \subseteq T \Leftrightarrow \forall x \in \mathcal{U}\,(x \in S \Rightarrow x \in T)$.
- Consequently, to show that $S \subseteq T$, we have to start with an arbitrary element x in S, and show that x also belongs to T.
- The definition of subset relationship implies that for any set S, we always have $\emptyset \subseteq S$ and $S \subseteq S$.
- The power set of a set S, denoted $\wp(S)$, contains all the subsets of S.
- If $|S| = n$, then $|\wp(S)| = 2^n$. Hence, an n-element set has 2^n subsets.
- To construct $\wp(S)$, list the subsets of S according to their sizes. Be sure to use a pair of curly braces for each subset, and enclose all of them within a pair of outer curly braces.

Exercises 4.2

1. Determine which of the following statements are true and which are false.

   (a) $\{1,2,3\} \subseteq \{0,1,2,3,4\}$ (b) $\{1,2,3\} \subseteq \mathbb{N}$
   (c) $\{1,2\} \subset [1,2]$ (d) $[2,4] \subseteq (0,6)$
   (e) $[2,4) \subset [2,4]$ (f) $[2,4) \subseteq (2,4]$

2. Determine which of the following statements are true and which are false.

   (a) $a \subseteq \{a\}$ (b) $\{a\} \subseteq \{a,b\}$
   (c) $\emptyset \subseteq \emptyset$ (d) $\emptyset \subseteq \{\emptyset\}$
   (e) $\emptyset \subset \{\emptyset\}$ (f) $\{a\} \subseteq \wp(\{\{a\},\{b\}\})$

3. Explain why $\mathbb{Z} \subseteq \mathbb{Q}$. In particular, explain how to express an integer as a rational number.
4. True or false: $\mathbb{N} \subseteq 6\mathbb{N}$? Explain.
5. If $A \subseteq B$, $B \subseteq C$, and $C \subseteq D$, is it true that $A \subseteq D$? What do you call this property?
6. Determine whether the following statements are true or false:
   (a) The empty set $\emptyset$ is a subset of $\{1, 2, 3\}$.
   (b) If $A = \{1, 2, 3\}$, then $\{1\}$ is a subset of $\wp(A)$.
7. Find the power set of the following sets.
   (a) $\{a, b\}$ (b) $\{4, 7\}$ (c) $\{x, y, z, w\}$
   (d) $\big\{\{a\}\big\}$ (e) $\big\{a, \{b\}\big\}$ (f) $\big\{\{x\}, \{y\}\big\}$
8. Evaluate the following sets.
   (a) $\wp(\{\emptyset\})$ (b) $\wp(\wp(\{a, b\}))$ (c) $\wp(\wp(\wp(\emptyset)))$
9. We have learned that $A \subseteq A$ for any set A. Then, should we write $A \in \wp(A)$ or $A \subseteq \wp(A)$? Explain.
10. Prove that $X \in \wp(A)$ if and only if $X \subseteq A$.
11. Determine which of the following statements are true, and which are false. Explain!
    (a) $\{a\} \in \{a, b, c\}$ (b) $\{a\} \subseteq \{\{a\}, b, c\}$ (c) $\{a\} \in \wp(\{\{a\}, b, c\})$
12. Determine which of the following statements are true, and which are false. Explain!
    (a) $\{a\} \subseteq \{a, b, c\}$ (b) $\{a\} \subseteq \{\{a, b\}, c\}$ (c) $\{a\} \subseteq \wp(\{\{a\}, b, c\})$

4.3 Unions and Intersections

We can form a new set from existing sets by carrying out a set operation.

Definition. Given two sets A and B, define their ***intersection*** to be the set

$$\boxed{A \cap B = \{x \in \mathcal{U} \mid x \in A \wedge x \in B\}.}$$

Note the similarity between the symbols $\cap$ and $\wedge$.

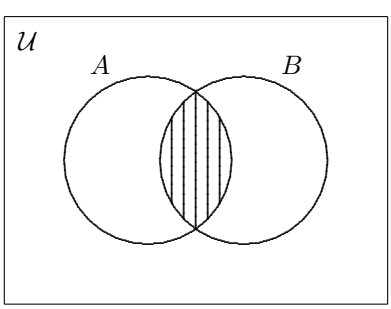
Loosely speaking, $A \cap B$ contains elements common to both A and B. $\diamondsuit$

Definition. The ***union*** of A and B is defined as

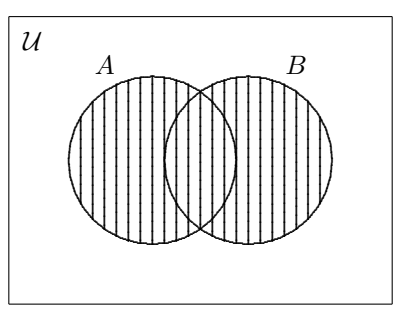
$$\boxed{A \cup B = \{x \in \mathcal{U} \mid x \in A \vee x \in B\}.}$$

Note the similarity between the symbols $\cup$ and $\vee$.

Thus $A \cup B$ is, as the name suggests, the set combining all the elements from A and B. $\diamondsuit$



$A \cap B$



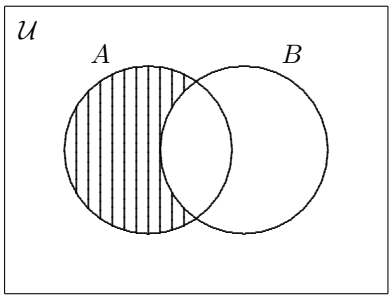
$A \cup B$

$x \in A - B$ means x belongs to A but not to B.

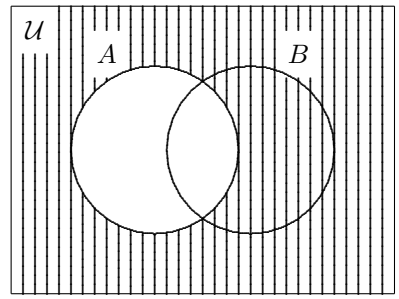
Definition. The ***set difference*** $A - B$, sometimes written as $A \setminus B$, is defined as

$$\boxed{A - B = \{x \in \mathcal{U} \mid x \in A \wedge x \notin B\}.}$$

In words, $A - B$ contains elements that can only be found in A but not in B. Operationally speaking, $A - B$ is the set obtained from A by removing the elements that also belong to B. Therefore, the set difference $A - B$ is also called the ***relative complement*** of B in A. In particular, $\mathcal{U} - A$ is called the ***complement*** of A, and is denoted by $\overline{A}$, A' or A^c. ♢



$A - B$



$\overline{A}$

Remark. We would like to remind the readers that it is not uncommon among authors to adopt different notations for the same mathematical concept. Likewise, the same notation could mean something different in another textbook or even another branch of mathematics. It is important to develop the habit of examining the context and making sure that you understand the meaning of the notations when you start reading a mathematical exposition. ♢

Example 4.3.1 Let $\mathcal{U} = \{1, 2, 3, 4, 5\}$, $A = \{1, 2, 3\}$, and $B = \{3, 4\}$. Find $A \cap B$, $A \cup B$, $A - B$, $B - A$, $\overline{A}$, and $\overline{B}$.

Solution: We have

$$\begin{aligned} A \cap B &= \{3\}, \\ A \cup B &= \{1, 2, 3, 4\}, \\ A - B &= \{1, 2\}, \\ B - A &= \{4\}. \end{aligned}$$

We also find $\overline{A} = \{4, 5\}$, and $\overline{B} = \{1, 2, 5\}$. ▲

Hands-On Exercise 4.3.1 Let $\mathcal{U} = \{\text{John}, \text{Mary}, \text{Dave}, \text{Lucy}, \text{Peter}, \text{Larry}\}$,

$$A = \{\text{John}, \text{Mary}, \text{Dave}\}, \qquad \text{and} \qquad B = \{\text{John}, \text{Larry}, \text{Lucy}\}.$$

Find $A \cap B$, $A \cup B$, $A - B$, $B - A$, $\overline{A}$, and $\overline{B}$.

△

Hands-On Exercise 4.3.2 If $A \subseteq B$, what would be $A - B$?

△

Example 4.3.2 The set of integers can be written as the

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

Can we replace $\{0\}$ with 0? Explain. ▲

Hands-On Exercise 4.3.3 Explain why the following expressions are syntactically incorrect.

(a) $\mathbb{Z} = \{-1, -2, -3, \ldots\} \cup\ 0\ \cup \{1, 2, 3, \ldots\}$.
(b) $\mathbb{Z} = \ldots, -3, -2, -1\ \cup\ 0\ \cup\ 1, 2, 3, \ldots$
(c) $\mathbb{Z} = \ldots, -3, -2, -1\ +\ 0\ +\ 1, 2, 3, \ldots$
(d) $\mathbb{Z} = \mathbb{Z}^-\ \cup\ 0\ \cup\ \mathbb{Z}^+$

How would you fix the errors in these expressions?

△

Example 4.3.3 For any set A, what are $A \cap \emptyset$, $A \cup \emptyset$, $A - \emptyset$, $\emptyset - A$ and $\overline{\overline{A}}$?

Solution: It is clear that

$$A \cap \emptyset = \emptyset, \qquad A \cup \emptyset = A, \qquad \text{and} \qquad A - \emptyset = A.$$

From the definition of set difference, we find $\emptyset - A = \emptyset$. Finally, $\overline{\overline{A}} = A$. ▲

Example 4.3.4 Write, in interval notation, $[5, 8) \cup (6, 9]$ and $[5, 8) \cap (6, 9]$.

Solution: The answers are

$$[5, 8) \cup (6, 9] = [5, 9], \qquad \text{and} \qquad [5, 8) \cap (6, 9] = (6, 8).$$

They are obtained by comparing the location of the two intervals on the real number line. ▲

Hands-On Exercise 4.3.4 Write, in interval notation, $(0, 3) \cup [-1, 2)$ and $(0, 3) \cap [-1, 2)$.

△

Example 4.3.5 We are now able to describe the following set

$$\{x \in \mathbb{R} \mid (x < 5) \vee (x > 7)\}$$

in the interval notation. It can be written as either $(-\infty, 5) \cup (7, \infty)$ or, using complement, $\mathbb{R} - [5, 7]$. Consequently, saying $x \notin [5, 7]$ is the same as saying $x \in (-\infty, 5) \cup (7, \infty)$, or equivalently, $x \in \mathbb{R} - [5, 7]$. ▲

Theorem 4.3.1 *The following properties hold for any sets A, B, and C in a universal set $\mathcal{U}$.*

1. **Commutative properties**: $A \cup B = B \cup A$,
   $A \cap B = B \cap A$.

2. **Associative properties**: $(A \cup B) \cup C = A \cup (B \cup C)$,
   $(A \cap B) \cap C = A \cap (B \cap C)$.

3. **Distributive laws**: $A \cup (B \cap C) = (A \cup B) \cap (A \cup C)$,
   $A \cap (B \cup C) = (A \cap B) \cup (A \cap C)$.

4. **Idempotent laws**: $A \cup A = A$,
   $A \cap A = A$.

5. **De Morgan's laws**: $\overline{A \cup B} = \overline{A} \cap \overline{B}$,
   $\overline{A \cap B} = \overline{A} \cup \overline{B}$.

6. **Laws of the excluded middle**, *or* **inverse laws**: $A \cup \overline{A} = \mathcal{U}$,
   $A \cap \overline{A} = \emptyset$.

As an illustration, we shall prove the distributive law

$$A \cup (B \cap C) = (A \cup B) \cap (A \cup C).$$

We need to show that

$$\forall x \in \mathcal{U} \left[x \in A \cup (B \cap C) \Leftrightarrow x \in (A \cup B) \cap (A \cup C) \right].$$

Equivalently, we need to show that

$$A \cup (B \cap C) \subseteq (A \cup B) \cap (A \cup C), \qquad \text{and} \qquad (A \cup B) \cap (A \cup C) \subseteq A \cup (B \cap C).$$

Either way, we need to establish the equality in two steps.

We now present two proofs of the distributive law $A \cup (B \cap C) = (A \cup B) \cap (A \cup C)$.

Proof 1: Let $x \in A \cup (B \cap C)$. Then $x \in A$, or $x \in B \cap C$. We know that $x \in B \cap C$ implies that $x \in B$ and $x \in C$. So we have

(i) $x \in A$ or $x \in B$, and
(ii) $x \in A$ or $x \in C$;

equivalently,

(i) $x \in A \cup B$, and
(ii) $x \in A \cup C$.

Thus, $x \in (A \cup B) \cap (A \cup C)$. We have proved that $A \cup (B \cap C) \subseteq (A \cup B) \cap (A \cup C)$.

Now let $x \in (A \cup B) \cap (A \cup C)$. Then $x \in A \cup B$ and $x \in A \cup C$. From the definition of union, we find

(i) $x \in A$ or $x \in B$, and
(ii) $x \in A$ or $x \in C$.

Both conditions require $x \in A$, so we can rewrite them as

(i) $x \in A$, or
(ii) $x \in B$ and $x \in C$;

equivalently,

(i) $x \in A$, or
(ii) $x \in B \cap C$.

Thus, $x \in A \cup (B \cap C)$. This proves that $(A \cup B) \cap (A \cup C) \subseteq A \cup (B \cap C)$. Together with $A \cup (B \cap C) \subseteq (A \cup B) \cap (A \cup C)$, we conclude that $A \cup (B \cap C) = (A \cup B) \cap (A \cup C)$. ■

Below is an alternate proof. This type of argument is shorter, but is more symbolic; hence, it is more difficult to follow.

Proof 2: Since

$$\begin{aligned}
x \in A \cup (B \cap C) &\Leftrightarrow x \in A \vee x \in (B \cap C) && \text{(defn. of union)}\\
&\Leftrightarrow x \in A \vee (x \in B \wedge x \in C) && \text{(defn. of intersection)}\\
&\Leftrightarrow (x \in A \vee x \in B) \wedge (x \in A \vee x \in C) && \text{(distributive law)}\\
&\Leftrightarrow (x \in A \cup B) \wedge (x \in A \cup C) && \text{(defn. of union)}\\
&\Leftrightarrow x \in (A \cup B) \cap (A \cup C) && \text{(defn. of intersection)}
\end{aligned}$$

it follows that $A \cup (B \cap C) = (A \cup B) \cap (A \cup C)$. ■

Hands-On Exercise 4.3.5 Prove that $A \cap (B \cup C) = (A \cap B) \cup (A \cap C)$.

△

Hands-On Exercise 4.3.6 Prove that if $A \subseteq B$ and $A \subseteq C$, then $A \subseteq B \cap C$.

Discussion. Let us start with a draft. The statement we want to prove takes the form of

$$(A \subseteq B) \wedge (A \subseteq C) \Rightarrow A \subseteq B \cap C.$$

Hence, what do we assume and what do we want to prove?

Assume:

Want to Prove:

Did you put down we assume $A \subseteq B$ and $A \subseteq C$, and we want to prove $A \subseteq B \cap C$? Great! Now, what does it mean by $A \subseteq B$? How about $A \subseteq C$? What is the meaning of $A \subseteq B \cap C$?

$A \subseteq B$ means: For any $x \in \mathcal{U}$, if $x \in A$, then $x \in B$ as well.

$A \subseteq C$ means:

$A \subseteq B \cap C$ means:

How can you use the first two pieces of information to obtain what we need to establish?

Now it is time to put everything together, and polish it into a final version. Remember three things:

(i) the outline of the proof,
(ii) the reason in each step of the main argument, and
(iii) the introduction and the conclusion.

Put the complete proof in the space below.

$\triangle$

Here are two results involving complements.

Theorem 4.3.2 *For any two sets A and B, we have $A \subseteq B \Leftrightarrow \overline{B} \subseteq \overline{A}$.*

Theorem 4.3.3 (Generalized De Morgan's Laws) *For any sets A, B and C,*

$$\begin{aligned} A - (B \cup C) &= (A - B) \cap (A - C), \\ A - (B \cap C) &= (A - B) \cup (A - C), \end{aligned}$$

Summary and Review

- Memorize the definitions of intersection, union, and set difference. We rely on them to prove or derive new results.
- The intersection of two sets A and B, denoted $A \cap B$, is the set of elements common to both A and B. In symbols, $\forall x \in \mathcal{U}\ \big[x \in A \cap B \Leftrightarrow (x \in A \wedge x \in B)\big]$.
- The union of two sets A and B, denoted $A \cup B$, is the set that combines all the elements in A and B. In symbols, $\forall x \in \mathcal{U}\ \big[x \in A \cap B \Leftrightarrow (x \in A \vee x \in B)\big]$.
- The set difference between two sets A and B, denoted by $A - B$, is the set of elements that can only be found in A but not in B. In symbol, it means $\forall x \in \mathcal{U}\ \big[x \in A - B \Leftrightarrow (x \in A \wedge x \notin B)\big]$.
- Know the properties of intersection, union, and set differences listed in Theorem 4.3.1.

Exercises 4.3

1. Write each of the following sets by listing its elements explicitly.

(a) $[-4, 4] \cap \mathbb{Z}$ (b) $(-4, 4] \cap \mathbb{Z}$ (c) $(-4, \infty) \cap \mathbb{Z}$
(d) $(-\infty, 4] \cap \mathbb{N}$ (e) $(-4, \infty) \cap \mathbb{Z}^-$ (f) $(4, 5) \cap \mathbb{Z}$

2. Assume $\mathcal{U} = \mathbb{Z}$, and let

$$\begin{aligned} A &= \{\ldots, -6, -4, -2, 0, 2, 4, 6, \ldots\} &&= 2\mathbb{Z}, \\ B &= \{\ldots, -9, -6, -3, 0, 3, 6, 9, \ldots\} &&= 3\mathbb{Z}, \\ C &= \{\ldots, -12, -8, -4, 0, 4, 8, 12, \ldots\} &&= 4\mathbb{Z}. \end{aligned}$$

Describe the following sets by listing their elements explicitly.

(a) $A \cap B$
(b) $C - A$
(c) $A - B$
(d) $A \cap \overline{B}$
(e) $B - A$
(f) $B \cup C$
(g) $(A \cup B) \cap C$
(h) $(A \cup B) - C$

3. Are these statements true or false?

(a) $[1, 2] \cap [2, 3] = \emptyset$
(b) $[1, 2) \cup (2, 3] = [2, 3]$

4. Let the universal set $\mathcal{U}$ be the set of people who voted in the 2012 U.S. presidential election. Define the subsets D, B, and W of $\mathcal{U}$ as follows:

$$\begin{aligned} D &= \{x \in \mathcal{U} \mid x \text{ registered as a Democrat}\}, \\ B &= \{x \in \mathcal{U} \mid x \text{ voted for Barack Obama}\}, \\ W &= \{x \in \mathcal{U} \mid x \text{ belonged to a union}\}. \end{aligned}$$

Express the following subsets of $\mathcal{U}$ in terms of D, B, and W.

(a) People who did not vote for Barack Obama.
(b) Union members who voted for Barack Obama.
(c) Registered Democrats who voted for Barack Obama but did not belong to a union.
(d) Union members who either were not registered as Democrats or voted for Barack Obama.
(e) People who voted for Barack Obama but were not registered as Democrats and were not union members.
(f) People who were either registered as Democrats and were union members, or did not vote for Barack Obama.

5. An insurance company classifies its set $\mathcal{U}$ of policy holders by the following sets:

$$\begin{aligned} A &= \{x \mid x \text{ drives a subcompact car}\}, \\ B &= \{x \mid x \text{ drives a car older than 5 years}\}, \\ C &= \{x \mid x \text{ is married}\}, \\ D &= \{x \mid x \text{ is over 21 years old}\}, \\ E &= \{x \mid x \text{ is a male}\}. \end{aligned}$$

Describe each of the following subsets of $\mathcal{U}$ in terms of A, B, C, D, and E.

(a) Male policy holders over 21 years old.
(b) Policy holders who are either female or drive cars more than 5 years old.
(c) Female policy holders over 21 years old who drive subcompact cars.
(d) Male policy holders who are either married or over 21 years old and do not drive subcompact cars.

6. Let A and B be arbitrary sets. Complete the following statements.

(a) $A \subseteq B \Leftrightarrow A \cap B =$ _____ .
(b) $A \subseteq B \Leftrightarrow A \cup B =$ _____ .
(c) $A \subseteq B \Leftrightarrow A - B =$ _____ .
(d) $A \subset B \Leftrightarrow (A - B =$ _____ $\wedge\ B - A \neq$ _____).
(e) $A \subset B \Leftrightarrow (A \cap B =$ _____ $\wedge\ A \cap B \neq$ _____).
(f) $A - B = B - A \Leftrightarrow$ _____ .

7. Give examples of sets A and B such that $A \in B$ and $A \subset B$.

8. Prove the De Morgan's laws.

9. Let A, B, and C be any three sets. Prove that if $A \subseteq C$ and $B \subseteq C$, then $A \cup B \subseteq C$.

10. Prove Theorem 4.3.2

11. Prove Theorem 4.3.3

12. Let A, B, and C be any three sets. Prove that

(a) $A - B = A \cap \overline{B}$
(b) $A = (A - B) \cup (A \cap B)$
(c) $A - (B - C) = A \cap (\overline{B} \cup C)$
(d) $(A - B) - C = A - (B \cup C)$

13. Comment on the following statements. Are they syntactically correct?

(a) $x \in A \cap x \in B \equiv x \in A \cap B$
(b) $x \in A \wedge B \Rightarrow x \in A \cap B$

14. Prove or disprove each of the following statements about arbitrary sets A and B. If you think a statement is true, prove it; if you think it is false, provide a counterexample.

(a) $\wp(A \cap B) = \wp(A) \cap \wp(B)$
(b) $\wp(A \cup B) = \wp(A) \cup \wp(B)$
(c) $\wp(A - B) = \wp(A) - \wp(B)$

Remark. To show that two sets U and V are equal, we usually want to prove that $x \in U \Leftrightarrow x \in V$. In this problem, the element x is actually a set. Since we usually use uppercase letters to denote sets, we should start the proof of (a) with "Let $S \in \wp(A \cap B)$." If you prefer to use the alternate approach, it looks like the following:

$$\begin{aligned} S \in \wp(A \cap B) &\Leftrightarrow \ldots \\ &\Leftrightarrow \ldots \\ &\vdots \\ &\Leftrightarrow S \in \wp(A) \cap \wp(B). \end{aligned}$$

These remarks also apply to (b) and (c). ♢

4.4 Cartesian Products

Another way to obtain a new set from two given sets A and B is to form ordered pairs. An ***ordered pair*** (x, y) consists of two values x and y. Their order of appearance is important, so we call them first and second elements respectively. Consequently, $(a, b) \neq (b, a)$ unless $a = b$. In general, $(a, b) = (c, d)$ if and only if $a = c$ and $b = d$.

Definition. The ***Cartesian product*** of A and B is the set

$$\boxed{A \times B = \{(a, b) \mid a \in A \wedge b \in B\}.}$$

*Enclose ordered pairs in parentheses. Write them as (x, y), and **not** as $\{x, y\}$.*

Thus, $A \times B$ (read as "A cross B") contains all the ordered pairs in which the first elements are selected from A, and the second elements are selected from B. ♢

Example 4.4.1 Let $A = \{\text{John}, \text{Jim}, \text{Dave}\}$ and $B = \{\text{Mary}, \text{Lucy}\}$. Determine $A \times B$ and $B \times A$.

Solution: We find

$$\begin{aligned} A \times B &= \{(\text{John}, \text{Mary}), (\text{John}, \text{Lucy}), (\text{Jim}, \text{Mary}), (\text{Jim}, \text{Lucy}), (\text{Dave}, \text{Mary}), (\text{Dave}, \text{Lucy})\}, \\ B \times A &= \{(\text{Mary}, \text{John}), (\text{Mary}, \text{Jim}), (\text{Mary}, \text{Dave}), (\text{Lucy}, \text{John}), (\text{Lucy}, \text{Jim}), (\text{Lucy}, \text{Dave})\}. \end{aligned}$$

In general, $A \times B \neq B \times A$. ▲

Example 4.4.2 Determine $A \times B$ and $A \times A$:

(a) $A = \{1, 2\}$ and $B = \{2, 5, 6\}$.

(b) $A = \{5\}$ and $B = \{0, 7\}$.

Solution: (a) We find

$$\begin{aligned} A \times B &= \{(1, 2), (1, 5), (1, 6), (2, 2), (2, 5), (2, 6)\}, \\ A \times A &= \{(1, 1), (1, 2), (2, 1), (2, 2)\}. \end{aligned}$$

(b) The answers are $A \times B = \{(5, 0), (5, 7)\}$, and $A \times A = \{(5, 5)\}$. ▲

Hands-On Exercise 4.4.1 Let $A = \{a, b, c, d\}$ and $B = \{r, s, t\}$. Find $A \times B$, $B \times A$, and $B \times B$.

△

Example 4.4.3 Determine $\wp(\{1, 2\}) \times \{3, 7\}$. Be sure to use correct notation.

Solution: For a complicated problem, divide it into smaller tasks and solve each one separately. Then assemble them to form the final answer. In this problem, we first evaluate

Divide and conquer!

$$\wp(\{1, 2\}) = \big\{\emptyset, \{1\}, \{2\}, \{1, 2\}\big\}.$$

This leads to

$$\begin{aligned} \wp(\{1, 2\}) \times \{3, 7\} &= \big\{\emptyset, \{1\}, \{2\}, \{1, 2\}\big\} \times \{3, 7\} \\ &= \big\{(\emptyset, 3), (\emptyset, 7), (\{1\}, 3), (\{1\}, 7), (\{2\}, 3), (\{2\}, 7), (\{1, 2\}, 3), (\{1, 2\}, 7)\big\}. \end{aligned}$$

Check to make sure that we have matching left and right parentheses, and matching left and right curly braces. ▲

Hands-On Exercise 4.4.2 Find $\{a, b, c\} \times \wp(\{d\})$.

△

Example 4.4.4 How could we describe the contents of the Cartesian product $[1, 3] \times \{2, 4\}$? Since $[1, 3]$ is an infinite set, it is impossible to list all the ordered pairs. We need to use the set-builder notation:

$$[1, 3] \times \{2, 4\} = \{(x, y) \mid 1 \leq x \leq 3, y = 2, 4\}.$$

We can also write $[1, 3] \times \{2, 4\} = \{(x, 2), (x, 4) \mid 1 \leq x \leq 3\}$. ▲

Hands-On Exercise 4.4.3 Describe, using the set-builder notation, the Cartesian product $[1, 3] \times [2, 4]$.

△

Cartesian products can be extended to more than two sets. Instead of ordered pairs, we need ***ordered n-tuples***. The ***n-fold Cartesian product*** of n sets $A_1, A_2, \ldots, A_n$ is the set

$$\boxed{A_1 \times A_2 \times \cdots \times A_n = \{(a_1, a_2, \ldots, a_n) \mid a_i \in A_i \text{ for each } i, 1 \leq i \leq n\}.}$$

In particular, when $A_i = A$ for all i, we abbreviate the Cartesian product as A^n.

Example 4.4.5 The n-dimensional space is denoted $\mathbb{R}^n$. It is the n-fold Cartesian product of $\mathbb{R}$. In special cases, $\mathbb{R}^2$ is the xy-plane, and $\mathbb{R}^3$ is the xyz-space. ▲

Hands-On Exercise 4.4.4 Let $A = \{1, 2\}$, $B = \{a, b\}$, and $C = \{r, s, t\}$. Find $A \times B \times C$.

△

$(A \times B) \times C$, $A \times (B \times C)$ and $A \times B \times C$ are three different sets.

Example 4.4.6 From a technical standpoint, $(A \times B) \times C$ is different from $A \times B \times C$. Can you explain why? Can you discuss the difference, if any, between $(A \times B) \times C$ and $A \times (B \times C)$? For instance, give some specific examples of the elements in $(A \times B) \times C$ and $A \times (B \times C)$ to illustrate their differences.

Solution: The elements of $(A \times B) \times C$ are ordered pairs in which the first coordinates are themselves ordered pairs. A typical element in $(A \times B) \times C$ takes the form of

$$\big((a, b), c\big).$$

The elements in $A \times B \times C$ are ordered triples of the form

$$(a, b, c).$$

Since their elements look different, it is clear that $(A \times B) \times C \neq A \times B \times C$. Likewise, a typical element in $A \times (B \times C)$ looks like

$$\big(a, (b, c)\big).$$

Therefore, $(A \times B) \times C \neq A \times (B \times C)$, and $A \times (B \times C) \neq A \times B \times C$. ▲

Theorem 4.4.1 *For any sets A, B, and C, we have*

$$\begin{aligned} A \times (B \cup C) &= (A \times B) \cup (A \times C), \\ A \times (B \cap C) &= (A \times B) \cap (A \times C), \\ A \times (B - C) &= (A \times B) - (A \times C). \end{aligned}$$

Remark. How would we show that the two sets S and T are equal? We need to show that

$$x \in S \Leftrightarrow x \in T.$$

The complication in this problem is that both S and T are Cartesian products, so x takes on a special form, namely, that of an ordered pair. Consider the first identity as an example; we need to show that

$$(u, v) \in A \times (B \cup C) \Leftrightarrow (u, v) \in (A \times B) \cup (A \times C).$$

We prove this in two steps: first showing $\Rightarrow$, then $\Leftarrow$, which is equivalent to first showing $\subseteq$, then $\supseteq$. Alternatively, we can use $\Leftrightarrow$ throughout the argument.

Proof 1: Let $(u, v) \in A \times (B \cup C)$. Then $u \in A$, and $v \in B \cup C$. The definition of union implies that $v \in B$ or $v \in C$. Thus far, we have found

(i) $u \in A$ and $v \in B$, or
(ii) $u \in A$ and $v \in C$.

This is equivalent to

(i) $(u, v) \in A \times B$, or
(ii) $(u, v) \in A \times C$.

Thus, $(u, v) \in (A \times B) \cup (A \times C)$. This proves that $A \times (B \cup C) \subseteq (A \times B) \cup (A \times C)$.

Next, let $(u, v) \in (A \times B) \cup (A \times C)$. Then $(u, v) \in A \times B$, or $(u, v) \in A \times C$. This means

(i) $u \in A$ and $v \in B$, or
(ii) $u \in A$ and $v \in C$.

Both conditions require $u \in A$, so we can rewrite them as

(i) $u \in A$, and
(ii) $v \in B$ or $v \in C$;

which is equivalent to

(i) $u \in A$, and
(ii) $v \in B \cup C$.

Thus, $(u, v) \in A \times (B \cup C)$. We have proved that $(A \times B) \cup (A \times C) \subseteq A \times (B \cup C)$. Together with $A \times (B \cup C) \subseteq (A \times B) \cup (A \times C)$ that we have proved earlier, we conclude that $A \times (B \cup C) = (A \times B) \cup (A \times C)$. ■

Proof 2: We shall only prove the first equality. Since

$$\begin{aligned} (u, v) \in A \times (B \cup C) &\Leftrightarrow u \in A \wedge v \in (B \cup C) && \text{(defn. of Cartesian product)} \\ &\Leftrightarrow u \in A \wedge (v \in B \vee v \in C) && \text{(defn. of union)} \\ &\Leftrightarrow (u \in A \wedge v \in B) \vee (u \in A \wedge v \in C) && \text{(distributive law)} \\ &\Leftrightarrow (u, v) \in A \times B \vee (u, v) \in A \times C && \text{(defn. of Cartesian product)} \\ &\Leftrightarrow (u, v) \in (A \times B) \cup (A \times C) && \text{(defn. of union)} \end{aligned}$$

we conclude that $A \times (B \cup C) = (A \times B) \cup (A \times C)$. ■

Theorem 4.4.2 *If A and B are finite sets, with $|A| = m$ and $|B| = n$, then $|A \times B| = mn$.*

Proof: The elements of $A \times B$ are ordered pairs of the form (a, b), where $a \in A$, and $b \in B$. There are m choices of a. For each fixed a, we can form the ordered pair (a, b) in n ways, because there are n choices for b. Together, the ordered pairs (a, b) can be formed in mn ways. ■

The argument we used in the proof is called ***multiplication principle***. We shall study it again in Chapter 8. In brief, it says that if a job can be completed in several steps, then the number of ways to finish the job is the product of the number of ways to finish each step.

Corollary 4.4.3 *If $A_1, A_2, \ldots, A_n$ are finite sets, then $|A_1 \times A_2 \times \cdots \times A_n| = |A_1| \cdot |A_2| \cdots |A_n|$.*

Corollary 4.4.4 *If A is a finite set with $|A| = n$, then $|\wp(A)| = 2^n$.*

Proof: Let the elements of A be $a_1, a_2, \ldots, a_n$. The elements of $\wp(A)$ are subsets of A. Each subset of A contains some elements from A. Associate to each subset S of A an ordered n-tuple $\big(b_1, b_2, \ldots, b_n\big)$ from $\{0, 1\}^n$ such that

$$b_i = \begin{cases} 0 & \text{if } a_i \notin S, \\ 1 & \text{if } a_i \in S. \end{cases}$$

The value of the ith element in this ordered n-tuple indicates whether the subset S contains the element a_i. It is clear that the subsets of A are in one-to-one correspondence with the n-tuples. This means the power set $\wp(A)$ and the Cartesian product $\{0, 1\}^n$ have the same cardinality. Since there are 2^n ordered n-tuples, we conclude that there are 2^n subsets as well. ■

This idea of one-to-one correspondence is a very important concept in mathematics. We shall study it again in Chapter 6.

Summary and Review

- The Cartesian product of two sets A and B, denoted $A \times B$, consists of ordered pairs of the form (a, b), where a comes from A, and b comes from B.
- Since ordered pairs are involved, $A \times B$ usually is not equal to $B \times A$.
- The notion of ordered pairs can be extended analogously to ordered n-tuples, thereby yielding an n-fold Cartesian product.
- If A and B are finite sets, then $|A \times B| = |A| \cdot |B|$.

Exercises 4.4

1. Let $X = \{-2, 2\}$, $Y = \{0, 4\}$ and $Z = \{-3, 0, 3\}$. Evaluate the following Cartesian products.

   (a) $X \times Y$ (b) $X \times Z$ (c) $Z \times Y \times Y$

2. Consider the sets X, Y and Z defined in Problem 1. Evaluate the following Cartesian products.

   (a) $X \times Y \times Z$ (b) $(X \times Y) \times Z$ (c) $X \times (Y \times Z)$

3. Without listing all the elements of $X \times Y \times X \times Z$, where X, Y, and Z are defined in Problem 1, determine $|X \times Y \times X \times Z|$.

4. Determine $|\wp(\wp(\wp(\{1, 2\})))|$.

5. Consider the set $X = \{-2, 2\}$. Evaluate the following Cartesian products.

   (a) $X \times \wp(X)$ (b) $\wp(X) \times \wp(X)$ (c) $\wp(X \times X)$

6. Let A and B be arbitrary nonempty sets.

   (a) Under what condition does $A \times B = B \times A$?
   (b) Under what condition is $(A \times B) \cap (B \times A)$ empty?

7. Let A, B, and C be any three sets. Prove that

   (a) $A \times (B \cap C) = (A \times B) \cap (A \times C)$
   (b) $A \times (B - C) = (A \times B) - (A \times C)$

8. Let A, B, and C be any three sets. Prove that if $A \subseteq B$, then $A \times C \subseteq B \times C$.

4.5 Index Sets

The notion of union can be extended to three sets:

$$A \cup B \cup C = \{x \in \mathcal{U} \mid (x \in A) \vee (x \in B) \vee (x \in C)\}.$$

It is obvious how to generalize it to the union of any number of sets. We use a notation that resembles the summation notation to describe such a union:

$$\bigcup_{i=1}^{n} A_i = A_1 \cup A_2 \cup \cdots \cup A_n.$$

We define

$$\bigcup_{i=1}^{n} A_i = \{x \in \mathcal{U} \mid (x \in A_1) \vee (x \in A_2) \vee \cdots \vee (x \in A_n)\}.$$

It looks messy! Here is a better alternative:

$$\boxed{\bigcup_{i=1}^{n} A_i = \{x \in \mathcal{U} \mid x \in A_i \text{ for } \textit{some } i, \text{where } 1 \leq i \leq n\}.}$$

In a similar manner, $\bigcap_{i=1}^{n} A_i = A_1 \cap A_2 \cap \cdots \cap A_n$, and we define

$$\boxed{\bigcap_{i=1}^{n} A_i = \{x \in \mathcal{U} \mid x \in A_i \text{ for } \textit{all } i, \text{where } 1 \leq i \leq n\}.}$$

In plain English, $\bigcup_{i=1}^{n} A_i$ is the collection of all elements in the A_i's, and $\bigcap_{i=1}^{n} A_n$ is the collection of all elements *common* to all A_i's.

Example 4.5.1 For $i = 1, 2, 3, \ldots$, let $A_i = [-i, i]$. First, construct several A_i for comparison, because it may help us detect any specific pattern. See Figure 4.6. It is clear that $A_1 \subset A_2 \subset \cdots$. Thus, $\bigcup_{i=1}^{n} A_i = [-n, n] = A_n$, and $\bigcap_{i=1}^{n} A_i = [-1, 1] = A_1$. ▲

Hands-On Exercise 4.5.1 Evaluate $\bigcup_{i=1}^{n} B_i$ and $\bigcap_{i=1}^{n} B_i$, where $B_i = [0, 2i)$.

△

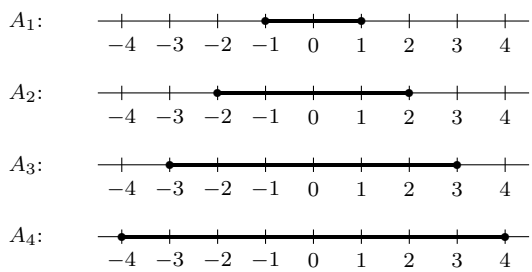


Figure 4.6: Comparing intervals to find their union and intersection.

It is obvious that we can also extend the upper bound to infinity.

$$\begin{aligned}\bigcup_{i=1}^{\infty} A_i &= A_1 \cup A_2 \cup \cdots = \{x \in \mathcal{U} \mid x \in A_i \text{ for } \textit{some } i \in \mathbb{N}\},\\ \bigcap_{i=1}^{\infty} A_i &= A_1 \cap A_2 \cap \cdots = \{x \in \mathcal{U} \mid x \in A_i \text{ for } \textit{all } i \in \mathbb{N}\}.\end{aligned}$$

In some situations, we may borrow the idea of partial sums from calculus. We first find the union or intersection of the first n sets, then take the limit as n approaches infinity. Thus, if the limit is well-defined, then

$$\bigcup_{i=1}^{\infty} A_i = \lim_{n\to\infty} \bigcup_{i=1}^{n} A_i, \qquad \text{and} \qquad \bigcap_{i=1}^{\infty} A_i = \lim_{n\to\infty} \bigcap_{i=1}^{n} A_i.$$

Inclusion or exclusion of endpoints may change when we take limit.

Example 4.5.2 Let $A_i = [-i, i]$. We have learned from the last example that $\bigcup_{i=1}^{n} A_i = [-n, n]$ and $\bigcap_{i=1}^{n} A_i = [-1, 1]$. Hence,

$$\bigcup_{i=1}^{\infty} A_i = \lim_{n\to\infty} [-n, n] = (-\infty, \infty), \qquad \text{and} \qquad \bigcap_{i=1}^{\infty} A_i = [-1, 1].$$

Recall that we write $(-\infty, \infty)$ instead of $[-\infty, \infty]$ because $\pm\infty$ are *not* numbers, they are merely symbols representing infinitely large values. ▲

Hands-On Exercise 4.5.2 Evaluate $\bigcup_{i=1}^{\infty} B_i$ and $\bigcap_{i=1}^{\infty} B_i$, where $B_i = [0, 2i)$.

△

Example 4.5.3 Let $B_i = \left(0, 1 - \frac{1}{2i}\right]$. Determine $\bigcup_{i=1}^{\infty} B_i$ and $\bigcap_{i=1}^{\infty} B_i$.

First take finite union and intersection.

Solution: Once again, we have $B_1 \subset B_2 \subset \cdots$. It is easy to check that

$$\bigcup_{i=1}^{n} B_i = B_n = \left(0, 1 - \frac{1}{2n}\right], \qquad \text{and} \qquad \bigcap_{i=1}^{n} B_i = B_1 = \left(0, \frac{1}{2}\right].$$

Then take limit.

It follows that

$$\bigcup_{i=1}^{\infty} B_i = \lim_{n\to\infty} \left(0, 1 - \frac{1}{2n}\right] = (0, 1), \qquad \text{and} \qquad \bigcap_{i=1}^{\infty} B_i = \left(0, \frac{1}{2}\right].$$

The last step: check the endpoints.

Note that $\lim_{n\to\infty} \left(0, 1 - \frac{1}{2n}\right] \neq (0, 1]$ because the endpoint 1 does not belong to any B_i. ▲

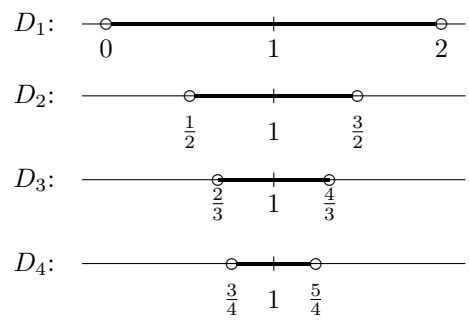
Hands-On Exercise 4.5.3 Let $C_i = \left[0, 1 - \frac{1}{i}\right]$. Determine $\bigcup_{i=1}^{\infty} C_i$ and $\bigcap_{i=1}^{\infty} C_i$.

△

Example 4.5.4 Let $D_i = \left(1 - \frac{1}{i}, 1 + \frac{1}{i}\right)$. Determine $\bigcup_{i=1}^{\infty} D_i$ and $\bigcap_{i=1}^{\infty} D_i$.

Solution: As the value of i increases, the value of $\frac{1}{i}$ decreases. Hence, the left endpoint $1 - \frac{1}{i}$ increases, and the right endpoint $1 + \frac{1}{i}$ decreases.

i	$D_i = \left(1 - \frac{1}{i}, 1 + \frac{1}{i}\right)$
1	$(0, 2)$
2	$\left(\frac{1}{2}, \frac{3}{2}\right)$
3	$\left(\frac{2}{3}, \frac{4}{3}\right)$
4	$\left(\frac{3}{4}, \frac{5}{4}\right)$



It is clear that $D_1 \supseteq D_2 \supseteq D_3 \supseteq \cdots$. Thus, $\bigcup_{i=1}^{\infty} D_i = D_1 = (0, 2)$, and $\bigcap_{i=1}^{\infty} D_i = \{1\}$. ▲

Hands-On Exercise 4.5.4 Let $E_i = \left[-i, 1 + \frac{1}{i}\right)$. Determine $\bigcup_{i=1}^{\infty} E_i$ and $\bigcap_{i=1}^{\infty} E_i$.

△

Hands-On Exercise 4.5.5 For each positive integer i, define $F_i = \{i, i+1, i+2, \ldots, 3i\}$. Determine $\bigcup_{i=1}^{\infty} F_i$ and $\bigcap_{i=1}^{\infty} F_i$.

△

The next two results are obvious.

Theorem 4.5.1 *If $A_1 \subseteq A_2 \subseteq A_3 \subseteq \cdots$, then $\bigcap_{i=1}^{\infty} A_i = A_1$.*

Theorem 4.5.2 *If $A_1 \supseteq A_2 \supseteq A_3 \supseteq \cdots$, then $\bigcup_{i=1}^{\infty} A_i = A_1$.*

How could we describe the union $A_2 \cup A_4 \cup A_6 \cup \cdots$? Well, we can write

$$\bigcup_{i \text{ even}} A_i,$$

which means that union of A_i, where i is even. Since the set of even positive integers is denoted by $2\mathbb{N}$, another way to describe the same union is

$$\bigcup_{i \in 2\mathbb{N}} A_i.$$

It means the union all A_i, where i is taken out from the set $2\mathbb{N}$. Accordingly,

$$\bigcup_{i=0}^{\infty} A_i = \bigcup_{i \in \mathbb{N}} A_i, \qquad \text{and} \qquad \bigcap_{i=0}^{\infty} A_i = \bigcap_{i \in \mathbb{N}} A_i.$$

We can even go one step further, by allowing i to be taken from any set of integers, or any set of real numbers, or even any set of objects. The only restriction is that A_i must exist, and its content must somehow depend on i.

In general, given a nonempty set I, if we could associate with each $i \in I$ a set A_i, we define the ***indexed family of sets*** $\mathcal{A}$ as

$$\mathcal{A} = \{A_i \mid i \in I\}.$$

We call I the ***index set***, and define

$$\begin{aligned} \bigcup_{i \in I} A_i &= \{x \mid x \in A_i \text{ for } \textit{some } i \in I\}, \\ \bigcap_{i \in I} A_i &= \{x \mid x \in A_i \text{ for } \textit{all } i \in I\}. \end{aligned}$$

Let us look at a few examples.

Example 4.5.5 To describe the union

$$A_1 \cup A_3 \cup A_7 \cup A_{11} \cup A_{23},$$

we first define the index set to be $I = \{1, 3, 7, 11, 23\}$, which is the set of all the subscripts used in the union. Now the union can be conveniently described as $\bigcup_{i \in I} A_i$. ▲

Example 4.5.6 Consider five sets

$$\begin{aligned} A_1 &= \{1, 4, 23\}, \\ A_2 &= \{7, 11, 23\}, \\ A_3 &= \{3, 6, 9\}, \\ A_4 &= \{5, 17, 22\}, \\ A_5 &= \{3, 6, 23\}. \end{aligned}$$

Let $I = \{2, 5\}$, then

$$\bigcup_{i \in I} A_i = A_2 \cup A_5 = \{7, 11, 23\} \cup \{3, 6, 23\} = \{3, 6, 7, 11, 23\}.$$

Likewise, $\bigcap_{i \in I} A_i = A_2 \cap A_5 = \{7, 11, 23\} \cap \{3, 6, 23\} = \{23\}$. ▲

Hands-On Exercise 4.5.6 Let $J = \{1, 4, 5\}$. Evaluate $\bigcup_{i \in J} A_i$ and $\bigcap_{i \in J} A_i$, where A_is are defined in the last example.

△

Hands-On Exercise 4.5.7 An index set could be a set of any objects. For instance, the sets of numbers in the last example could be the favorite Lotto numbers of five different students. We could index these sets according to the names of the students:

$$\begin{aligned} A_{\text{John}} &= \{1, 4, 23\}, \\ A_{\text{Mary}} &= \{7, 11, 23\}, \\ A_{\text{Joe}} &= \{3, 6, 9\}, \\ A_{\text{Pete}} &= \{5, 17, 22\}, \\ A_{\text{Lucy}} &= \{3, 6, 23\}. \end{aligned}$$

If $I = \{\text{Mary}, \text{Joe}, \text{Lucy}\}$, what is $\bigcup_{i \in I}$? How would you interpret its physical meaning?

△

Example 4.5.7 Let $I = \{x \mid x \text{ is a living human being}\}$, and define

$$\begin{aligned} B_i &= \{x \in I \mid x \text{ is a child of } i\}, \\ A_i &= \{i\} \cup B_i \end{aligned}$$

for each $i \in I$. Then

$$\bigcap_{i \in I} A_i = \emptyset, \qquad \bigcup_{i \in I} A_i = I, \qquad \bigcap_{i \in I} B_i = \emptyset,$$

and

$$\bigcup_{i \in I} B_i = I - \{x \mid x\text{'s parents are both deceased}\}.$$

We leave it as an exercise to verify these unions and intersections. ▲

Hands-On Exercise 4.5.8 Verify the intersection and union in the last example.

△

Hands-On Exercise 4.5.9 If I represents a set of students, and A_i represents the set of friends of student i, interpret the meaning of $\bigcup_{i\in I} A_i$ and $\bigcap_{i\in I} A_i$.

△

We close this section with yet another generalization of De Morgan's laws.

Theorem 4.5.3 (Extended De Morgan's Laws) *For any nonempty index set I, we have*

$$\overline{\bigcup_{i\in I} A_i} = \bigcap_{i\in I} \overline{A_i}, \qquad \textit{and} \qquad \overline{\bigcap_{i\in I} A_i} = \bigcup_{i\in I} \overline{A_i}.$$

Proof 1: Let $x \in \overline{\bigcup_{i\in I} A_i}$, then

$$x \notin \bigcup_{i\in I} A_i = \{x \mid x \in A_i \text{ for } \textit{some } i \in I\}.$$

This means $x \notin A_i$ for every $i \in I$. Hence, $x \in \overline{A_i}$ for each $i \in I$. Consequently,

$$x \in \bigcap_{i\in I} \overline{A_i}.$$

This proves that $\overline{\bigcup_{i\in I} A_i} \subseteq \bigcap_{i\in I} \overline{A_i}$.

Next, let $x \in \bigcap_{i\in I} \overline{A_i}$. Then $x \in \overline{A_i}$ for each $i \in I$. This means $x \notin A_i$ for each $i \in I$. Then

$$x \notin \{x \mid x \in A_i \text{ for } \textit{some } i \in I\} = \bigcup_{i\in I} A_i.$$

Thus, $x \in \overline{\bigcup_{i\in I} A_i}$, proving that $\bigcap_{i\in I} \overline{A_i} \subseteq \overline{\bigcup_{i\in I} A_i}$. We proved earlier that $\overline{\bigcup_{i\in I} A_i} \subseteq \bigcap_{i\in I} \overline{A_i}$. Therefore, the two sets must be equal.

The proof of $\overline{\bigcap_{i\in I} A_i} = \bigcup_{i\in I} \overline{A_i}$ proceeds in a similar manner, and is left as an exercise. ■

Proof 2: We shall prove $\overline{\bigcup_{i\in I} A_i} = \bigcap_{i\in I} \overline{A_i}$. We leave out the explanations for you to fill in:

$$\begin{aligned}
x \in \overline{\bigcup_{i\in I} A_i} \quad &\Leftrightarrow \quad \overline{x \in \bigcup_{i\in I} A_i} \\
&\Leftrightarrow \quad \overline{x \in A_i \text{ for some } i} \\
&\Leftrightarrow \quad x \notin A_i \text{ for all } i \\
&\Leftrightarrow \quad x \in \overline{A_i} \text{ for all } i \\
&\Leftrightarrow \quad x \in \bigcap_{i\in I} \overline{A_i}.
\end{aligned}$$

The proof of $\overline{\bigcap_{i\in I} A_i} = \bigcup_{i\in I} \overline{A_i}$ is left as an exercise. ■

Summary and Review

- When dealing with arbitrary intersection or union of intervals, first identify the endpoints, then analyze the sets involved in the operation to determine whether an endpoint should be included or excluded.
- Intersection and union can be performed on a group of similar sets identified by subscripts belonging to an index set.
- Consequently, intersection or union can be formed by naming a specific index set.

Exercises 4.5

1. For each $n \in \mathbb{Z}^+$, define $A_n = \left(-\frac{1}{n}, 2n\right)$. Find $\bigcap_{n=1}^{\infty} A_n$ and $\bigcup_{n=1}^{\infty} A_n$.
2. For each $n \in \mathbb{Z}^+$, define $B_n = \{m \in \mathbb{Z} \mid -\frac{n}{2} \leq m \leq 3n\}$. Evaluate $\bigcap_{n=1}^{\infty} B_n$ and $\bigcup_{n=1}^{\infty} B_n$.
3. Define $C_n = \{n, n+1, n+2, \ldots, 2n+1\}$ for each integer $n \geq 0$. Evaluate $\bigcap_{n=0}^{\infty} C_n$ and $\bigcup_{n=0}^{\infty} C_n$.
4. For each $n \in I = \{1, 2, 3, \ldots, 100\}$, define $D_n = [-n, 2n] \cap \mathbb{Z}$. Evaluate $\bigcap_{n \in I} D_n$ and $\bigcup_{n \in I} D_n$.
5. For each $n \in \mathbb{N}$, define $E_n = \{-n, -n+1, -n+2, \ldots, n^2\}$. Evaluate $\bigcap_{n \in \mathbb{N}} E_n$ and $\bigcup_{n \in \mathbb{N}} E_n$.
6. For each $n \in \mathbb{N}$, define $F_n = \left\{\frac{m}{n} \mid m \in \mathbb{Z}\right\}$. Evaluate $\bigcap_{n \in \mathbb{N}} F_n$ and $\bigcup_{n \in \mathbb{N}} F_n$.
7. Let $I = (0, 1)$, and define $A_i = \left[1, \frac{1}{i}\right]$ for each $i \in I$. For instance $A_{0.5} = [1, 2]$ and $A_{\frac{\pi}{4}} = \left[1, \frac{4}{\pi}\right]$. Evaluate $\bigcup_{i \in I} A_i$ and $\bigcap_{i \in I} A_i$.
8. Define $I = (0, 1)$, and for each $i \in I$, let $B_i = (-i, \frac{1}{i})$. Evaluate $\bigcup_{i \in I} B_i = (-1, \infty)$ and $\bigcap_{i \in I} B_i$.
9. Evaluate $\bigcap_{x \in (1,2)} (1 - 2x, x^2)$ and $\bigcup_{x \in (1,2)} (1 - 2x, x^2)$.
10. Evaluate $\bigcap_{x \in (0,1)} \left(x, \frac{1}{x}\right)$ and $\bigcup_{x \in (0,1)} \left(x, \frac{1}{x}\right)$.
11. Let the universal set be $\mathbb{R}^2$. For each $r \in (0, \infty)$, define

$$A_r = \{(x, y) \mid y = rx^2\};$$

that is, A_r is the set of points on the parabola $y = rx^2$, where $r > 0$. Evaluate $\bigcap_{r \in (0,\infty)} A_r$ and $\bigcup_{r \in (0,\infty)} A_r$.

12. Prove that $\overline{\bigcap_{i \in I} A_i} = \bigcup_{i \in I} \overline{A_i}$ for any nonempty index set I.

Chapter 4

SWDM 6: Functions

Chapter 6

Functions

6.1 Functions: An Introduction

The functions we studied in calculus are real functions, which are defined over a set of real numbers, and the results they produce are also real. In this chapter, we shall study their generalization over other sets. The definition could be difficult to grasp at the beginning, so we would start with a brief introduction.

Most students view real functions as computational devices. However, in the generalization, functions are not restricted to computation only. A better way to look at functions is their input-output relationship. Let f denote a function. Given an element (which need not be a number), we call the result from f the ***image of x under f***, and write $f(x)$, which is read as "f of x."

Imagine f as a machine. It takes the input value x, and returns $f(x)$ as the output value. This input-output relationship is depicted in Figure 6.1 in two different ways.

Figure 6.1: Two pictorial views of a function as a machine.

The question is: how could we obtain $f(x)$? A function need not involve any computation. Consequently, we cannot speak of "computing" the value of $f(x)$. Instead, we talk about what is the rule we follow to obtain $f(x)$. This rule can be described in many forms. We can, of course, use a computational rule. But a table, an algorithm, or even a verbal description also work as well.

When we say a real function is defined over the real numbers, we mean the input values must be real numbers. The output values are also real numbers. In general, the input and output values need not be of the same type. The ***nearest integer function***, denoted $[x]$, rounds the real number x to the nearest integer. Here, the images (the output values) are integers. Consequently, we need to distinguish the set of input values from the set of possible output values. We call them the ***domain*** and the ***codomain***, respectively, of the function.

Example 6.1.1 When a professor reports the final letter grades for the students in her class, we can regard this as a function g. The domain is the set of students in her class, and the codomain could be the set of letter grades $\{A, B, C, D, F\}$. ▲

We said the codomain is the set of *possible* output values, because not every element in the codomain needs to appear as the image of some element from the domain. If no student fails the professor's class in Example 6.1.1, no one will receive the final grade F. The collection of the images (the final letter grades) form a subset of the codomain. We call this subset the ***range*** of the function g. The range of a function can be a *proper* subset of the codomain. Hence, the codomain of a function is different from the set of its images. If the range of a function does equal to the codomain, we say that the function is ***onto***.

Example 6.1.2 For the nearest integer function $h(x) = [x]$, the domain is $\mathbb{R}$. The codomain is $\mathbb{Z}$, and the range is also $\mathbb{Z}$. Hence, the nearest integer function is onto. ▲

Example 6.1.3 Let x be a real number. The ***greatest integer function*** $\lfloor x \rfloor$ returns the greatest integer less than or equal to x. For example,

$$\left\lfloor \sqrt{50} \right\rfloor = 7, \qquad \lfloor -6.34 \rfloor = -7, \qquad \text{and} \qquad \lfloor 15 \rfloor = 15.$$

Therefore, $\lfloor x \rfloor$ returns x if it is an integer, otherwise, it rounds x *down* to the next closest integer. Hence, it is also called the ***floor function*** of x. It is clear that its domain is $\mathbb{R}$, and the codomain and range are both $\mathbb{Z}$. ▲

Hands-On Exercise 6.1.1 Let x be a real number. The ***least integer function*** $\lceil x \rceil$ returns the least integer greater than or equal to x. For example,

$$\left\lceil \sqrt{50} \right\rceil = 8, \qquad \lceil -6.34 \rceil = -6, \qquad \text{and} \qquad \lceil 15 \rceil = 15.$$

Thus, $\lceil x \rceil$ returns x if it is an integer, otherwise, it rounds x *up* to the next closest integer. Hence, it is also called the ***ceiling function*** of x. What is its domain and codomain?

△

We impose two restrictions on the input-output relationships that we call functions. For any fixed input value x, the output from a function must be the same every time we use the function. As a machine, it spits out the same answer every time we feed the same value x to it. As a calculator, it displays the same answer on its screen every time we enter the same value x, and push the button for the function. We call the output value the image of x, and write $f(x)$. The first important requirement for a function f to be well-defined is: the image $f(x)$ is *unique* for any fixed x-value.

A good machine must perform properly. In terms of a function f, we must be able to obtain $f(x)$ for any value x (and, of course, produce only one result for each x). This is perhaps a little bit too demanding. A remedy is to restrict our attention to those x's over which f would work. The set of legitimate input values is precisely what we call the domain of the function. Consequently, the second requirement says: for every element x from the domain, the output value $f(x)$ should be well-defined. This is the mathematical way of saying that the value $f(x)$ can be obtained.

Example 6.1.4 Compare this to a calculator. If you enter a negative number and press the $\sqrt{\ }$ button, an error message will appear. To be able to compute the square root of a number, the number must be nonnegative. The domain of a function is the set of acceptable input values for which meaningful results can be found. For the square root function, the domain is $\mathbb{R}^+ \cup \{0\}$, which is the set of nonnegative real numbers. ▲

Hands-On Exercise 6.1.2 For the square root function, we may regard its codomain as $\mathbb{R}$. What is its range? Is the function onto?

△

Hands-On Exercise 6.1.3 For the square root function, can we say its domain is $\mathbb{R}^+ \cup 0$? Explain.

△

The two conditions for a function to be well-defined are often combined and written as if it were only one condition:

> *A function f is well-defined if every element x from the domain has a unique image in the codomain.*

Every element in the domain has a unique image.

When you examine this definition closer, you will find the two separate requirements:

- every element in the domain has an image under f, and
- the image is unique.

In the next section, we shall present the complete formal definition.

Summary and Review

- A function is a rule that assigns to every element in the domain a unique image in the codomain.

Exercises 6.1

1. Complete the following table:

x	5.7	π	e	-7.2	-0.8	9
$\lfloor x \rfloor$						
$\lceil x \rceil$						
$[x]$						

2. What is the domain and the codomain of the cube root function? Is it onto?

3. For the square root function, how would you use the interval notation to describe the domain?

4. For the square root function, which set complement would you use to describe the domain?

6.2 Definition of Functions

Definition. Let A and B be nonempty sets. A ***function*** from A to B is a rule that assigns to *every element* of A a *unique* element in B. We call A the ***domain***, and B the ***codomain***, of the function. If the function is called f, we write $f\colon A \to B$. Given $x \in A$, its associated element in B is called its ***image*** under f. We denote it $f(x)$, which is pronounced as "f of x." ♢

A function is sometimes called a ***map*** or ***mapping***. Hence, we sometimes say f ***maps*** x to its image $f(x)$. Functions are also called ***transformations***.

Example 6.2.1 The function $f\colon \{a, b, c\} \to \{1, 3, 5, 9\}$ is defined according to the rule

$$f(a) = 1, \qquad f(b) = 5, \qquad \text{and} \qquad f(c) = 9.$$

It is a well-defined function. The rule of assignment can be summarized in a table:

x	a	b	c
$f(x)$	1	5	9

We can also describe the assignment rule pictorially with an ***arrow diagram***, as shown in Figure 6.2. ▲

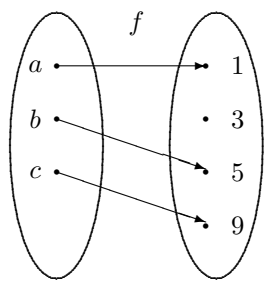


Figure 6.2: An example of a well-defined function.

The two key requirements of a function are

Every element in the domain has a unique image.

- every element in the domain has an image under f, and
- the image is unique.

You may want to remember that every element in A has exactly one "partner" in B.

Example 6.2.2 Figure 6.3 depicts two examples of non-functions. In the one on the left, one of the elements in the domain has no image associated with it. In the one on the right, one of the elements in the domain has two images assigned to it. Both are not functions. ▲

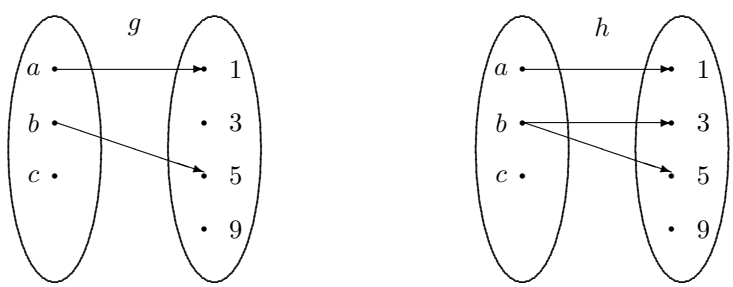


Figure 6.3: Two types of non-functions.

Hands-On Exercise 6.2.1 Do these rules

x	a	b	c
$f(x)$	5	3	3

x	b	c
$g(x)$	9	5

x	a	b	b	c
$h(x)$	1	5	3	9

produce well-defined functions from $\{a, b, c\}$ to $\{1, 3, 5, 9\}$? Explain.

△

Hands-On Exercise 6.2.2 Does the definition

$$r(x) = \begin{cases} x & \text{if today is Monday,} \\ 2x & \text{if today is not Monday} \end{cases}$$

produce a well-defined function from $\mathbb{R}$ to $\mathbb{R}$? Explain.

△

Hands-On Exercise 6.2.3 Does the definition

$$s(x) = \begin{cases} 5 & \text{if } x < 2, \\ 7 & \text{if } x > 3, \end{cases}$$

produce a well-defined function from $\mathbb{R}$ to $\mathbb{R}$? Explain.

△

Example 6.2.3 The function $f\colon [0, \infty) \to \mathbb{R}$ defined by

$$f(x) = \sqrt{x}$$

is well-defined. So is the function $g\colon [2, \infty) \to \mathbb{R}$ defined as

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

Can you explain why the domain is $[2, \infty)$? ▲

Example 6.2.4 Let A denote the set of students taking Discrete Mathematics, and $G = \{A, B, C, D, F\}$, and $\ell(x)$ is the final grade of student x in Discrete Mathematics. Every student should receive a final grade, and the instructor has to report one and only one final grade for each student. This is precisely what we call a function. ▲

Example 6.2.5 The function $n\colon \wp(\{a, b, c, d\}) \to \mathbb{Z}$ is defined as $n(S) = |S|$. It evaluates the cardinality of a subset of $\{a, b, c, d\}$. For example,

$$n\big(\{a, c\}\big) = n\big(\{b, d\}\big) = 2.$$

Note that $n(\emptyset) = 0$. ▲

Hands-On Exercise 6.2.4 Consider Example 6.2.5. What other subsets S of $\{a, b, c, d\}$ also yield $n(S) = 2$? What are the smallest and the largest images the function n can produce?

△

Example 6.2.6 Consider a function $f\colon \mathbb{Z}_7 \to \mathbb{Z}_5$. The domain and the codomain are,

$$\mathbb{Z}_7 = \{0, 1, 2, 3, 4, 5, 6\}, \qquad \text{and} \qquad \mathbb{Z}_5 = \{0, 1, 2, 3, 4\},$$

respectively. Not only are their elements different, their binary operations are different too. In the domain $\mathbb{Z}_7$, the arithmetic is performed modulo 7, but the arithmetic in the codomain $\mathbb{Z}_5$ is done modulo 5. So we need to be careful in describing the rule of assignment if a computation is involved. We could say, for example,

Exercise caution when the domain and codomain involve different moduli.

$$f(x) = z, \quad \text{where } z \equiv 3x \pmod 5.$$

Consequently, starting with any element x in $\mathbb{Z}_7$, we consider x as an ordinary integer, multiply by 3, and reduce the answer modulo 5 to obtain the image $f(x)$. For brevity, we shall write

$$f(x) \equiv 3x \pmod 5.$$

We summarize the images in the following table:

n	0	1	2	3	4	5	6
$f(n)$	0	3	1	4	2	0	3

Take note that the images start repeating after $f(4) = 2$. ▲

Hands-On Exercise 6.2.5 Tabulate the images of $g\colon \mathbb{Z}_{10} \to \mathbb{Z}_5$ defined by

$$g(x) \equiv 3x \pmod 5.$$

△

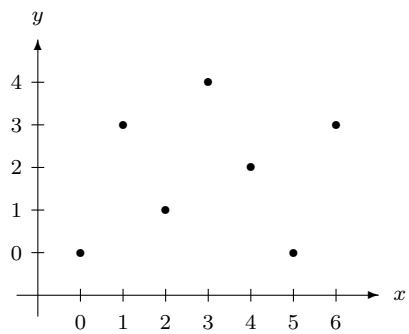
Definition. The ***graph*** of a function $f\colon A \to B$ is the set of ordered pairs (x, y) from $A \times B$ such that $y = f(x)$. ♢

The graph of a function, in this general definition, may not look like the kind of graphs we expected from real functions. A graph is, by definition, a set of *ordered pairs*.

Example 6.2.7 The graph of the function f in Example 6.2.6 is the set of ordered pairs

$$\{(0, 0), (1, 3), (2, 1), (3, 4), (4, 2), (5, 0), (6, 3)\}.$$

If one insists, we could display the graph of a function using an xy-plane that resembles the usual Cartesian plane. Keep in mind: the elements x and y come from A and B, respectively. We can "plot" the graph for f in Example 6.2.6 as shown below.



Besides using a graphical representation, we can also use a $(0,1)$-matrix. A $(0,1)$-matrix is a matrix whose entries are 0 and 1. For the function f, we use a 7×5 matrix, whose rows and columns correspond to the elements of A and B, respectively, and put one in the (i,j)th entry if $j = f(i)$, and zero otherwise. The resulting matrix is

$$\begin{array}{c} \\ 0 \\ 1 \\ 2 \\ 3 \\ 4 \\ 5 \\ 6 \end{array} \begin{array}{c} \begin{array}{ccccc} 0 & 1 & 2 & 3 & 4 \end{array} \\ \begin{pmatrix} 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 \\ 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 1 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 \end{pmatrix} \end{array}$$

We call it the ***incidence matrix*** for the function f. ▲

Hands-On Exercise 6.2.6 "Plot" the graph of g in Hands-On Exercise 6.2.5. Also construct its incidence matrix.

△

Summary and Review

- A function f from a set A to a set B (called the domain and the codomain, respectively) is a rule that describes how a value in the codomain B is assigned to an element from the domain A.
- But it is not just any rule; rather, the rule must assign to every element x in the domain a unique value in the codomain.
- This unique value is called the image of x under the function f, and is denoted $f(x)$.
- We use the notation $f\colon A \to B$ to indicate that the name of the function is f, the domain is A, and the codomain is B.
- The graph of a function $f\colon A \to B$ is the collection of all ordered pairs (x,y) from $A \times B$ such that $y = f(x)$.
- The graph of a function may not be a curve, as in the case of a real function. It can be just a collection of points.
- We can also display the images of a function in a table, or represent the function with an incidence matrix.

Exercises 6.2

1. What subset A of $\mathbb{R}$ would you use to make $f\colon A \to \mathbb{R}$ defined by $f(x) = \sqrt{3x-7}$ a well-defined function?

2. What subset A of $\mathbb{R}$ would you use to make

   (a) $g\colon A \to \mathbb{R}$, where $g(x) = \sqrt{(x-3)(x-7)}$

   (b) $h\colon A \to \mathbb{R}$, where $h(x) = \dfrac{x+2}{\sqrt{(x-2)(5-x)}}$

   well-defined functions?

3. Which of these data support a well-defined function from $\{1,2,3,4\}$ to $\{1,2,3,4\}$? Explain.

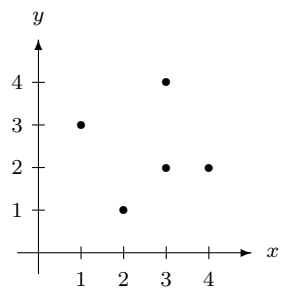
x	1	2	3
$f(x)$	3	4	2

x	1	2	3	4
$g(x)$	2	4	3	2

x	1	2	3	3	4
$h(x)$	2	4	3	2	3

4. Which of the following are the graphical representation or incidence matrix of well-defined functions from $\{1,2,3,4\}$ to $\{1,2,3,4\}$? Explain.

f:



g:
$$\begin{array}{c} \\ 1 \\ 2 \\ 3 \\ 4 \end{array} \begin{array}{c} \begin{array}{cccc} 1 & 2 & 3 & 4 \end{array} \\ \begin{pmatrix} 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 \end{pmatrix} \end{array}$$

5. Determine whether these are well-defined functions. Explain.

   (a) $f\colon \mathbb{R} \to \mathbb{R}$, where $f(x) = \dfrac{3}{x^2+5}$.

   (b) $g\colon (5,\infty) \to \mathbb{R}$, where $g(x) = \dfrac{7}{\sqrt{x-4}}$.

   (c) $h\colon \mathbb{R} \to \mathbb{R}$, where $h(x) = -\sqrt{7-4x+4x^2}$.

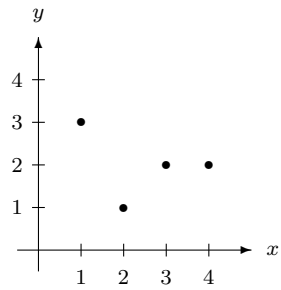
6. Determine whether these are well-defined functions. Explain.

   (a) $s\colon \mathbb{R} \to \mathbb{R}$, where $x^2 + [s(x)]^2 = 9$.

   (b) $t\colon \mathbb{R} \to \mathbb{R}$, where $|x - t(x)| = 4$.

7. Below are the graph of the function p and the incidence matrix for the function q, respectively, from $\{1,2,3,4\}$ to $\{1,2,3,4\}$.

p:



q:
$$\begin{array}{c} \\ 1 \\ 2 \\ 3 \\ 4 \end{array} \begin{array}{c} \begin{array}{cccc} 1 & 2 & 3 & 4 \end{array} \\ \begin{pmatrix} 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 \end{pmatrix} \end{array}$$

Complete the following table:

x	1	2	3	4
$p(x)$				

x	1	2	3	4
$q(x)$				

8. Let T be your family tree that includes your biological mother, your maternal grandmother, your maternal great-grandmother, and so on, and all of their female descendants. Determine which of the following define a function from T to T.
   (a) $h_1: T \to T$, where $h_1(x)$ is the mother of x.
   (b) $h_2: T \to T$, where $h_2(x)$ is x's sister.
   (c) $h_3: T \to T$, where $h_3(x)$ is an aunt of x.
   (d) $h_4: T \to T$, where $h_4(x)$ is the eldest daughter of x's maternal grandmother.
9. For each of the following functions, determine the image of the given x.
   (a) $k_1: \mathbb{N} - \{1\} \to \mathbb{N}, \quad k_1(x) = \text{smallest prime factor of } x, \quad x = 217.$
   (b) $k_2: \mathbb{Z}_{11} \to \mathbb{Z}_{11}, \quad k_2(x) \equiv 3x \pmod{11}, \quad x = 6.$
   (c) $k_3: \mathbb{Z}_{15} \to \mathbb{Z}_{15}, \quad k_3(x) \equiv 3x \pmod{15}, \quad x = 6.$
10. For each of the following functions, determine the images of the given x-values.
   (a) $\ell_1: \mathbb{Z} \to \mathbb{Z}, \quad \ell_1(x) = x \bmod 7, \quad x = 250$, $x = 0$, and $x = -16$.

   Remark: Recall that, without parentheses, the notation "mod" means the binary operation mod.
   (b) $\ell_2: \mathbb{Z} \to \mathbb{Z}, \quad \ell_2(x) = \gcd(x, 24), \quad x = 100$, $x = 0$, and $x = -21$.

6.3 One-to-One Functions

We distinguish two special families of functions: the one-to-one functions and the onto functions. We shall discuss one-to-one functions in this section, and onto functions in the next.

Definition. A function $f: A \to B$ is said to be ***one-to-one*** if

$$x_1 \neq x_2 \Rightarrow f(x_1) \neq f(x_2)$$

To be one-to-one, distinct elements must have distinct images.

for all elements $x_1, x_2 \in A$. A one-to-one function is also called an ***injection***, and we call a function ***injective*** if it is one-to-one. A function that is not one-to-one is referred to as ***many-to-one***. ♢

Any well-defined function is either one-to-one or many-to-one. A function cannot be one-to-many because no element can have multiple images. The difference between one-to-one and many-to-one functions is whether there exist distinct elements that share the same image. There are no repeated images in a one-to-one function.

Example 6.3.1 The ***identity function*** on any nonempty set A

$$i_A: A \to A, \qquad i_A(x) = x,$$

maps any element back to itself. It is clear that all identity functions are one-to-one. ▲

Example 6.3.2 The function $h: A \to A$ defined by $h(x) = c$ for some fixed element $c \in A$, is an example of a ***constant function***. It is a function with only one image. This is the exact opposite of an identity function. It is clearly *not* one-to-one unless $|A| = 1$. ▲

For domains with a small number of elements, one can use inspection on the images to determine if the function is one-to-one. This becomes impossible if the domain contains a larger number of elements.

Use this definition to prove that a function is one-to-one.

In practice, it is easier to use the contrapositive of the definition to test whether a function is one-to-one:

$$\boxed{f(x_1) = f(x_2) \Rightarrow x_1 = x_2.}$$

Example 6.3.3 Is the function $f\colon \mathbb{R} \to \mathbb{R}$ defined by $f(x) = 3x + 2$ one-to-one?

Solution: Assume $f(x_1) = f(x_2)$, which means

$$3x_1 + 2 = 3x_2 + 2.$$

Thus $3x_1 = 3x_2$, which implies that $x_1 = x_2$. Therefore f is one-to-one. ▲

Hands-On Exercise 6.3.1 Determine whether the function $g\colon \mathbb{R} \to \mathbb{R}$ defined by $g(x) = 5 - 7x$ is one-to-one.

△

Hands-On Exercise 6.3.2 Determine whether the function $h\colon [2, \infty) \to \mathbb{R}$ defined by $h(x) = \sqrt{x - 2}$ is one-to-one.

△

Interestingly, sometimes we can use calculus to determine if a real function is one-to-one. A real function f is ***increasing*** if

$$x_1 < x_2 \Rightarrow f(x_1) < f(x_2),$$

and ***decreasing*** if

$$x_1 < x_2 \Rightarrow f(x_1) > f(x_2).$$

Obviously, both increasing and decreasing functions are one-to-one. From calculus, we know that

- A function is increasing over an open interval (a, b) if $f'(x) > 0$ for all $x \in (a, b)$.
- A function is decreasing over an open interval (a, b) if $f'(x) < 0$ for all $x \in (a, b)$.

Therefore, if the derivative of a function is always positive, or always negative, then the function must be one-to-one.

Example 6.3.4 The function $p\colon \mathbb{R} \to \mathbb{R}$ defined by

$$p(x) = 2x^3 - 5$$

is one-to-one, because $p'(x) = 6x^2 > 0$ for any $x \in \mathbb{R}^*$. Likewise, the function $q\colon \left(-\frac{\pi}{2}, \frac{\pi}{2}\right) \to \mathbb{R}$ defined by

$$q(x) = \tan x$$

is also one-to-one, because $q'(x) = \sec^2 x > 0$ for any $x \in \left(-\frac{\pi}{2}, \frac{\pi}{2}\right)$. ▲

Hands-On Exercise 6.3.3 Use both methods to show that the function $k\colon (0,\infty) \to \mathbb{R}$ defined by $k(x) = \ln x$ is one-to-one.

△

Example 6.3.5 The function $h\colon \mathbb{R} \to \mathbb{R}$ given by $h(x) = x^2$ is not one-to-one because some of its images are identical. For example, $h(3) = h(-3) = 9$. It is a many-to-one function. Likewise, the absolute value function $|x|$ is not one-to-one.

You can use counterexamples to show that a function is **not** *one-to-one.*

The functions $p\colon [\,0,\infty) \to \mathbb{R}$ defined by $p(x) = x^2$ and $q\colon [\,0,\infty) \to \mathbb{R}$ defined by $q(x) = |x|$ are one-to-one. Whether a function is one-to-one depends not only on its formula, but also on its domain. Consequently, sometimes we may be able to convert a many-to-one function into a one-to-one function by modifying its domain. ▲

Example 6.3.6 Construct a one-to-one function from $[\,1,3\,]$ to $[\,2,5\,]$.

Solution: There are many possible solutions. In any event, start with a graph. We can use a straight line graph. The domain $[\,1,3\,]$ lies on the x-axis, and the codomain $[\,2,5\,]$ lies on the y-axis. Hence the graph should cover the boxed region in Figure 6.4.

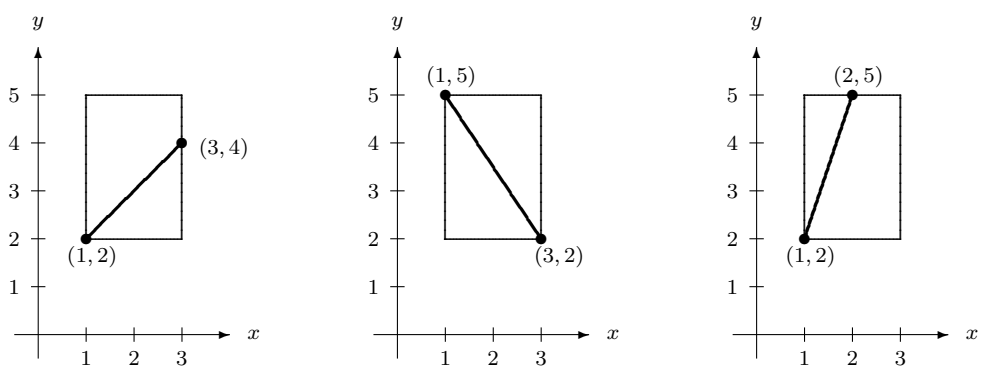


Figure 6.4: Three candidates for one-to-one functions from $[\,1,3\,]$ to $[\,2,5\,]$.

All three graphs do not produce duplicate images. We need to cover all x-values from 1 to 3 in order for the function to be well-defined. This leaves only the first two graphs as legitimate examples.

Every number in the domain must have an image.

To determine the formula for f, we need to derive the equation of the line. Take the first graph as our choice. The line joins the point $(1,2)$ to the point $(3,4)$. Thus, its equation is

$$\frac{y-2}{x-1} = \frac{4-2}{3-1} = 1.$$

The last step is to write the answer in the form of $f(x) = \ldots$. We have to express y in terms of x. We find $y = x + 1$. Hence,

$$f\colon [\,1,3\,] \to [\,2,5\,], \qquad f(x) = x + 1$$

is an example of a one-to-one function. ▲

Hands-On Exercise 6.3.4 Construct a one-to-one function from $[1,3]$ to $[2,5]$ based on the second graph in Example 6.3.6.

△

Hands-On Exercise 6.3.5 Construct a one-to-one function from $[3,8]$ to $[2,5]$.

△

Example 6.3.7 Determine whether the function $g\colon \mathbb{Z}_{43} \to \mathbb{Z}_{43}$ defined by

$$g(x) \equiv 11x - 5 \pmod{43}$$

is one-to-one.

Solution: Assume $g(x_1) = g(x_2)$. This means

$$11x_1 - 5 \equiv 11x_2 - 5 \pmod{43},$$

which implies

$$11x_1 \equiv 11x_2 \pmod{43}.$$

Notice that $4 \cdot 11 = 44 \equiv 1 \pmod{43}$, hence $11^{-1} \equiv 4 \pmod{43}$. Multiplying 4 to both sides of the last congruence yields

$$44x_1 \equiv 44x_2 \pmod{43},$$

which is equivalent to, since $44 \equiv 1 \pmod{43}$,

$$x_1 \equiv x_2 \pmod{43}.$$

Therefore, $x_1 = x_2$ in $\mathbb{Z}_{43}$. This proves that g is one-to-one. ▲

Hands-On Exercise 6.3.6 Is the function $h\colon \mathbb{Z}_{15} \to \mathbb{Z}_{15}$ defined by

$$h(x) \equiv 4x - 11 \pmod{15}$$

a one-to-one function?

△

Hands-On Exercise 6.3.7 Show that the function $k\colon \mathbb{Z}_{15} \to \mathbb{Z}_{15}$ defined by

$$k(x) \equiv 5x - 11 \pmod{15}$$

is not one-to-one by finding $x_1 \neq x_2$ such that $k(x_1) = k(x_2)$.

You can use a counterexample to show that a function is not *one-to-one.*

△

Example 6.3.8 In the last hands-on exercise, we should not rely on the non-existence of 5^{-1} in $\mathbb{Z}_{15}$ to prove that k is not one-to-one. One must consider the interaction between the domain, the codomain, and the definition of the function. For example, despite the fact that 5^{-1} does not exist in $\mathbb{Z}_{15}$, the function $p\colon \mathbb{Z}_3 \to \mathbb{Z}_{15}$ defined by

$$p(x) \equiv 5x - 11 \pmod{15}$$

is one-to-one, because $p(0) = 4$, $p(1) = 9$, and $p(2) = 14$ are distinct images. ▲

The last example illustrates the trickiness in a function with different moduli in its domain and codomain. Use caution when you deal with such functions! Sometimes, infinite sets also pose a challenge. Because there is an infinite supply of elements, we may obtain results that appear to be impossible for finite sets.

Example 6.3.9 The function $f\colon \mathbb{Z} \to \mathbb{Z}$ defined by

$$f(n) = \begin{cases} \frac{n}{2} & \text{if } n \text{ is even} \\ \frac{n+1}{2} & \text{if } n \text{ is odd} \end{cases}$$

is not one-to-one, because, for example, $f(0) = f(-1) = 0$. The function $g\colon \mathbb{Z} \to \mathbb{Z}$ defined by

$$g(n) = 2n$$

is one-to-one, because if $g(n_1) = g(n_2)$, then $2n_1 = 2n_2$ implies that $n_1 = n_2$. ▲

Hands-On Exercise 6.3.8 Show that the function $h\colon \mathbb{Z} \to \mathbb{N}$ defined by

$$h(n) = \begin{cases} 2n + 1 & \text{if } n \geq 0, \\ -2n & \text{if } n < 0, \end{cases}$$

is one-to-one.

△

Example 6.3.10 Let A be the set of all married individuals from a monogamous community who are neither divorced nor widowed. Then the function $s\colon A \to A$ defined by

$$s(x) = \text{ spouse of } x$$

is one-to-one. The reason is, it is impossible to have $x_1 \neq x_2$ and yet $s(x_1) = s(x_2)$. ▲

Summary and Review

- A function f is said to be one-to-one if $f(x_1) = f(x_2) \Rightarrow x_1 = x_2$.
- No two images of a one-to-one function are the same.
- To show that a function f is *not* one-to-one, all we need is to find two different x-values that produce the same image; that is, find $x_1 \neq x_2$ such that $f(x_1) = f(x_2)$.

Exercises 6.3

1. Which of the following functions are one-to-one? Explain.

   (a) $f\colon \mathbb{R} \to \mathbb{R}$, $f(x) = x^3 - 2x^2 + 1$.

   (b) $g\colon [\,2, \infty) \to \mathbb{R}$, $f(x) = x^3 - 2x^2 + 1$.

2. Which of the following functions are one-to-one? Explain.

   (a) $p\colon \mathbb{R} \to \mathbb{R}$, $h(x) = e^{1-2x}$.

   (b) $q\colon \mathbb{R} \to \mathbb{R}$, $p(x) = |1 - 3x|$.

3. Construct a one-to-one function $f\colon (1, 3) \to (2, 5)$ so that $f\colon [\,1, 3) \to [\,2, 5)$ is still one-to-one.

4. Construct a one-to-one function $g\colon [\,2, 5) \to (1, 4\,]$.

5. Determine which of the following are one-to-one functions.

   (a) $f\colon \mathbb{Z} \to \mathbb{Z}$; $\quad f(n) = n^3 + 1$
   (b) $g\colon \mathbb{Q} \to \mathbb{Q}$; $\quad g(x) = n^2$
   (c) $h\colon \mathbb{R} \to \mathbb{R}$; $\quad h(x) = x^3 - x$
   (d) $k\colon \mathbb{R} \to \mathbb{R}$; $\quad k(x) = 5^x$

6. Determine which of the following are one-to-one functions.

   (a) $p\colon \wp(\{1, 2, 3, \ldots, n\}) \to \{0, 1, 2, \ldots, n\}$; $\quad p(S) = |S|$
   (b) $q\colon \wp(\{1, 2, 3, \ldots, n\}) \to \wp(\{1, 2, 3, \ldots, n\})$; $\quad q(S) = \overline{S}$

7. Determine which of the following functions are one-to-one.

   (a) $f_1\colon \{1, 2, 3, 4, 5\} \to \{a, b, c, d\}$; $\quad f_1(1) = b,\ f_1(2) = c,\ f_1(3) = a,\ f_1(4) = a,\ f_1(5) = c$
   (b) $f_2\colon \{1, 2, 3, 4\} \to \{a, b, c, d, e\}$; $\quad f_2(1) = c,\ f_2(2) = b,\ f_2(3) = a,\ f_2(4) = d$
   (c) $f_3\colon \mathbb{Z} \to \mathbb{Z}$; $\quad f_5(n) = -n$
   (d) $f_4\colon \mathbb{Z} \to \mathbb{Z}$; $\quad f_4(n) = \begin{cases} 2n & \text{if } n < 0 \\ -3n & \text{if } n \geq 0 \end{cases}$

8. Determine which of the following functions are one-to-one.

   (a) $g_1\colon \{1, 2, 3, 4, 5\} \to \{a, b, c, d, e\}$; $\quad g_1(1) = b,\ g_1(2) = b,\ g_1(3) = b,\ g_1(4) = a,\ g_1(5) = d$
   (b) $g_2\colon \{1, 2, 3, 4, 5\} \to \{a, b, c, d, e\}$; $\quad g_2(1) = d,\ g_2(2) = b,\ g_2(3) = e,\ g_2(4) = a,\ g_2(5) = c$
   (c) $g_3\colon \mathbb{N} \to \mathbb{N}$; $\quad g_3(n) = \begin{cases} (n+1)/2 & \text{if } n \text{ is odd} \\ n/2 & \text{if } n \text{ is even} \end{cases}$
   (d) $g_4\colon \mathbb{N} \to \mathbb{N}$; $\quad g_4(n) = \begin{cases} n+1 & \text{if } n \text{ is odd} \\ n-1 & \text{if } n \text{ is even} \end{cases}$

9. List all the one-to-one functions from $\{1, 2\}$ to $\{a, b, c, d\}$.

   Hint: List the images of each function.

10. Is it possible to find a one-to-one function from $\{1, 2, 3, 4\}$ to $\{1, 2\}$? Explain.

11. Determine which of the following functions are one-to-one.

(a) $f\colon \mathbb{Z}_{10} \to \mathbb{Z}_{10}$; $h(n) \equiv 3n \pmod{10}$.
(b) $g\colon \mathbb{Z}_{10} \to \mathbb{Z}_{10}$; $g(n) \equiv 5n \pmod{10}$.
(c) $h\colon \mathbb{Z}_{36} \to \mathbb{Z}_{36}$; $h(n) \equiv 3n \pmod{36}$.

12. Determine which of the following functions are one-to-one.

(a) $r\colon \mathbb{Z}_{36} \to \mathbb{Z}_{36}$; $r(n) \equiv 5n \pmod{36}$.
(b) $s\colon \mathbb{Z}_{10} \to \mathbb{Z}_{10}$; $s(n) \equiv n + 5 \pmod{10}$.
(c) $t\colon \mathbb{Z}_{10} \to \mathbb{Z}_{10}$; $t(n) \equiv 3n + 5 \pmod{10}$.

13. Determine which of the following functions are one-to-one.

(a) $\alpha\colon \mathbb{Z}_{12} \to \mathbb{Z}_{7}$; $\alpha(n) \equiv 2n \pmod{7}$.
(b) $\beta\colon \mathbb{Z}_{8} \to \mathbb{Z}_{12}$; $\beta(n) \equiv 3n \pmod{12}$.
(c) $\gamma\colon \mathbb{Z}_{6} \to \mathbb{Z}_{12}$; $\gamma(n) \equiv 2n \pmod{12}$.
(d) $\delta\colon \mathbb{Z}_{12} \to \mathbb{Z}_{36}$; $\delta(n) \equiv 6n \pmod{36}$.

14. Give an example of a one-to-one function f from $\mathbb{N}$ to $\mathbb{N}$ that is not the identity function.

6.4 Onto Functions

One-to-one functions focus on the elements in the domain. We do not want any two of them sharing a common image. Onto functions focus on the codomain. We want to know if it contains elements not associated with any element in the domain.

Definition. A function $f\colon A \to B$ is ***onto*** if, for every element $b \in B$, there exists an element $a \in A$ such that

$$f(a) = b.$$

An onto function is also called a ***surjection***, and we say it is ***surjective***. ♢

Example 6.4.1 The graph of the piecewise-defined functions $h\colon [1,3] \to [2,5]$ defined by

$$h(x) = \begin{cases} 3x - 1 & \text{if } 1 \le x \le 2, \\ -3x + 11 & \text{if } 2 < x \le 3, \end{cases}$$

is displayed on the left in Figure 6.5. It is clearly onto, because, given any $y \in [2,5]$, we can find at least one $x \in [1,3]$ such that $h(x) = y$. Likewise, the function $k\colon [1,3] \to [2,5]$ defined by

$$k(x) = \begin{cases} 3x - 1 & \text{if } 1 \le x \le 2, \\ 5 & \text{if } 2 < x \le 3, \end{cases}$$

is also onto. Its graph is displayed on the right of Figure 6.5. ▲

Hands-On Exercise 6.4.1 The two functions in Example 6.4.1 are onto but not one-to-one. Construct a one-to-one and onto function f from $[1,3]$ to $[2,5]$.

△

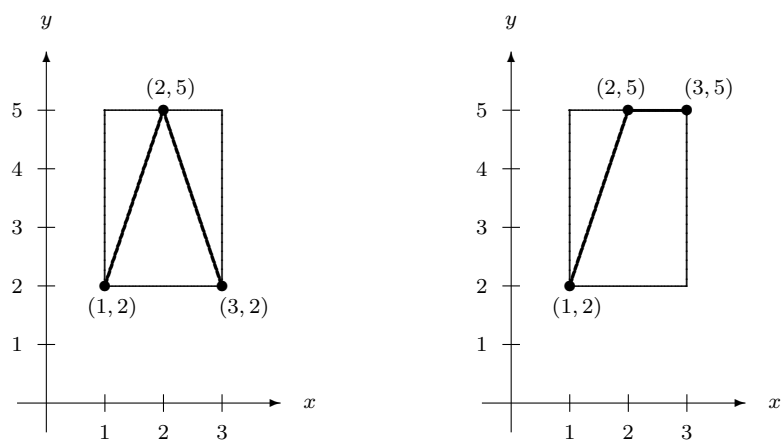


Figure 6.5: Two onto functions from $[1, 3]$ to $[2, 5]$.

Hands-On Exercise 6.4.2 Construct a function $g\colon [1, 3] \to [2, 5]$ that is one-to-one but not onto.

△

Hands-On Exercise 6.4.3 Find a subset B of $\mathbb{R}$ that would make the function $s\colon \mathbb{R} \to B$ defined by $s(x) = x^2$ an onto function.

△

Example 6.4.2 Construct a function $g\colon (5, 8) \to \mathbb{R}$ that is both one-to-one and onto.

Remark. This is a challenging problem. Since the domain is an open interval, a straight line graph does not work, because it will not cover every number in the codomain. ♢

Solution: The solution is based on the observation that the function $h\colon (-\frac{\pi}{2}, \frac{\pi}{2}) \to \mathbb{R}$ defined by $h(x) = \tan x$ is one-to-one and onto. For this to work in this problem, we need to shift and scale the interval $(5, 8)$ to the same size as $(-\frac{\pi}{2}, \frac{\pi}{2})$.

First, we have to shift the center of the interval $(5, 8)$ to the center of the interval $(-\frac{\pi}{2}, \frac{\pi}{2})$. The midpoint of the interval $(5, 8)$ is $\frac{5+8}{2} = \frac{13}{2}$, and the midpoint of $(-\frac{\pi}{2}, \frac{\pi}{2})$ is 0. Hence, we need to shift the interval $(5, 8)$ to the left $\frac{13}{2}$ units. This means we need to use the transformation $x - \frac{13}{2}$. The two endpoints 5 and 8 become $-\frac{3}{2}$ and $\frac{3}{2}$, respectively:

x	5	$\frac{13}{2}$	8
$x - \frac{13}{2}$	$-\frac{3}{2}$	0	$\frac{3}{2}$

After the transformation $x - \frac{13}{2}$, the original interval $(5, 8)$ becomes the interval $(-\frac{3}{2}, \frac{3}{2})$. Next,

we want to stretch the interval $(-\frac{3}{2}, \frac{3}{2})$ into $(-\frac{\pi}{2}, \frac{\pi}{2})$. This calls for a scaling factor of $\frac{\pi}{3}$.

x	5	$\frac{13}{2}$	8
$\frac{\pi}{3}\left(x - \frac{13}{2}\right)$	$-\frac{\pi}{2}$	0	$\frac{\pi}{2}$

Putting these transformations together, we conclude that

$$g(x) = \tan\left[\frac{\pi}{3}\left(x - \frac{13}{2}\right)\right]$$

gives a one-to-one and onto function from $(5, 8)$ to $\mathbb{R}$. ▲

Hands-On Exercise 6.4.4 Construct a function $h\colon (2, 9) \to \mathbb{R}$ that is both one-to-one and onto.

△

In general, how can we tell if a function $f\colon A \to B$ is onto? The key question is: given an element y in the codomain, is it the image of some element x in the domain? If it is, we must be able to find an element x in the domain such that $f(x) = y$. Mathematically, if the rule of assignment is in the form of a computation, then we need to solve the equation $y = f(x)$ for x. If we can *always* express x in terms of y, and if the resulting x-value is in the domain, the function is onto.

Example 6.4.3 Is the function $p\colon \mathbb{R} \to \mathbb{R}$ defined by $p(x) = 3x^2 - 4x + 5$ onto?

Solution 1: Let $y = 3x^2 - 4x + 5$, we want to know if we can always express x in terms of y. Rearranging the equation, we find

$$3x^2 - 4x + (5 - y) = 0.$$

We want this equation to be solvable over $\mathbb{R}$, that is, we want its solutions to be real. This requires its discriminant to be nonnegative. So we need

$$(-4)^2 - 4 \cdot 3 \cdot (5 - y) = 12y - 44 \geq 0.$$

We have real solutions only when $y \geq \frac{11}{3}$. This means, when $y < \frac{11}{3}$, we cannot find an x-value such that $p(x) = y$. Therefore, p is not onto. ▲

Solution 2: By completing the square, we find

$$p(x) = 3x^2 - 4x + 5 = 3\left(x - \frac{2}{3}\right)^2 + \frac{11}{3} \geq \frac{11}{3}.$$

Since $p(x) \not< \frac{11}{3}$, it is clear that p is not onto. ▲

Hands-On Exercise 6.4.5 The function $g\colon \mathbb{R} \to \mathbb{R}$ is defined as $g(x) = 3x + 11$. Prove that it is onto.

△

Example 6.4.4 Is the function $p\colon \mathbb{R} \to \mathbb{R}$ defined by

$$p(x) = \begin{cases} 4x+1 & \text{if } x \le 3 \\ \frac{1}{2}\,x & \text{if } x > 3 \end{cases}$$

an onto function?

Solution: The graphs $y = 4x+1$ and $y = \frac{1}{2}\,x$ are both increasing. For $x \le 3$, the y-values cover the range $(-\infty, 13)$, and for $x > 3$, the y-values cover the range $\left(\frac{3}{2}, \infty\right)$. Since these two y-ranges overlap, all the y-values are being covered by the images. Therefore, p is onto. ▲

Hands-On Exercise 6.4.6 Determine whether

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

is an onto function.

△

Example 6.4.5 Consider the function $g\colon \mathbb{Z}_{43} \to \mathbb{Z}_{43}$ defined by

$$g(x) \equiv 11x - 5 \pmod{43}.$$

Let

$$y = g(x) \equiv 11x - 5 \pmod{43},$$

then

$$x \equiv 11^{-1}(y+5) \equiv 4(y+5) \pmod{43}.$$

This shows that g is onto. ▲

Hands-On Exercise 6.4.7 Show that the function $h\colon \mathbb{Z}_{23} \to \mathbb{Z}_{23}$ defined by $h(x) \equiv 5x + 8 \pmod{23}$ is onto.

△

Example 6.4.6 Is the function $u\colon \mathbb{Z} \to \mathbb{Z}$ defined by

$$u(n) = \begin{cases} 2n & \text{if } n \ge 0 \\ -n & \text{if } n < 0 \end{cases}$$

one-to-one? Is it onto?

Solution: Since $u(-2) = u(1) = 2$, the function u is not one-to-one. Since $u(n) \ge 0$ for any $n \in \mathbb{Z}$, the function u is not onto. ▲

Hands-On Exercise 6.4.8 Is the function $v\colon \mathbb{N} \to \mathbb{N}$ defined by $v(n) = n + 1$ onto? Explain.

△

Example 6.4.7 The function s in Example 6.3.10 is both one-to-one and onto. It provides a one-to-one correspondence between the elements of A by matching a married individual to his/her spouse. ▲

Hands-On Exercise 6.4.9 Is the function h_1 in Exercises 6.2, Problem 8, an onto function? Explain.

△

Summary and Review

- A function $f\colon A \to B$ is onto if, for every element $b \in B$, there exists an element $a \in A$ such that $f(a) = b$.
- To show that f is an onto function, set $y = f(x)$, and solve for x, or show that we can always express x in terms of y for any $y \in B$.
- To show that a function is *not* onto, all we need is to find an element $y \in B$, and show that no x-value from A would satisfy $f(x) = y$.

Exercises 6.4

1. Which of the following functions are onto? Explain!

   (a) $f\colon \mathbb{R} \to \mathbb{R}$, $f(x) = x^3 - 2x^2 + 1$.

   (b) $g\colon [\,2, \infty) \to \mathbb{R}$, $g(x) = x^3 - 2x^2 + 1$.

2. Which of the following functions are onto? Explain!

   (a) $p\colon \mathbb{R} \to \mathbb{R}$, $p(x) = e^{1-2x}$.

   (b) $q\colon \mathbb{R} \to \mathbb{R}$, $q(x) = |1 - 3x|$.

3. Construct a one-to-one function $f\colon [1, 3] \to [2, 5]$ that is not onto.

4. Construct an onto function $g\colon [\,2, 5) \to (1, 4\,]$ that is not one-to-one.

5. Determine which of the following are onto functions.

   (a) $f\colon \mathbb{Z} \to \mathbb{Z}$; $\quad f(n) = n^3 + 1$
   (b) $g\colon \mathbb{Q} \to \mathbb{Q}$; $\quad g(x) = n^2$
   (c) $h\colon \mathbb{R} \to \mathbb{R}$; $\quad h(x) = x^3 - x$
   (d) $k\colon \mathbb{R} \to \mathbb{R}$; $\quad k(x) = 5^x$

6. Determine which of the following are onto functions.

   (a) $p\colon \wp(\{1, 2, 3, \ldots, n\}) \to \{0, 1, 2, \ldots, n\}$; $\quad p(S) = |S|$
   (b) $q\colon \wp(\{1, 2, 3, \ldots, n\}) \to \wp(\{1, 2, 3, \ldots, n\})$; $\quad q(S) = \overline{S}$

7. Determine which of the following functions are onto.

   (a) $f_1\colon \{1, 2, 3, 4, 5\} \to \{a, b, c, d\}$; $\quad f_1(1) = b,\ f_1(2) = c,\ f_1(3) = a,\ f_1(4) = a,\ f_1(5) = c$
   (b) $f_2\colon \{1, 2, 3, 4\} \to \{a, b, c, d, e\}$; $\quad f_2(1) = c,\ f_2(2) = b,\ f_2(3) = a,\ f_2(4) = d$

(c) $f_3\colon \mathbb{Z} \to \mathbb{Z}$; $\quad f_5(n) = -n$

(d) $f_4\colon \mathbb{Z} \to \mathbb{Z}$; $\quad f_4(n) = \begin{cases} 2n & \text{if } n < 0 \\ -3n & \text{if } n \geq 0 \end{cases}$

8. Determine which of the following functions are onto.

   (a) $g_1\colon \{1,2,3,4,5\} \to \{a,b,c,d,e\}$; $\quad g_1(1) = b,\ g_1(2) = b,\ g_1(3) = b,\ g_1(4) = a,\ g_1(5) = d$
   (b) $g_2\colon \{1,2,3,4,5\} \to \{a,b,c,d,e\}$; $\quad g_2(1) = d,\ g_2(2) = b,\ g_2(3) = e,\ g_2(4) = a,\ g_2(5) = c$
   (c) $g_3\colon \mathbb{N} \to \mathbb{N}$; $\quad g_3(n) = \begin{cases} (n+1)/2 & \text{if } n \text{ is odd} \\ n/2 & \text{if } n \text{ is even} \end{cases}$
   (d) $g_4\colon \mathbb{N} \to \mathbb{N}$; $\quad g_4(n) = \begin{cases} n+1 & \text{if } n \text{ is odd} \\ n-1 & \text{if } n \text{ is even} \end{cases}$

9. Is it possible for a function from $\{1,2\}$ to $\{a,b,c,d\}$ to be onto? Explain.

10. List all the onto functions from $\{1,2,3,4\}$ to $\{a,b\}$?

    Hint: List the images of each function.

11. Determine which of the following functions are onto.

    (a) $f\colon \mathbb{Z}_{10} \to \mathbb{Z}_{10}$; $\quad h(n) \equiv 3n \pmod{10}$.
    (b) $g\colon \mathbb{Z}_{10} \to \mathbb{Z}_{10}$; $\quad g(n) \equiv 5n \pmod{10}$.
    (c) $h\colon \mathbb{Z}_{36} \to \mathbb{Z}_{36}$; $\quad h(n) \equiv 3n \pmod{36}$.

12. Determine which of the following functions are onto.

    (a) $r\colon \mathbb{Z}_{36} \to \mathbb{Z}_{36}$; $\quad r(n) \equiv 5n \pmod{36}$.
    (b) $s\colon \mathbb{Z}_{10} \to \mathbb{Z}_{10}$; $\quad s(n) \equiv n + 5 \pmod{10}$.
    (c) $t\colon \mathbb{Z}_{10} \to \mathbb{Z}_{10}$; $\quad t(n) \equiv 3n + 5 \pmod{10}$.

13. Determine which of the following functions are onto.

    (a) $\alpha\colon \mathbb{Z}_{12} \to \mathbb{Z}_7$; $\quad \alpha(n) \equiv 2n \pmod{7}$.
    (b) $\beta\colon \mathbb{Z}_8 \to \mathbb{Z}_{12}$; $\quad \beta(n) \equiv 3n \pmod{12}$.
    (c) $\gamma\colon \mathbb{Z}_6 \to \mathbb{Z}_{12}$; $\quad \gamma(n) \equiv 2n \pmod{12}$.
    (d) $\delta\colon \mathbb{Z}_{12} \to \mathbb{Z}_{36}$; $\quad \delta(n) \equiv 6n \pmod{36}$.

14. Give an example of a function $f\colon \mathbb{N} \to \mathbb{N}$ that is

    (a) neither one-to-one nor onto
    (b) one-to-one but not onto
    (c) onto but not one-to-one
    (d) both one-to-one and onto

6.5 Properties of Functions

In this section, we will study some properties of functions. To facilitate our discussion, we need to introduce some notations. Some students may find them confusing and difficult to use. Besides memorizing the definitions, try to understand what they really mean.

Definition. Given a function $f\colon A \to B$, and $C \subseteq A$, the ***image of C under f*** is defined as

$$f(C) = \{f(x) \mid x \in C\}.$$

In words, $f(C)$ is the set of all the images of the elements of C. $\diamondsuit$

Remark. A few remarks about the definition:

1. It is about the image of a *subset* C of the domain of A. Do not confuse it with the image of an *element* x from A.
2. Therefore, do not merely say "the image." Be specific: the image of an element, or the image of a subset.
3. Better yet: include the notation $f(x)$ or $f(C)$ in the discussion.
4. While $f(x)$ is an *element* in the codomain, $f(C)$ is a *subset* of the codomain.
5. Perhaps, the most important thing to remember is:

$$\boxed{\text{If } y \in f(C), \text{ then } y \in B, \text{ and there exists an } x \in C \text{ such that } f(x) = y.}$$

Remember this!

This key observation is often what we need to start a proof with. ♢

Definition. Let $f\colon A \to B$ be a function. The ***image*** or ***range*** of f, denoted $\operatorname{im} f$, is defined as the set $f(A)$. Hence, $\operatorname{im} f$ is the set of all possible images that f can assume. ♢

The definition implies that a function $f\colon A \to B$ is onto if $\operatorname{im} f = B$. Unfortunately, this observation is of limited use, because it is not always easy to find $\operatorname{im} f$.

Example 6.5.1 For the function $f\colon \mathbb{R} \to \mathbb{R}$ defined by

$$f(x) = x^2,$$

we find $\operatorname{im} f = [0, \infty)$. We also have, for example, $f\big([2, \infty)\big) = [4, \infty)$. It is clear that f is neither one-to-one nor onto. ▲

Example 6.5.2 For the function $g\colon \mathbb{Z} \to \mathbb{Z}$ defined by

$$g(n) = n + 3,$$

we find $\operatorname{im} g = \mathbb{Z}$, and $g(\mathbb{N}) = \{4, 5, 6, \ldots\}$. The function g is both one-to-one and onto. ▲

Hands-On Exercise 6.5.1 The function $p\colon \mathbb{R} \to \mathbb{R}$ is defined as $p(x) = 3x + 11$. Find $p(\mathbb{R}^+)$ and $\operatorname{im} p$.

△

Hands-On Exercise 6.5.2 The function $q\colon \mathbb{R} \to \mathbb{R}$ is defined as $q(x) = x^2 - x - 7$. Find $\operatorname{im} q$.

△

Example 6.5.3 The function $h\colon \mathbb{Z}_{15} \to \mathbb{Z}_{15}$ is defined by

$$h(x) \equiv 5x - 11 \pmod{15}.$$

From the tabulated data

x	0	1	2	3	4	5	6	$\cdots$	14
$f(x)$	4	9	14	4	9	14	4	$\cdots$	14

it becomes clear that the images repeat the pattern 4, 9, 14 five times. Therefore, we determine that $\operatorname{im} h = \{4, 9, 14\}$. ▲

Hands-On Exercise 6.5.3 Determine $h(\{0, 3, 4\})$, where h is defined in Example 6.5.3.

△

Example 6.5.4 Determine $f(\{(0, 2), (1, 3)\})$, where the function $f\colon \{0, 1, 2\} \times \{0, 1, 2, 3\} \to \mathbb{Z}$ is defined according to

$$f(a, b) = a + b.$$

Remark: Strictly speaking, we should write $f((a, b))$ because the argument is an ordered pair of the form (a, b). However, we often write $f(a, b)$, because f can be viewed as a two-variable function. The first variable comes from $\{0, 1, 2\}$, the second comes from $\{0, 1, 2, 3\}$, and we add them to form the image. ♢

Solution: Because

$$f(0, 2) = 0 + 2 = 2, \qquad \text{and} \qquad f(1, 3) = 1 + 3 = 4,$$

we determine that $f(\{(0, 2), (1, 3)\}) = \{2, 4\}$. ▲

Hands-On Exercise 6.5.4 Find $\operatorname{im} f$, where f is defined in Example 6.5.4.

△

We are now ready to present the first collection of properties of functions.

Theorem 6.5.1 *Given $f\colon A \to B$, the following properties hold for any $C_1, C_2 \subseteq A$.*

(a) $f(C_1 \cup C_2) = f(C_1) \cup f(C_2)$

(b) $f(C_1 \cap C_2) \subseteq f(C_1) \cap f(C_2)$

(c) $f(C_1 - C_2) \supseteq f(C_1) - f(C_2)$

(d) $C_1 \subseteq C_2 \Rightarrow f(C_1) \subseteq f(C_2)$

Remark. These results provide excellent opportunities to learn how to write mathematical proofs. We only provide the proof of (a) below, and leave the proofs of (b)–(d) as exercises. In (a), we want to establish the equality of two sets. One way to prove that $S = T$ is to show that $S \subseteq T$, and $T \subseteq S$. Now, in order to prove that $S \subseteq T$, we need to show that $z \in S$ implies $z \in$T; to show that $T \subseteq S$, we want to prove that $z \in T$ implies $z \in s$. ♢

Proof of (a): First, we want to show that $f(C_1 \cup C_2) \subseteq f(C_1) \cup f(C_2)$. Let $y \in f(C_1 \cup C_2)$, then there exists $x \in C_1 \cup C_2$ such that $f(x) = y$. Having $x \in C_1 \cup C_2$ means either $x \in C_1$ or $x \in C_2$, so we have to consider two cases.

- If $x \in C_1$, then $f(x) \in f(C_1)$.
- If $x \in C_2$, then $f(x) \in f(C_2)$.

Thus, $y = f(x)$ belongs to either $f(C_1)$ or $f(C_2)$, which means $y = f(x) \in f(C_1) \cup f(C_2)$. This proves that $f(C_1 \cup C_2) \subseteq f(C_1) \cup f(C_2)$.

Next, we want to show that $f(C_1) \cup f(C_2) \subseteq f(C_1 \cup C_2)$. Let $y \in f(C_1) \cup f(C_2)$, then y belongs to either $f(C_1)$ or $f(C_2)$.

- If $y \in f(C_1)$, then there exists $x_1 \in C_1$ such that $f(x_1) = y$.
- If $y \in f(C_2)$, then there exists $x_2 \in C_2$ such that $f(x_2) = y$.

These two possibilities together imply that there exists an element x belonging to either C_1 or C_2, that is, $x \in C_1 \cup C_2$, such that $f(x) = y$. This means $f(x) \in f(C_1 \cup C_2)$. This proves that $f(C_1) \cup f(C_2) \subseteq f(C_1 \cup C_2)$. This concludes the proof of $f(C_1 \cup C_2) = f(C_1) \cup f(C_2)$. ■

Hands-On Exercise 6.5.5 Prove part (b) of Theorem 6.5.1.

△

Remark. Part (b) of Theorem 6.5.1 only gives a subset relationship. The reason is: having $y \in f(C_1)$ and $y \in f(C_2)$ does not necessarily mean that y is the image of the same element. Since f can be many-to-one, it is possible to have $x_1 \in C_1 - C_2$ and $x_2 \in C_2 - C_1$ such that $f(x_1) = f(x_2) = y$. Consider $f\colon \{1, 2, 3\} \to \{a, b\}$ defined by

$$f(1) = f(3) = a, \qquad \text{and} \qquad f(2) = b.$$

If $C_1 = \{1, 2\}$ and $C_2 = \{2, 3\}$, then $f(C_1) = f(C_2) = \{a, b\}$, and

$$f(C_1 \cap C_2) = f(\{2\}) = \{b\} \subset \{a, b\} = f(C_1) \cap f(C_2).$$

Therefore, we can only conclude that $y \in f(C_1 \cap C_2) \Rightarrow y \in f(C_1) \cap f(C_2)$. ♢

Definition. Given a function $f\colon A \to B$, and $D \subseteq B$, the ***preimage of D under f*** is defined as

$$f^{-1}(D) = \{x \in A \mid f(x) \in D\}.$$

Hence, $f^{-1}(D)$ is the set of elements in the domain whose images are in C. The symbol $f^{-1}(D)$ is also pronounced as "f inverse of D." ♢

Remark. Some remarks about the definition:

1. The preimage of D is a subset of the domain A.
2. In particular, the preimage of B is always A.
3. The key thing to remember is:

$$\boxed{\text{If } x \in f^{-1}(D), \text{ then } x \in A, \text{ and } f(x) \in D.}$$

4. It is possible that $f^{-1}(D) = \emptyset$ for some subset D. If this happens, f is not onto.
5. Therefore, f is onto if and only if $f^{-1}(\{b\}) \neq \emptyset$ for every $b \in B$. ♢

Example 6.5.5 If $t: \mathbb{R} \to \mathbb{R}$ is defined by $t(x) = x^2 - 5x + 5$, find $t^{-1}(\{-1\})$.

Solution: We want to find x such that $t(x) = x^2 - 5x + 5 = -1$. Hence, we have to solve the equation

$$0 = x^2 - 5x + 6 = (x-2)(x-3).$$

The solutions are $x = 2$ and $x = 3$. Therefore, $t^{-1}(\{-1\}) = \{2, 3\}$. ▲

Hands-On Exercise 6.5.6 If $k: \mathbb{Q} \to \mathbb{R}$ is defined by $k(x) = x^2 - x - 7$, find $k^{-1}(\{3\})$.

△

Example 6.5.6 For the function $f: \{0,1,2\} \times \{0,1,2,3\} \to \mathbb{Z}$ defined by

$$f(a,b) = a + b,$$

we find

$$\begin{aligned} f^{-1}(\{3\}) &= \{(0,3),(1,2),(2,1)\}, \\ f^{-1}(\{4\}) &= \{(1,3),(2,2)\}. \end{aligned}$$

Since preimages are sets, we need to write the answers in set notation. ▲

Hands-On Exercise 6.5.7 Find $h^{-1}(\{4\})$ and $h^{-1}(\{2\})$, where the function h is defined in Example 6.5.3.

△

Theorem 6.5.2 . *Given $f: A \to B$, and $D_1, D_2 \subseteq B$, the following properties hold.*

(a) $f^{-1}(D_1 \cup D_2) = f^{-1}(D_1) \cup f^{-1}(D_2)$

(b) $f^{-1}(D_1 \cap D_2) = f^{-1}(D_1) \cap f^{-1}(D_2)$

(c) $f^{-1}(D_1 - D_2) = f^{-1}(D_1) - f^{-1}(D_2)$

(d) $D_1 \subseteq D_2 \Rightarrow f^{-1}(D_1) \subseteq f^{-1}(D_2)$

Proof of (a): First, we want to prove that $f^{-1}(D_1 \cup D_2) \subseteq f^{-1}(D_1) \cup f^{-1}(D_2)$. Let $x \in f^{-1}(D_1 \cup D_2)$, then $f(x) \in D_1 \cup D_2$. This means either $f(x) \in D_1$ or $f(x) \in D_2$.

- If $f(x) \in D_1$, then $x \in f^{-1}(D_1)$.
- If $f(x) \in D_2$, then $x \in f^{-1}(D_2)$.

Since x belongs to either $f^{-1}(D_1)$ or $f^{-1}(D_2)$, we determine that $x \in f^{-1}(D_1) \cup f^{-1}(D_2)$. Therefore, $f^{-1}(D_1 \cup D_2) \subseteq f^{-1}(D_1) \cup f^{-1}(D_2)$.

Next, we want to prove that $f^{-1}(D_1) \cup f^{-1}(D_2) \subseteq f^{-1}(D_1 \cup D_2)$. Let $x \in f^{-1}(D_1) \cup f^{-1}(D_2)$. Then x belongs to either $f^{-1}(D_1)$ or $x \in f^{-1}(D_2)$.

- If $x \in f^{-1}(D_1)$, then $f(x) \in D_1$.
- If $x \in f^{-1}(D_2)$, then $f(x) \in D_2$.

Hence, $f(x)$ belongs to either D_1 or D_2, which means $f(x) \in D_1 \cup D_2$. Thus, $x \in f^{-1}(D_1 \cup D_2)$. We have proved that $f^{-1}(D_1) \cup f^{-1}(D_2) \subseteq f^{-1}(D_1 \cup D_2)$. Together with $f^{-1}(D_1 \cup D_2) \subseteq f^{-1}(D_1) \cup f^{-1}(D_2)$, we conclude that $f^{-1}(D_1 \cup D_2) = f^{-1}(D_1) \cup f^{-1}(D_2)$. ■

Hands-On Exercise 6.5.8 Prove part (b) of Theorem 6.5.2.

△

Whether a function $f: A \to B$ is one-to-one or onto can be determined by the cardinality of the preimages.

- f is one-to-one if and only if $|f^{-1}(\{b\})| \leq 1$ for every $b \in B$.
- f is onto if and only if $|f^{-1}(\{b\})| \geq 1$ for every $b \in B$.

If A and B are *finite* sets, then

- $|A| \leq |B|$ if f is one-to-one, and
- $|A| \geq |B|$ if f is onto.

In particular, if f is one-to-one and onto, we have $|A| = |B|$.

Example 6.5.7 A function $f: \mathbb{Z}_{14} \to \mathbb{Z}_{10}$ cannot be one-to-one because in order for it to be one-to-one, we need 14 distinct images. Since the codomain has only 10 elements, it is impossible for it to come up with 14 different images.

Likewise, a function $g: \mathbb{Z}_{23} \to \mathbb{Z}_{57}$ cannot be onto because the domain has 23 elements, hence, we can have at most 23 different images. But the codomain has 57 elements, therefore, some of its elements must be left unused. ▲

Example 6.5.8 Consider the function $h: \mathbb{Z}_{23} \to \mathbb{Z}_{57}$ defined by

$$h(x) \equiv 43x \pmod{57}.$$

If $y \equiv 43x \pmod{57}$, then, since $43^{-1} \equiv 4 \pmod{57}$, we find, in $\mathbb{Z}_{23}$,

$$x = 43^{-1}y = 4y.$$

Since we can also express x in terms of y, we declare that f is onto. Yet, we have learned from the previous example that f cannot be onto. Is there any contradiction?

Be sure you know which modulus you are using.

Solution: There is an error in the argument. We should have said

$$x \equiv 43^{-1}y \equiv 4y \pmod{57}.$$

Since x is reduced modulo 57, its value may exceed 23. If this happens, $x \notin \mathbb{Z}_{23}$. For example, if $y = 11$, we would have $x = 44 \notin \mathbb{Z}_{23}$. Even if we reduce 44 modulo 23, we obtain $x \equiv 21 \pmod{23}$, we would have

$$43 \cdot 21 \equiv 48 \not\equiv 11 \pmod{57}.$$

So it is still not the correct preimage. This example again illustrates the importance of taking caution when a function involves different moduli in its domain and codomain. ▲

Summary and Review

- Given a function $f\colon A \to B$, the image of $C \subseteq A$ is defined as $f(C) = \{f(x) \mid x \in C\}$.
- If $y \in f(C)$, then $y \in B$, and there exists an $x \in C$ such that $f(x) = y$.
- See Theorem 6.5.1 for a list of properties of the image of a set.
- The preimage of $D \subseteq B$ is defined as $f^{-1}(D) = \{x \in A \mid f(x) \in D\}$.
- If $x \in f^{-1}(D)$, then $x \in A$, and $f(x) \in D$.
- See Theorem 6.5.2 for a list of properties of the preimage of a set.

Exercises 6.5

1. For each of the following functions, find the image of C, and the preimage of D.
   (a) $f_1\colon \{1,2,3,4,5\} \to \{a,b,c,d\}$; $f_1(1) = b$, $f_1(2) = c$, $f_1(3) = a$, $f_1(4) = a$, $f_1(5) = c$; $C = \{1,3\}$, $D = \{a.c\}$.
   (b) $f_2\colon \{1,2,3,4\} \to \{a,b,c,d,e\}$; $f_2(1) = c$, $f_2(2) = b$, $f_2(3) = a$, $f_2(4) = d$; $C = \{1,3\}$, $D = \{b,d\}$.
   (c) $f_3\colon \{1,2,3,4,5\} \to \{a,b,c,d,e\}$; $f_3(1) = b$, $f_3(2) = b$, $f_3(3) = b$, $f_3(4) = a$, $f_3(5) = d$; $C = \{1,3,5\}$, $D = \{c\}$.
   (d) $f_4\colon \{1,2,3,4,5\} \to \{a,b,c,d,e\}$; $f_4(1) = d$, $f_4(2) = b$, $f_4(3) = e$, $f_4(4) = a$, $f_4(5) = c$; $C = \{3\}$, $D = \{c\}$.

2. For each of the following functions, find the image of C, and the preimage of D.
   (a) $f_5\colon \mathbb{Z} \to \mathbb{Z}$; $f_5(n) = -n$; $C = 2\mathbb{Z}$, $D = \mathbb{N}$.
   (b) $f_6\colon \mathbb{Z} \to \mathbb{Z}$; $f_6(n) = \begin{cases} 2n & \text{if } n < 0, \\ -3n & \text{if } n \geq 0; \end{cases}$ $C = \mathbb{N}$, $D = 2\mathbb{Z}$.
   (c) $f_7\colon \mathbb{N} \to \mathbb{N}$; $f_7(n) = \begin{cases} (n+1)/2 & \text{if } n \text{ is odd}, \\ n/2 & \text{if } n \text{ is even}; \end{cases}$ $C = D = 2\mathbb{N}$.
   (d) $f_8\colon \mathbb{N} \to \mathbb{N}$; $f_8(n) = \begin{cases} n+1 & \text{if } n \text{ is odd}, \\ n-1 & \text{if } n \text{ is even}; \end{cases}$ $C = D = 2\mathbb{N}$.

3. The function $s\colon \mathbb{Z}_{12} \to \mathbb{Z}_{12}$ is defined as

$$s(x) \equiv 4x + 7 \pmod{12}.$$

   (a) Find $s(\{2,5,7\})$.
   (b) Find $s^{-1}(\{2,5,7\})$.
   (c) Find $\operatorname{im} s$.

4. The function $t\colon \mathbb{Z}_{15} \to \mathbb{Z}_{15}$ is defined as

$$t(x) \equiv 3x^2 - 5 \pmod{15}.$$

   (a) Find $t(\{2, 3, 5, 13\})$.
   (b) Find $t^{-1}(\{1, 5, 7\})$.
   (c) Find $\operatorname{im} t$.

5. The function $u\colon \mathbb{R} \to \mathbb{R}$ is defined as $u(x) = 3x + 11$, and the function $v\colon \mathbb{Z} \to \mathbb{R}$ is defined as $v(x) = 3x + 11$.

   (a) Find $u([3, 5))$ and $v(\{3, 4, 5\})$.
   (b) Find $u^{-1}((2, 7])$ and $v^{-1}((2, 7])$.

6. Is the function $h\colon \mathbb{Z} \to \mathbb{Z}$ defined by

$$h(n) = \begin{cases} 2n & \text{if } n \geq 0 \\ -n & \text{if } n < 0 \end{cases}$$

   one-to-one? Is it onto?

7. Define the $r\colon \mathbb{Z} \times \mathbb{Z} \to \mathbb{Q}$ according to $r(m, n) = 3^m 5^n$.

   (a) Find $r(\{1, 2, 3\} \times \{-1, 0, 1\})$.
   (b) Find $r^{-1}\left(\left\{\frac{25}{27}\right\}\right)$.
   (c) Find $r^{-1}(D)$, where $D = \{3, 9, 27, 81, \ldots\}$.

8. Define the function $p\colon \mathbb{Z} \times \mathbb{Z} \to \mathbb{Z}$ according to $p(x, y) = 12x + 15y$.

   (a) Find $p^{-1}(\{18\})$. You may use the set-builder notation to describe your answer.
   (b) Find $\operatorname{im} p$.

9. The sum of the entries in a particular row in a matrix is called a row sum, and the sum of the entries in a particular column is called a column sum. Discuss how can we use the row sums and column sums of the incidence matrix of a function to determine if the function is well-defined, one-to-one, and onto.

10. Below is the incidence matrix of the function $f\colon \{a, b, c, d, e\} \to \{\alpha, \beta, \gamma, \delta, \epsilon\}$:

$$\begin{array}{c} \\ a \\ b \\ c \\ d \\ e \end{array} \begin{array}{c} \begin{array}{ccccc} \alpha & \beta & \gamma & \delta & \epsilon \end{array} \\ \begin{pmatrix} 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 1 & 0 \\ 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 \end{pmatrix} \end{array}$$

   (a) Find $f(\{a, d, e\})$.
   (b) Find $f^{-1}(\{\alpha, \beta, \epsilon\})$.
   (c) Find $\operatorname{im} f$.

11. Consider the function h_1 defined in Problem 8a in Exercises 6.2. What is $h_1^{-1}(\{m\})$, if m represents your mother?

12. Let S denote the maternal family tree, that includes you, your mother, your maternal grandmother, your maternal great-grandmother, and so on. Define a function $M\colon S \to S$ by letting $M(x)$ be the mother of x. Determine $\operatorname{im} M$.

13. Prove part (c) of Theorem 6.5.1.

14. Prove part (c) of Theorem 6.5.2.

15. (a) Prove part (d) of Theorem 6.5.1.

    (b) Prove part (d) of Theorem 6.5.2.

16. Construct an example of a function $f\colon A \to B$, and $C_1, C_2 \subseteq A$ such that $f(C_1 - C_2) \supsetneq f(C_1) - f(C_2)$. See part (c) of Theorem 6.5.1.

17. Given a function $f\colon A \to B$, and $C \subset A$, since $f(C)$ is a subset of B, the preimage of this subset is indicated by the notation $f^{-1}(f(C))$. Consider the function $f\colon \mathbb{Z} \to \mathbb{Z}$ defined by $f(x) = x^2$, and $C = \{0, 1, 2, 3\}$.

    (a) Find $f(C)$.

    (b) Find $f^{-1}(f(C))$.

18. Prove that $C \subseteq f^{-1}(f(C))$ for any function $f\colon A \to B$, and $C \subseteq A$.

6.6 Inverse Functions

A ***bijection*** is a function that is both one-to-one and onto. Naturally, if a function is a bijection, we say that it is ***bijective***. If a function $f\colon A \to B$ is a bijection, we can define another function g that essentially reverses the assignment rule associated with f. Then, applying the function g to any element y from the codomain B, we are able to obtain an element x from the domain A such that $f(x) = y$. Let us refine this idea into a more concrete definition.

Definition. Let $f\colon A \to B$ be a bijective function. Its ***inverse function*** is the function $f^{-1}\colon B \to A$ with the property that

$$f^{-1}(b) = a \Leftrightarrow b = f(a).$$

Pronounce f^{-1} as "f inverse."

The notation f^{-1} is pronounced as "f inverse." See Figure 6.6 for a pictorial view of an inverse function.

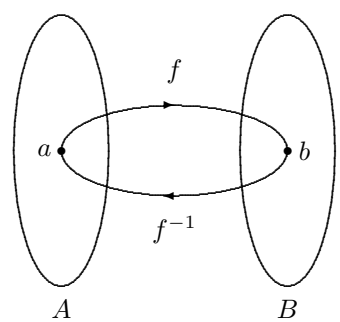


Figure 6.6: The pictorial view of an inverse function.

Why is $f^{-1}\colon B \to A$ a well-defined function? For it to be well-defined, every element $b \in B$ must have a unique image. This means given any element $b \in B$, we must be able to find one and only one element $a \in A$ such that $f(a) = b$. Such an a exists, because f is onto, and there is only one such element a because f is one-to-one. Therefore, f^{-1} is a well-defined function.

If a function f is defined by a computational rule, then the input value x and the output value y are related by the equation $y = f(x)$. In an inverse function, the role of the input and output are switched. Therefore, we can find the inverse function f^{-1} by following these steps:

(i) Interchange the role of x and y in the equation $y = f(x)$. That is, write $x = f(y)$.

(ii) Solve for y. That is, express y in terms of x. The resulting expression is $f^{-1}(x)$.

Be sure to write the final answer in the form $f^{-1}(x) = \ldots$. Do not forget to include the domain and the codomain, and describe them properly.

Example 6.6.1 To find the inverse function of $f\colon \mathbb{R} \to \mathbb{R}$ defined by $f(x) = 2x + 1$, we start with the equation $y = 2x + 1$. Next, interchange x with y to obtain the new equation

$$x = 2y + 1.$$

Solving for y, we find $y = \frac{1}{2}(x - 1)$. Therefore, the inverse function is

$$f^{-1}\colon \mathbb{R} \to \mathbb{R}, \qquad f^{-1}(x) = \frac{1}{2}(x - 1).$$

It is important to describe the domain and the codomain, because they may not be the same as the original function. ▲

Example 6.6.2 The function $s\colon \left[-\frac{\pi}{2}, \frac{\pi}{2}\right] \to [-1, 1]$ defined by $s(x) = \sin x$ is a bijection. Its inverse function is

$$s^{-1}\colon [-1, 1] \to \left[-\frac{\pi}{2}, \frac{\pi}{2}\right], \qquad s^{-1}(x) = \arcsin x.$$

The function $\arcsin x$ is also written as $\sin^{-1} x$, which follows the same notation we use for inverse functions. ▲

Hands-On Exercise 6.6.1 The function $f\colon [-3, \infty) \to [0, \infty)$ is defined as $f(x) = \sqrt{x + 3}$. Show that it is a bijection, and find its inverse function.

△

Hands-On Exercise 6.6.2 Find the inverse function of $g\colon \mathbb{R} \to (0, \infty)$ defined by $g(x) = e^x$.

△

Remark. Exercise caution with the notation. Assume the function $f\colon \mathbb{Z} \to \mathbb{Z}$ is a bijection. The notation $f^{-1}(3)$ means the image of 3 under the inverse function f^{-1}. If $f^{-1}(3) = 5$, we know that $f(5) = 3$. The notation $f^{-1}(\{3\})$ means the preimage of the set $\{3\}$. In this case, we find $f^{-1}(\{3\}) = \{5\}$. The results are essentially the same *if the function is bijective*.

If a function $g\colon \mathbb{Z} \to \mathbb{Z}$ is many-to-one, then it does not have an inverse function. This makes the notation $g^{-1}(3)$ meaningless. Nonetheless, $g^{-1}(\{3\})$ is well-defined, because it means the preimage of $\{3\}$. If $g^{-1}(\{3\}) = \{1, 2, 5\}$, we know $g(1) = g(2) = g(5) = 3$.

In general, $f^{-1}(D)$ means the preimage of the subset D under the function f. Here, the function f can be any function. If f is a bijection, then $f^{-1}(D)$ can also mean the image of the subset D under the inverse function f^{-1}. There is no confusion here, because the results are the same. ◇

Example 6.6.3 The function $f\colon \mathbb{R} \to \mathbb{R}$ is defined as

$$f(x) = \begin{cases} 3x & \text{if } x \le 1, \\ 2x + 1 & \text{if } x > 1. \end{cases}$$

Find its inverse function.

Solution: Since f is a piecewise-defined function, we expect its inverse function to be piecewise-defined as well. First, we need to find the two ranges of input values in f^{-1}. The images for $x \leq 1$ are $y \leq 3$, and the images for $x > 1$ are $y > 3$. Hence, the codomain of f, which becomes the domain of f^{-1}, is split into two halves at 3. The inverse function should look like

$$f^{-1}(x) = \begin{cases} ??? & \text{if } x \leq 3, \\ ??? & \text{if } x > 3. \end{cases}$$

Next, we determine the formulas in the two ranges. We find

$$f^{-1}(x) = \begin{cases} \frac{1}{3}\,x & \text{if } x \leq 3, \\ \frac{1}{2}(x-1) & \text{if } x > 3. \end{cases}$$

The details are left to you as an exercise. ▲

Hands-On Exercise 6.6.3 Find the inverse function of $g\colon \mathbb{R} \to \mathbb{R}$ defined by

$$g(x) = \begin{cases} 3x+5 & \text{if } x \leq 6, \\ 5x-7 & \text{if } x > 6. \end{cases}$$

Be sure you describe g^{-1} properly.

△

Example 6.6.4 The function $g\colon \mathbb{Z}_{10} \to \mathbb{Z}_{10}$ is defined by $g(x) \equiv 7x+2 \pmod{10}$. Find its inverse function.

Solution: From $x = g(y) \equiv 7y+2 \pmod{10}$, we obtain

$$y \equiv 7^{-1}(x-2) \equiv 3(x-2) \pmod{10}.$$

Hence, the inverse function $g^{-1}\colon \mathbb{Z}_{10} \to \mathbb{Z}_{10}$ is defined by $g^{-1}(x) \equiv 3(x-2) \pmod{10}$. ▲

Hands-On Exercise 6.6.4 The function $h\colon \mathbb{Z}_{57} \to \mathbb{Z}_{57}$ defined by $h(x) \equiv 49x-3 \pmod{57}$. Find its inverse function.

△

Example 6.6.5 Define $h\colon \mathbb{Z}_{10} \to \mathbb{Z}_{10}$ according to $h(x) = 2(x+3) \bmod 10$. Does h^{-1} exist?

Solution: Since 2^{-1} does not exist, we suspect the answer is no. In fact, $h(x)$ is always even, and it is easy to verify that $\operatorname{im} h = \{0,2,4,6,8\}$. Since h is not onto, h^{-1} does not exist. ▲

Example 6.6.6 Find the inverse function of $f\colon \mathbb{Z} \to \mathbb{N} \cup \{0\}$ defined by

$$f(n) = \begin{cases} 2n & \text{if } n \geq 0, \\ -2n-1 & \text{if } n < 0. \end{cases}$$

Solution: In an inverse function, the domain and the codomain are switched, so we have to start with $f^{-1}\colon \mathbb{N} \cup \{0\} \to \mathbb{Z}$ before we describe the formula that defines f^{-1}. Writing $n = f(m)$, we find

$$n = \begin{cases} 2m & \text{if } m \geq 0, \\ -2m-1 & \text{if } m < 0. \end{cases}$$

We need to consider two cases.

(i) If $n = 2m$, then n is even, and $m = \frac{n}{2}$.

(ii) If $n = -2m-1$, then n is odd, and $m = -\frac{n+1}{2}$.

Therefore, the inverse function is defined by

$$f^{-1}\colon \mathbb{N} \cup \{0\} \to \mathbb{Z}, \qquad f^{-1}(n) = \begin{cases} \frac{n}{2} & \text{if } n \text{ is even}, \\ -\frac{n+1}{2} & \text{if } n \text{ is odd}. \end{cases}$$

Verify this with some numeric examples. ▲

Hands-On Exercise 6.6.5 The function $f\colon \mathbb{Z} \to \mathbb{N}$ is defined as

$$f(n) = \begin{cases} -2n & \text{if } n < 0, \\ 2n+1 & \text{if } n \geq 0. \end{cases}$$

Find its inverse.

△

Let A and B be finite sets. If there exists a bijection $f\colon A \to B$, then the elements of A and B are in one-to-one correspondence via f. Hence, $|A| = |B|$. This idea provides the basis for some interesting proofs.

Example 6.6.7 Let $A = \{a_1, a_2, \ldots, a_n\}$ be an n-element sets. Recall that the power set $\wp(A)$ contains all the subsets of A, and

$$\{0,1\}^n = \{(b_1, b_2, \ldots, b_n) \mid b_i \in \{0,1\} \text{ for each } i, \text{ where } 1 \leq i \leq n\}.$$

Define $F\colon \wp(A) \to \{0,1\}^n$ according to $F(S) = (x_1, x_2, \ldots, x_n)$, where

$$x_i = \begin{cases} 1 & \text{if } a_i \in S, \\ 0 & \text{if } a_i \notin S. \end{cases}$$

Simply put, $F(S)$ is an ordered n-tuple whose ith entry is either 1 or 0, indicating whether S contains the ith element of A (1 for yes, and 0 for no).

It is clear that F is a bijection. For $n = 8$, we have, for example,

$$F(\{a_2, a_5, a_8\}) = (0,1,0,0,1,0,0,1),$$

and

$$F^{-1}\big((1,1,0,0,0,1,1,0)\big) = \{a_1, a_2, a_6, a_7\}.$$

The function F defines a one-to-one correspondence between the subsets of A and the ordered n-tuples in $\{0,1\}^n$. Since there are two choices for each entry in these ordered n-tuples, we have 2^n such ordered n-tuples. This proves that $|\wp(A)| = 2^n$, that is, A has 2^n subsets. ▲

Hands-On Exercise 6.6.6 Consider the function F defined in Example 6.6.7. Assume $n = 8$. Find $F(\emptyset)$ and $F^{-1}\big((1,0,1,1,1,0,0,0)\big)$.

△

Summary and Review

- A bijection is a function that is both one-to-one and onto.
- The inverse of a bijection $f\colon A \to B$ is the function $f^{-1}\colon B \to A$ with the property that

$$f(x) = y \Leftrightarrow x = f^{-1}(y).$$

- In brief, an inverse function reverses the assignment rule of f. It starts with an element y in the codomain of f, and recovers the element x in the domain of f such that $f(x) = y$.

Exercises 6.6

1. Which of the following functions are bijections? Explain!
   (a) $f\colon \mathbb{R} \to \mathbb{R}$, $f(x) = x^3 - 2x^2 + 1$.
   (b) $g\colon [\,2,\infty) \to \mathbb{R}$, $g(x) = x^3 - 2x^2 + 1$.
   (c) $h\colon \mathbb{R} \to \mathbb{R}$, $h(x) = e^{1-2x}$.
   (d) $p\colon \mathbb{R} \to \mathbb{R}$, $p(x) = |1 - 3x|$.
   (e) $q\colon [\,2,\infty) \to [\,0,\infty)$, $q(x) = \sqrt{x-2}$.
2. For those functions that are not bijections in the last problem, can we modify their codomains to change them into bijections?
3. Let f and g be the functions from $(1,3)$ to $(4,7)$ defined by

$$f(x) = \frac{3}{2}x + \frac{5}{2}, \qquad \text{and} \qquad g(x) = -\frac{3}{2}x + \frac{17}{2}.$$

   Find their inverse functions. Be sure to describe their domains and codomains.
4. Find the inverse function $f\colon \mathbb{R} \to \mathbb{R}$ defined by

$$f(x) = \begin{cases} 3x + 5 & \text{if } x \le 6, \\ 5x - 7 & \text{if } x > 6. \end{cases}$$

   Be sure you describe f^{-1} correctly and properly.
5. The function $g\colon [\,1,3\,] \to [\,4,\,7]$ is defined according to

$$g(x) = \begin{cases} x + 3 & \text{if } 1 \le x < 2, \\ 11 - 2x & \text{if } 2 \le x \le 3. \end{cases}$$

   Find its inverse function. Be sure you describe it correctly and properly.

6. Find the inverse of the function $r\colon (0,\infty) \to \mathbb{R}$ defined by $r(x) = 4 + 3\ln x$.

7. Find the inverse of the function $s\colon \mathbb{R} \to (-\infty, -3)$ defined by $s(x) = 4 - 7e^{2x}$.

8. Find the inverse of each of the following bijections.

   (a) $h\colon \{1,2,3,4,5\} \to \{a,b,c,d,e\}$, $\quad h(1) = e,\ h(2) = c,\ h(3) = b,\ h(4) = a,\ h(5) = d.$

   (b) $k\colon \{1,2,3,4,5\} \to \{1,2,3,4,5\}$, $\quad k(1) = 3,\ k(2) = 1,\ k(3) = 5,\ k(4) = 4,\ k(5) = 2.$

9. Find the inverse of each of the following bijections.

   (a) $u\colon \mathbb{Q} \to \mathbb{Q}$, $\quad u(x) = 3x - 2.$

   (b) $v\colon \mathbb{Q} - \{1\} \to \mathbb{Q} - \{2\}$, $\quad v(x) = \frac{2x}{x-1}.$

   (c) $w\colon \mathbb{Z} \to \mathbb{Z}$, $\quad w(n) = n + 3.$

10. Find the inverse of each of the following bijections.

    (a) $r\colon \mathbb{Z}_{12} \to \mathbb{Z}_{12}$, $\quad r(n) \equiv 7n \pmod{12}.$

    (b) $s\colon \mathbb{Z}_{33} \to \mathbb{Z}_{33}$, $\quad s(n) \equiv 7n + 5 \pmod{33}.$

    (c) $t\colon \mathbb{Z} \to \mathbb{N} \cup \{0\}$, $\quad t(n) = \begin{cases} 2n - 1 & \text{if } n > 0 \\ -2n & \text{if } n \le 0 \end{cases}$

11. The images of the bijection $\alpha\colon \{1,2,3,4,5,6,7,8\} \to \{a,b,c,d,e,f,g,h\}$ are given below.

x	1	2	3	4	5	6	7	8
$\alpha(x)$	g	a	d	h	b	e	f	c

    Find its inverse function.

12. Below is the incidence matrix for the bijection $\beta\colon \{a,b,c,d,e,f\} \to \{x,y,z,u,v,w\}$.

$$\begin{array}{c} \\ a \\ b \\ c \\ d \\ e \\ f \end{array} \begin{array}{c} \begin{array}{cccccc} u & v & w & x & y & z \end{array} \\ \begin{pmatrix} 0 & 1 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 1 & 0 & 0 \end{pmatrix} \end{array}$$

    Find its inverse function.

6.7 Composite Functions

Given functions $f\colon A \to B$ and $g\colon B \to C$, the ***composite function***, $g \circ f$, which is pronounced as "g circle f", is defined as

Write $(g \circ f)(x)$ instead of $g \circ f(x)$.

$$g \circ f\colon A \to C, \qquad (g \circ f)(x) = g(f(x)).$$

The image is obtained in two steps. First, $f(x)$ is obtained. Next, it is passed to g to obtain the final result. It works like connecting two machines to form a bigger one, see Figure 6.7. We can also use an arrow diagram to provide another pictorial view, see Figure 6.8.

Numeric value of $(g \circ f)(x)$ can be computed in two steps. For example, to compute $(g \circ f)(5)$, we first compute the value of $f(5)$, and then the value of $g(f(5))$. To find the algebraic description of $(g \circ f)(x)$, we need to compute and simplify the formula for $g(f(x))$. In this case, it is often easier to start from the "outside" function. More precisely, start with g, and write the

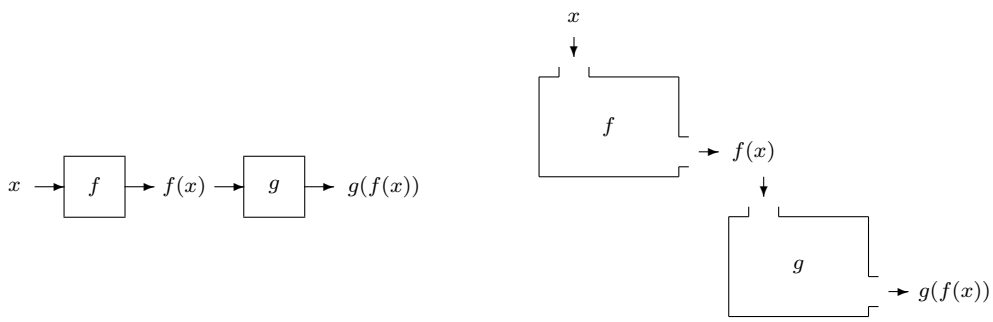


Figure 6.7: A composite function, viewed as input-output machines.

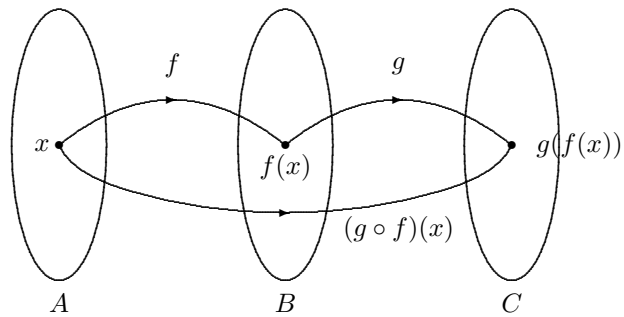


Figure 6.8: Another pictorial view of a composite function.

intermediate answer in terms of $f(x)$, then substitute in the definition of $f(x)$ and simplify the result.

Example 6.7.1 Assume $f, g\colon \mathbb{R} \to \mathbb{R}$ are defined as $f(x) = x^2$, and $g(x) = 3x + 1$. We find

$$(g \circ f)(x) = g(f(x)) = 3[f(x)] + 1 = 3x^2 + 1,$$
$$(f \circ g)(x) = f(g(x)) = [g(x)]^2 = (3x + 1)^2.$$

Therefore,

$$g \circ f\colon \mathbb{R} \to \mathbb{R}, \qquad (g \circ f)(x) = 3x^2 + 1,$$
$$f \circ g\colon \mathbb{R} \to \mathbb{R}, \qquad (f \circ g)(x) = (3x + 1)^2.$$

We note that, in general, $f \circ g \neq g \circ f$. ▲

Hands-On Exercise 6.7.1 If $p, q\colon \mathbb{R} \to \mathbb{R}$ are defined as $p(x) = 2x + 5$, and $q(x) = x^2 + 1$, determine $p \circ q$ and $q \circ p$. Do not forget to describe the domain and the codomain.

△

Hands-On Exercise 6.7.2 The functions $f, g\colon \mathbb{Z}_{12} \to \mathbb{Z}_{12}$ are defined by

$$f(x) \equiv 7x + 2 \pmod{12}, \qquad \text{and} \qquad g(x) \equiv 5x - 3 \pmod{12}.$$

Compute the composite function $f \circ g$.

△

Example 6.7.2 Define $f, g\colon \mathbb{R} \to \mathbb{R}$ as

$$f(x) = \begin{cases} 3x+1 & \text{if } x < 0, \\ 2x+5 & \text{if } x \geq 0, \end{cases}$$

and $g(x) = 5x - 7$. Find $g \circ f$.

Solution: Since f is a piecewise-defined function, we expect the composite function $g \circ f$ is also a piecewise-defined function. It is defined by

$$(g \circ f)(x) = g(f(x)) = 5f(x) - 7 = \begin{cases} 5(3x+1) - 7 & \text{if } x < 0, \\ 5(2x+5) - 7 & \text{if } x \geq 0. \end{cases}$$

After simplification, we find

$$g \circ f\colon \mathbb{R} \to \mathbb{R}, \qquad (g \circ f)(x) = \begin{cases} 15x - 2 & \text{if } x < 0, \\ 10x + 18 & \text{if } x \geq 0. \end{cases}$$

In this example, it is rather obvious what the domain and codomain are. Nevertheless, it is always a good practice to include them when we describe a function. ▲

Hands-On Exercise 6.7.3 The functions $f\colon \mathbb{R} \to \mathbb{R}$ and $g\colon \mathbb{R} \to \mathbb{R}$ are defined by

$$f(x) = 3x + 2, \qquad \text{and} \qquad g(x) = \begin{cases} x^2 & \text{if } x \leq 5, \\ 2x - 1 & \text{if } x > 5. \end{cases}$$

Determine $f \circ g$.

△

The next example further illustrates why it is often easier to start with the outside function g in the derivation of the formula for $g(f(x))$.

Example 6.7.3 The function $p\colon [1, 5] \to \mathbb{R}$ is defined by

$$p(x) = \begin{cases} 2x+3 & \text{if } 1 \leq x < 3, \\ 5x-2 & \text{if } 3 \leq x \leq 5; \end{cases}$$

and the function $q\colon \mathbb{R} \to \mathbb{R}$ by

$$q(x) = \begin{cases} 4x & \text{if } x < 7, \\ 3x & \text{if } x \geq 7. \end{cases}$$

Describe the function $q \circ p$.

Solution: Since

$$(q \circ p)(x) = q(p(x)) = \begin{cases} 4p(x) & \text{if } p(x) < 7, \\ 3p(x) & \text{if } p(x) \geq 7, \end{cases}$$

we have to find out when will $p(x) < 7$, and when will $p(x) \geq 7$, because these conditions determine what we need to do next to continue the computation. Since $p(x)$ is computed in two different ways, we have to analyze two cases.

- ***Case 1***: $1 \leq x < 3$. In this case, $p(x)$ is defined as $2x + 3$. This is an increasing function, hence,

  $$p(x) \geq p(1) = 2 \cdot 1 + 3 = 5, \qquad \text{and} \qquad p(x) < p(3) = 2 \cdot 3 + 3 = 9.$$

  For some xs in this range, we have $p(x) < 7$, but for other x-values, we have $p(x) \geq 7$. We need to know the cut-off point. This happens when $p(x) = 2x + 3 = 7$, that is, when $x = 2$. This leads to two subcases.

  - ***Case 1a***: When $1 \leq x < 2$, we have $p(x) = 2x + 3 < 7$. Thus,

    $$q(p(x)) = 4p(x) = 4(2x + 3) = 8x + 12.$$

  - ***Case 1b***: When $2 \leq x < 3$, we have $p(x) = 2x + 3 \geq 7$. Thus,

    $$q(p(x)) = 3p(x) = 3(2x + 3) = 6x + 9.$$

- ***Case 2***: $3 \leq x \leq 5$. In this case, $p(x)$ is computed as $5x - 2$. This is an increasing function, hence $p(x) \geq p(3) = 5 \cdot 3 - 2 = 13$. Since $p(x)$ is always greater than 7, we find

  $$q(p(x)) = 3p(x) = 3(5x - 2) = 15x - 6.$$

Combining these cases, we determine that the composite function $q \circ p : [1, 5] \to \mathbb{R}$ is defined by

$$(q \circ p)(x) = \begin{cases} 8x + 12 & \text{if } 1 \leq x < 2, \\ 6x + 9 & \text{if } 2 \leq x < 3, \\ 15x - 6 & \text{if } 3 \leq x \leq 5. \end{cases}$$

Study this example again to make sure that you understand it thoroughly. ▲

Hands-On Exercise 6.7.4 The functions $f, g : \mathbb{Z} \to \mathbb{Z}$ are defined by

$$f(n) = \begin{cases} n + 1 & \text{if } n \text{ is even} \\ n - 1 & \text{if } n \text{ is odd} \end{cases} \qquad g(n) = \begin{cases} n + 3 & \text{if } n \text{ is even} \\ n - 7 & \text{if } n \text{ is odd} \end{cases}$$

Determine $f \circ g$.

△

Strictly speaking, $g \circ f$ is well-defined if the codomain of f equals to the domain of g. It is clear that $g \circ f$ is still well-defined if $\operatorname{im} f$ is a subset of the domain of g. Hence, if

$$f : A \to B, \quad g : C \to D,$$

then $g \circ f$ is well-defined if $B \subseteq C$, or more generally, $\operatorname{im} f \subseteq C$.

Example 6.7.4 Let $\mathbb{R}^*$ denote the set of nonzero real numbers. Suppose

$$\begin{aligned} f\colon \mathbb{R}^* \to \mathbb{R}, &\qquad f(x) = 1/x, \\ g\colon \mathbb{R} \to (0,\infty), &\qquad g(x) = 3x^2 + 11. \end{aligned}$$

Determine $f \circ g$ and $g \circ f$. Be sure to specify their domains and codomains.

Solution: To compute $f \circ g$, we start with g, whose domain is $\mathbb{R}$. Hence, $\mathbb{R}$ is the domain of $f \circ g$. The result from g is a number in $(0,\infty)$. The interval $(0,\infty)$ contains positive numbers only, so it is a subset of $\mathbb{R}^*$. Therefore, we can continue our computation with f, and the final result is a number in $\mathbb{R}$. Hence, the codomain of $f \circ g$ is $\mathbb{R}$. The image is computed according to $f(g(x)) = 1/g(x) = 1/(3x^2 + 11)$. We are now ready to present our answer:

$$f \circ g\colon \mathbb{R} \to \mathbb{R}, \qquad (f \circ g)(x) = \frac{1}{3x^2 + 11}.$$

In a similar manner, the composite function $g \circ f\colon \mathbb{R}^* \to (0,\infty)$ is defined as

$$(g \circ f)(x) = \frac{3}{x^2} + 11.$$

Be sure you understand how we determine the domain and codomain of $g \circ f$. ▲

Hands-On Exercise 6.7.5 Let $\mathbb{Z}$ denote the set of integers. Determine $h \circ g$, where

$$\begin{aligned} g\colon \mathbb{Z} \to \mathbb{R}, &\qquad g(x) = \sqrt{|x|}, \\ h\colon \mathbb{R} \to \mathbb{R}, &\qquad h(x) = (x-5)^2. \end{aligned}$$

Is $g \circ h$ well-defined? Explain!

△

As usual, take extra caution with modular arithmetic.

Example 6.7.5 Define $f\colon \mathbb{Z}_{15} \to \mathbb{Z}_{23}$ and $g\colon \mathbb{Z}_{23} \to \mathbb{Z}_{32}$ according to

$$\begin{aligned} f(x) &\equiv 3x + 5 \pmod{23}, \\ g(x) &\equiv 2x + 1 \pmod{32}. \end{aligned}$$

We may expect $g \circ f\colon \mathbb{Z}_{15} \to \mathbb{Z}_{23}$ to be defined as

$$(g \circ f)(x) \equiv 2(3x+5) + 1 \equiv 6x + 11) \pmod{32}.$$

In particular, $(g \circ f)(8) \equiv 59 \equiv 27 \pmod{32}$.

If we perform the computation one step at a time, we find $f(8) \equiv 29 \equiv 6 \pmod{23}$, from which we obtain

$$(g \circ f)(8) = g(f(8)) = g(6) \equiv 13 \pmod{32}.$$

which is not what we have just found. Can you explain why?

Solution: The source of the problem is the different moduli used in f and g. The composite function should be defined as

$$(g \circ f)(x) \equiv 2r + 1 \pmod{32}, \qquad \text{where } r \equiv 3x + 5 \pmod{23}.$$

In a way, this definition forces us to carry out the computation in two steps. Consequently, we will obtain the correct answer $(g \circ f)(8) = 13$. ▲

There is a closed connection between a bijection and its inverse function, from the perspective of composition.

Theorem 6.7.1 *For a bijective function $f: A \to B$,*

$$f^{-1} \circ f = i_A, \qquad \text{and} \qquad f \circ f^{-1} = i_B,$$

where i_A and i_B denote the identity function on A and B, respectively.

Proof: To prove that $f^{-1} \circ f = i_A$, we need to show that $(f^{-1} \circ f)(a) = a$ for all $a \in A$. Assume $f(a) = b$. Then, because f^{-1} is the inverse function of f, we know that $f^{-1}(b) = a$. Therefore,

$$(f^{-1} \circ f)(a) = f^{-1}(f(a)) = f^{-1}(b) = a,$$

which is what we want to show. The proof of $f \circ f^{-1} = i_B$ procceds in the exact same manner, and is omitted here. ■

Example 6.7.6 Show that the functions $f, g: \mathbb{R} \to \mathbb{R}$ defined by $f(x) = 2x + 1$ and $g(x) = \frac{1}{2}(x - 1)$ are inverse functions of each other.

Remark. The problem does not ask you to *find* the inverse function of f or the inverse function of g. Instead, the answers are given to you already. You job is to *verify* that the answers are indeed correct, that the functions are inverse functions of each other. ♢

To **verify** *that g is the inverse function of f, you just have to check whether* **both** *$f \circ g$ and $g \circ f$ are identity functions.*

Solution: Form the two composite functions $f \circ g$ and $g \circ f$, and check whether they *both* equal to the identity function:

$$(f \circ g)(x) = f(g(x)) = 2g(x) + 1 = 2\left[\tfrac{1}{2}(x - 1)\right] + 1 = x,$$
$$(g \circ f)(x) = g(f(x)) = \tfrac{1}{2}\big[f(x) - 1\big] = \tfrac{1}{2}\left[(2x + 1) - 1\right] = x.$$

We conclude that f and g are inverse functions of each other. ▲

Hands-On Exercise 6.7.6 Verify that $f: \mathbb{R} \to \mathbb{R}^+$ defined by $f(x) = e^x$, and $g: \mathbb{R}^+ \to \mathbb{R}$ defined by $g(x) = \ln x$, are inverse functions of each other.

△

Theorem 6.7.2 *Suppose $f: A \to B$ and $g: B \to C$. Let i_A and i_B denote the identity function on A and B, respectively. We have the following results.*

(a) $f \circ i_A = f$ and $i_B \circ f = f$.
(b) If both f and g are one-to-one, then $g \circ f$ is also one-to-one.
(c) If both f and g are onto, then $g \circ f$ is also onto.
(d) If both f and g are bijective, then $g \circ f$ is also bijective. In fact, $(g \circ f)^{-1} = f^{-1} \circ g^{-1}$.

Proof: (a) To show that $f \circ i_A = f$, we need to show that $(f \circ i_A)(a) = f(a)$ for all $a \in A$. This follows from direct computation:

$$(f \circ i_A)(a) = f(i_A(a)) = f(a).$$

The proofs of $i_B \circ f = f$ and (b)–(d) are left as exercises. ■

Example 6.7.7 The converses of (b) and (c) in Theorem 6.7.2 are false, as demonstrated in the functions

$$\begin{array}{ll} f\colon \mathbb{Z} \to \mathbb{Z}, & f(x) = 2x, \\ g\colon \mathbb{Z} \to \mathbb{Z}, & g(x) = \lfloor x/2 \rfloor. \end{array}$$

Here, $g \circ f = i_{\mathbb{Z}}$, so $g \circ f$ is one-to-one, and it is obvious that f is also one-to-one, but g is not one-to-one. It is easy to see that both g and $g \circ f$ are onto, but f is not. ▲

Summary and Review

- The composition of two functions $f\colon A \to B$ and $g\colon B \to C$ is the function $g \circ f\colon A \to C$ defined by $(g \circ f)(x) = g(f(x))$.
- If $f\colon A \to B$ is bijective, then $f^{-1} \circ f = i_A$ and $f \circ f^{-1} = i_B$.
- To check whether $f\colon A \to B$ and $g\colon B \to A$ are inverse of each other, we need to show that
  - $(g \circ f)(x) = g(f(x)) = x$ for all $x \in A$, *and*
  - $(f \circ g)(y) = f(g(y)) = y$ for all $y \in B$.

Exercises 6.7

1. The functions $g, f\colon \mathbb{R} \to \mathbb{R}$ are defined by $f(x) = 5x - 1$ and $g(x) = 3x^2 + 4$. Determine $f \circ g$ and $g \circ f$.

2. The function $h\colon (0, \infty) \to (0, \infty)$ is defined by $h(x) = x + \frac{1}{x}$. Determine $h \circ h$. Simplify your answer as much as possible.

3. The functions $g, f\colon \mathbb{R} \to \mathbb{R}$ are defined by $f(x) = 1 - 3x$ and $g(x) = x^2 + 1$. Evaluate $f(g(f(0)))$.

4. The functions $p\colon (2, 8] \to \mathbb{R}$ and $q\colon \mathbb{R} \to \mathbb{R}$ are defined by

$$p(x) = \begin{cases} 3x - 1 & \text{if } 2 < x \le 4, \\ 17 - 2x & \text{if } 4 < x \le 8, \end{cases}$$

$$q(x) = \begin{cases} 4x - 1 & \text{if } x < 3, \\ 3x + 1 & \text{if } x \ge 3. \end{cases}$$

   Evaluate $q \circ p$.

5. Describe $g \circ f$.
   (a) $f\colon \mathbb{Z} \to \mathbb{N}$, $f(n) = n^2 + 1$; $g\colon \mathbb{N} \to \mathbb{Q}$, $g(n) = \frac{1}{n}$.
   (b) $f\colon \mathbb{R} \to (0, 1)$, $f(x) = 1/(x^2 + 1)$; $g\colon (0, 1) \to (0, 1)$, $g(x) = 1 - x$.
   (c) $f\colon \mathbb{Q} - \{2\} \to \mathbb{Q}^*$, $f(x) = 1/(x - 2)$; $g\colon \mathbb{Q}^* \to \mathbb{Q}^*$, $g(x) = 1/x$.
   (d) $f\colon \mathbb{R} \to [1, \infty)$, $f(x) = x^2 + 1$; $g\colon [1, \infty) \to [0, \infty)$ $g(x) = \sqrt{x - 1}$.
   (e) $f\colon \mathbb{Q} - \{10/3\} \to \mathbb{Q} - \{3\}$, $f(x) = 3x - 7$; $g\colon \mathbb{Q} - \{3\} \to \mathbb{Q} - \{2\}$, $g(x) = 2x/(x - 3)$.

6. Describe $g \circ f$.
   (a) $f\colon \mathbb{Z} \to \mathbb{Z}_5$, $f(n) \equiv n \pmod 5$; $g\colon \mathbb{Z}_5 \to \mathbb{Z}_5$, $g(n) \equiv n + 1 \pmod 5$.
   (b) $f\colon \mathbb{Z}_8 \to \mathbb{Z}_{12}$, $f(n) \equiv 3n \pmod{12}$; $g\colon \mathbb{Z}_{12} \to \mathbb{Z}_6$, $g(n) \equiv 2n \pmod 6$.

7. Describe $g \circ f$.
   (a) $f\colon \{1, 2, 3, 4, 5\} \to \{1, 2, 3, 4, 5\}$, $f(1) = 5$, $f(2) = 3$, $f(3) = 2$, $f(4) = 1$, $f(5) = 4$;
   $g\colon \{1, 2, 3, 4, 5\} \to \{1, 2, 3, 4, 5\}$; $g(1) = 3$, $g(2) = 1$, $g(3) = 5$, $g(4) = 4$, $g(5) = 2$

(b) $f\colon \{a,b,c,d,e\} \to \{1,2,3,4,5\}$; $f(a)=5,\ f(b)=1,\ f(c)=2,\ f(d)=4,\ f(e)=3$;
$g\colon \{1,2,3,4,5\} \to \{a,b,c,d,e\}$; $g(1)=e,\ g(2)=d,\ g(3)=a,\ g(4)=c,\ g(5)=b$

8. Verify that $f, g\colon \mathbb{R} \to \mathbb{R}$ defined by

$$f(x) = \begin{cases} 11-2x & \text{if } x<4 \\ 15-3x & \text{if } x \geq 4 \end{cases} \qquad \text{and} \qquad g(x) = \begin{cases} \frac{1}{3}(15-x) & \text{if } x \leq 3 \\ \frac{1}{2}(11-x) & \text{if } x > 3 \end{cases}$$

are inverse to each other.

9. The functions $f, g\colon \mathbb{Z} \to \mathbb{Z}$ are defined by

$$f(n) = \begin{cases} 2n-1 & \text{if } n \geq 0 \\ 2n & \text{if } n < 0 \end{cases} \qquad \text{and} \qquad g(n) = \begin{cases} n+1 & \text{if } n \text{ is even} \\ 3n & \text{if } n \text{ is odd} \end{cases}$$

Determine $g \circ f$.

10. Define the functions f and g on your maternal family tree (see Problem 8 in Exercises 6.2) according to

$$\begin{aligned} f(x) &= \text{the mother of } x, \\ g(x) &= \text{the eldest daughter of the mother of } x. \end{aligned}$$

Describe these functions.

(a) $f \circ g$ (b) $g \circ f$
(c) $f \circ f$ (d) $g \circ g$

11. Given the bijections f and g, find $f \circ g$, $(f \circ g)^{-1}$ and $g^{-1} \circ f^{-1}$.

(a) $f\colon \mathbb{Z} \to \mathbb{Z}$, $f(n) = n+1$; $g\colon \mathbb{Z} \to \mathbb{Z}$, $g(n) = 2-n$.
(b) $f\colon \mathbb{Q} \to \mathbb{Q}$, $f(x) = 5x$; $g\colon \mathbb{Q} \to \mathbb{Q}$, $g(x) = \frac{x-2}{5}$.
(c) $f\colon \mathbb{Q}-\{2\} \to \mathbb{Q}-\{2\}$, $f(x) = 3x-4$; $g\colon \mathbb{Q}-\{2\} \to \mathbb{Q}-\{2\}$, $g(x) = \frac{x}{x-2}$.
(d) $f\colon \mathbb{Z}_7 \to \mathbb{Z}_7$, $f(n) \equiv 2n+5 \pmod 7$; $g\colon \mathbb{Z}_7 \to \mathbb{Z}_7$, $g(n) \equiv 3n-2 \pmod 7$.

12. Give an example of sets A, B, and C, and of functions $f\colon A \to B$ and $g\colon B \to C$, such that $g \circ f$ and f are both one-to-one, but g is not one-to-one.

13. Prove part (b) of Theorem 6.7.2.

14. Prove part (c) of Theorem 6.7.2.

15. Prove part (d) of Theorem 6.7.2.

16. The incidence matrices for the functions $f\colon \{a,b,c,d,e\} \to \{x,y,z,w\}$ and $g\colon \{x,y,z,w\} \to \{1,2,3,4,5,6\}$ are

$$\begin{array}{c} \\ a \\ b \\ c \\ d \\ e \end{array} \begin{array}{c} \begin{array}{cccc} x & y & z & w \end{array} \\ \begin{pmatrix} 0 & 0 & 1 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 \end{pmatrix} \end{array}, \qquad \text{and} \qquad \begin{array}{c} \\ x \\ y \\ z \\ w \end{array} \begin{array}{c} \begin{array}{cccccc} 1 & 2 & 3 & 4 & 5 & 6 \end{array} \\ \begin{pmatrix} 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 \\ 0 & 1 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 & 0 \end{pmatrix} \end{array},$$

respectively. Construct the incidence matrix for the composition $g \circ f$.

Chapter 5

SWDM 7: Relations

Chapter 7

Relations

7.1 Definition of Relations

Given two nonempty sets A and B, a function tells us how to obtain a unique element $b \in B$ from any element $a \in A$. Very often, we are only interested in some sort of relationship between the elements from these two sets. A familiar example is the equality of two numbers. By saying $a = b$, we are proclaiming that the two numbers a and b are related by being equal in value. Likewise, $a \geq b$ is another example of a relation.

Example 7.1.1 Given $a, b \in \mathbb{R}^*$, declare a and b to be related if they have the same sign. For instance, 7.14 and e are related, so are $-\pi$ and $-\sqrt{2}$. However, 5 and -2 are not. Note that a is related to b implies that b is also related to a. ▲

Example 7.1.2 For $a, b \in \mathbb{R}$, define "a is related to b" if and only if $a < b$. Take note that $3 < 5$, but $5 \not< 3$. This demonstrates that a is related to b does not necessarily imply that b is also related to a. ▲

A relationship can be one-way only.

Example 7.1.3 Let A be a set of students, and let B be a set of courses. Given $a \in A$ and $b \in B$, define "a is related to b" if and only if student a is taking course b. While it could be possible that "John Smith is related to MATH 210" because John is taking MATH 210, it is certainly absurd to say that "MATH 210 is related to John Smith," because it does not make much sense to say that MATH 210 is taking John Smith. This again illustrates that a is related to b does not necessarily imply that b is also related to a. ▲

The two sets need not be the same.

In these examples, we see that when we say "a is related to b," the order in which a and b appear may make a difference. This suggests the following definition.

Definition. A ***relation*** from a set A to a set B is a subset of $A \times B$. Hence, a relation R consists of ordered pairs (a, b), where $a \in A$ and $b \in B$. If $(a, b) \in R$, we say that ***a is related to b***, and we also write $a\,R\,b$. ♢

A relation is a set of ordered pairs, we also write $a\,R\,b$ if a is related to b.

Remark. We can also replace R by a symbol, especially when one is readily available. This is exactly what we do in, for example, $a < b$. To say it is not true that $a < b$, we can write $a \not< b$. Likewise, if $(a, b) \notin R$, then a is not related to b, and we could write $a \not R\, b$. But the slash may not be easy to recognize when it is written over an uppercase letter. In this regard, it may be a good practice to avoid using the slash notation over a letter. Alternatively, one may use the "bar" notation $\overline{a\,R\,b}$ to indicate that a and b are not related. ♢

Example 7.1.4 Define $R = \{(a, b) \in \mathbb{R}^2 \mid a < b\}$, hence $(a, b) \in R$ if and only if $a < b$. Obviously, saying "$a < b$" is much clearer than "$a\,R\,b$." If a and b are not related, we could say $(a, b) \notin R$, or $a \not< b$. ▲

Example 7.1.5 Define

$$F = \left\{ (x, y) \in \mathbb{R}^2 \;\middle|\; y = \frac{1}{x^2+1} \right\}.$$

Therefore x is related to y if and only if $y = \frac{1}{x^2+1}$. We can also write

$$F = \left\{ \left(x, \frac{1}{x^2+1} \right) \;\middle|\; x \in \mathbb{R} \right\},$$

which may look a bit simpler.

For instance, $(1, 0.5) \in F$, but $(1, 0) \notin F$. In this case, $(2, 0.2) \in F$ is probably easier to understand than $2\,F\,0.2$. Likewise, $(1, 2) \notin F$ may be easier to read than $1 \not F\, 2$. ▲

Hands-On Exercise 7.1.1 Define the relation H as $\{(x, x^2+1) \mid x \in \mathbb{R}\}$. Determine whether the following statements

$$2\,H\,3, \quad (-4, 17) \notin H, \quad \left(\tfrac{1}{2}, \tfrac{3}{2}\right) \notin H, \quad (\sqrt{2}, 3) \in H, \quad (1, 2) \in H,$$

are true or false.

△

Hands-On Exercise 7.1.2 Let $G = \{(x, y) \in \mathbb{R}^2 \mid xy = 1\}$. Is 2 related to 0.5? How would you write it? Repeat with 4 and 0.5, and with 10 and 3.

△

Hands-On Exercise 7.1.3 In the last example, is 0 related to 3? How would you write it? Repeat with 1 and -1. Again with $\frac{1}{\sqrt{2}}$ and $\sqrt{2}$.

△

Since a relation is a set, we can describe a relation by listing its elements (that is, using the roster method).

Example 7.1.6 Let $A = \{1, 2, 3, 4, 5, 6\}$ and $B = \{1, 2, 3, 4\}$. Define $(a, b) \in R$ if and only if $(a - b) \bmod 2 = 0$. Then

$$R = \{(1,1), (1,3), (2,2), (2,4), (3,1), (3,3), (4,2), (4,4), (5,1), (5,3), (6,2), (6,4)\}.$$

We note that R consists of ordered pairs (a, b) where a and b have the same parity. Be cautious, that $1 \le a \le 6$ and $1 \le b \le 4$. Hence, it is meaningless to talk about whether $(1, 5) \in R$ or $(1, 5) \notin R$. ▲

Hands-On Exercise 7.1.4 Let $A = \{2, 3, 4, 7\}$ and $B = \{1, 2, 3, \ldots, 12\}$. Define $a\,S\,b$ if and only if $a \mid b$. Use the roster method to describe S.

△

In the last example, 7 never appears as the first element (in the first coordinate) of any ordered pair. Likewise, 1, 5, 7, and 11 never appear as the second element (in the second coordinate) of any ordered pair.

Definition. The ***domain*** of a relation $R \subseteq A \times B$ is defined as

$$\operatorname{dom} R = \{a \in A \mid (a, b) \in R \text{ for some } b \in B\},$$

and the ***image*** or ***range*** is defined as

$$\operatorname{im} R = \{b \in B \mid (a, b) \in R \text{ for some } a \in A\}.$$

Hands-On Exercise 7.1.5 Find $\operatorname{dom} S$ and $\operatorname{im} S$, where S in Hands-On Exercise 7.1.4.

△

A relation $R \subseteq A \times B$ can be displayed graphically on a ***digraph*** which is also called a ***directed graph***. Represent the elements from A and B by ***vertices*** or ***dots***, and use ***directed lines*** (also called ***directed edges*** or ***arcs***) to connect two vertices if the corresponding elements are related. Figure 7.1 displays a graphical representation of the relation in Example 7.1.6.

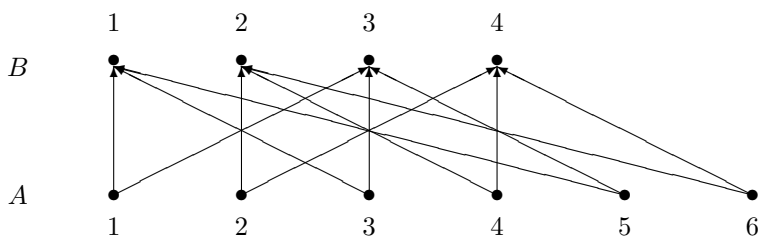


Figure 7.1: The graphical representation of the a relation.

Although a digraph gives us a clear and precise visual representation of a relation, it could become very confusing and hard to read when the relation contains many ordered pairs. As we will see in Section 7.4, we can sometimes simplify the digraphs in some special situations. Otherwise, the graphical representation is only effective for relations with a small number of ordered pairs.

We can use a ***matrix representation*** to describe a relation. A matrix consists of values arranged in rows and columns. A relation R from $A = \{a_1, \ldots, a_m\}$ to $B = \{b_1, \ldots, b_n\}$ can be described by an m-by-n matrix $M = (m_{ij})$ whose entry at row i and column j is defined by

$$m_{ij} = \begin{cases} 1 & \text{if } a_i \, R \, b_j, \\ 0 & \text{otherwise.} \end{cases}$$

The matrix M is called the ***incidence matrix*** for R.

Example 7.1.7 The incidence matrix for the relation R in Example 7.1.6 is

$$\begin{array}{c} \\ 1 \\ 2 \\ 3 \\ 4 \\ 5 \\ 6 \end{array} \begin{array}{c} \begin{array}{cccc} 1 & 2 & 3 & 4 \end{array} \\ \begin{pmatrix} 1 & 0 & 1 & 0 \\ 0 & 1 & 0 & 1 \\ 1 & 0 & 1 & 0 \\ 0 & 1 & 0 & 1 \\ 1 & 0 & 1 & 0 \\ 0 & 1 & 0 & 1 \end{pmatrix} \end{array}$$

in which we label the rows and columns with the elements involved in the relation. ▲

Hands-On Exercise 7.1.6 Determine the incidence matrix for the relation S in Hands-On Exercise 7.1.4.

△

Hands-On Exercise 7.1.7 The courses taken by John, Mary, Paul, and Sally are listed below.

John:	MATH 210, CSIT 121, MATH 223
Mary:	MATH 231, CSIT 121, MATH 210
Paul:	CSIT 120, MATH 231, MATH 223
Sally:	MATH 210, CSIT 120

Represent, using a graph and a matrix, the relation R defined as $a\,R\,b$ if student a is taking course b.

△

Summary and Review

- Relations are generalizations of functions. A relation merely states that the elements from two sets A and B are related in a certain way.
- More formally, a relation is defined as a subset of $A \times B$.
- The domain of a relation is the set of elements in A that appear in the first coordinates of some ordered pairs, and the image or range is the set of elements in B that appear in the second coordinates of some ordered pairs.
- For brevity and for clarity, we often write $x\,R\,y$ if $(x, y) \in R$.
- Under this convention, the mathematical notations $\leq$, $\geq$, $=$, $\subseteq$, and their like, can be regarded as relational operators.

Exercises 7.1

1. Represent each of the following relations from $\{1, 2, 3, 6\}$ to $\{1, 2, 3, 6\}$ using a digraph and an incidence matrix.

   (a) $\{(x, y) \mid x = y\}$ (b) $\{(x, y) \mid x \neq y\}$ (c) $\{(x, y) \mid x < y\}$

2. Find the domain and image of each relation in Problem 1.

3. Represent each of the following relations from $\{1, 2, 3, 6\}$ to $\{1, 2, 3, 6\}$ using a digraph and an incidence matrix.

   (a) $\{(x, y) \mid x^2 \leq y\}$ (d) $\{(x, y) \mid x \text{ divides } y\}$ (c) $\{(x, y) \mid x + y \text{ is even}\}$

4. Find the domain and image of each relation in Problem 3.

5. Find the incidence matrix for each of the following relations from $\{1, 2, 3, 4\}$ to $\{1, 2, 3, 4, 5\}$.

   (a) $R = \{(1, 1), (2, 2), (2, 3), (3, 3), (3, 4), (4, 5)\}$

   (b) $S = \{(1, 1), (1, 2), (2, 2), (2, 3), (3, 3), (3, 4), (4, 4)\}$

   (c) $T = \{(1, 5), (2, 4), (3, 3), (4, 1), (4, 4)\}$

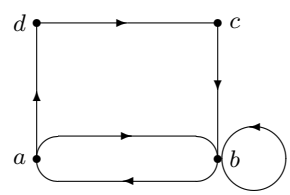
6. Determine the incidence matrix and the digraph that represent the relation R defined on $\{x \in \mathbb{Z} \mid -3 \leq x \leq 3\}$ by
$$x\,R\,y \Leftrightarrow 3 \mid (x - y).$$

7. Determine the incidence matrix and the digraph that represent the relation S defined on $\{1, 2, 4, 5, 10, 20\}$ by
$$x\,S\,y \Leftrightarrow (x < y \text{ and } x \text{ divides } y).$$

8. Let $D = \{1, 2, 3, \ldots, 30\}$ be the set of dates in November, and let $W = \{$Sunday, Monday, Tuesday, Wednesday, Thursday, Friday, Saturday$\}$ be the set of days of the week. For November of this year, define the relation T from D to W by
$$(x, y) \in T \Leftrightarrow x \text{ falls on } y.$$
List the ordered pairs in T. Is T a function from T to W?

9. Find the incidence matrix for the relation $I \subseteq \wp(\{1, 2\}) \times \wp(\{1, 2\})$, where
$$(S, T) \in I \Leftrightarrow S \cap T \neq \emptyset.$$

10. For a relation $R \subseteq A \times A$, instead of using two rows of vertices in a digraph, we can use a digraph on the vertices that represent the elements of A. Hence, it is possible to have two directed arcs between a pair of vertices, and a loop may appear around a vertex x if $(x, x) \in R$. Find the incidence matrix for the relation represented by the following digraph:



7.2 Properties of Relations

If R is a relation from A to A, then $R \subseteq A \times A$; we say that R is a ***relation on A***.

Definition. A relation R on A is said to be

- ***reflexive*** if $(a, a) \in R\,a$ for all $a \in A$,
- ***irreflexive*** if $(a, a) \notin R$ for all $a \in A$,
- ***symmetric*** if $(a, b) \in R \Rightarrow (b, a) \in R$ for all $a, b \in A$,
- ***antisymmetric*** if $[(a, b) \in R \wedge (b, a) \in R] \Rightarrow a = b$ for all $a, b \in A$,
- ***transitive*** if $[(a, b) \in R \wedge (b, c) \in R] \Rightarrow (a, c) \in R$ for all $a, b, c \in A$.

These are important definitions, so let us repeat them using the relational notation $a\,R\,b$:

- ***reflexive*** if $a\,R\,a$ for all $a \in A$,
- ***irreflexive*** if $a \not R\, a$ (that is, $\overline{a\,R\,a}$) for all $a \in A$,
- ***symmetric*** if $a\,R\,b \Rightarrow b\,R\,a$ for all $a, b \in A$,
- ***antisymmetric*** if $[(a\,R\,b) \wedge (b\,R\,a)] \Rightarrow a = b$ for all $a, b \in A$,
- ***transitive*** if $[(a\,R\,b) \wedge (b\,R\,c)] \Rightarrow a\,R\,c$ for all $a, b, c \in A$. ♢

Remark. A relation cannot be both reflexive and irreflexive. Hence, these two properties are mutually exclusive. If it is reflexive, then it is not irreflexive. If it is irreflexive, then it cannot be reflexive. Nonetheless, it is possible for a relation to be neither reflexive nor irreflexive. ♢

Remark. Many students find the concept of symmetry and antisymmetry confusing. Even though the name may suggest so, antisymmetry is *not* the opposite of symmetry. It is possible for a relation to be both symmetric and antisymmetric, and it is also possible for a relation to be both non-symmetric and non-antisymmetric. A good way to understand antisymmetry is to look at its contrapositive:

$$a \neq b \Rightarrow \overline{(a, b) \in R \wedge (b, a) \in R}.$$

Thus, if two distinct elements a and b are related (not every pair of elements need to be related), then either a is related to b, or b is related to a, *but not both*. Consequently, if we find *distinct* elements a and b such that $(a, b) \in R$ *and* $(b, a) \in R$, then R is not antisymmetric. ♢

Example 7.2.1 The ***empty relation*** is the subset $\emptyset$. It is clearly irreflexive, hence not reflexive. To check symmetry, we want to know whether $a\,R\,b \Rightarrow b\,R\,a$ for all $a, b \in A$. More specifically, we want to know whether $(a, b) \in \emptyset \Rightarrow (b, a) \in \emptyset$. Since $(a, b) \in \emptyset$ is always false, the implication is always true. Thus the relation is symmetric. Likewise, it is antisymmetric and transitive.

The ***complete relation*** is the entire set $A \times A$. It is clearly reflexive, hence not irreflexive. It is also trivial that it is symmetric and transitive. It is not antisymmetric unless $|A| = 1$.

The ***identity relation*** consists of ordered pairs of the form (a, a), where $a \in A$. In other words, $a\,R\,b$ if and only if $a = b$. It is reflexive (hence not irreflexive), symmetric, antisymmetric, and transitive. ▲

Example 7.2.2 Consider the relation R on the set $A = \{1, 2, 3, 4\}$ defined by

$$R = \{(1, 1), (2, 3), (2, 4), (3, 3), (3, 4)\}.$$

- Since $(2, 2) \notin R$, and $(1, 1) \in R$, the relation is neither reflexive nor irreflexive.
- We have $(2, 3) \in R$ but $(3, 2) \notin R$, thus R is not symmetric.
- For any $a \neq b$, only one of the four possibilities $(a, b) \notin R$, $(b, a) \notin R$, $(a, b) \in R$, or $(b, a) \in R$ can occur, so R is antisymmetric.
- By going through all the ordered pairs in R, we verify that whether $(a, b) \in R$ and $(b, c) \in R$, we always have $(a, c) \in R$ as well. This shows that R is transitive.

Therefore, R is antisymmetric and transitive. ▲

Example 7.2.3 Define the relation S on the set $A = \{1, 2, 3, 4\}$ according to

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

- Since $(1, 1), (2, 2), (3, 3), (4, 4) \notin S$, the relation S is irreflexive, hence, it is not reflexive.
- Since we have only two ordered pairs, and it is clear that whenever $(a, b) \in S$, we also have $(b, a) \in S$. Hence, S is symmetric.
- We have both $(2, 3) \in S$ and $(3, 2) \in S$, but $2 \neq 3$. Hence, S is not antisymmetric.
- Since $(2, 3) \in S$ and $(3, 2) \in S$, but $(2, 2) \notin S$, the relation S is not transitive.

We conclude that S is irreflexive and symmetric. ▲

Hands-On Exercise 7.2.1 Define the relation R on the set $\mathbb{R}$ as

$$a\,R\,b \Leftrightarrow a \leq b.$$

Determine whether R is reflexive, irreflexive, symmetric, antisymmetric, or transitive.

△

Hands-On Exercise 7.2.2 The relation S on the set $\mathbb{R}^*$ is defined as

$$a\,S\,b \Leftrightarrow ab > 0.$$

Determine whether S is reflexive, irreflexive, symmetric, antisymmetric, or transitive.

△

Example 7.2.4 Here are two examples from geometry. Let $\mathcal{T}$ be the set of triangles that can be drawn on a plane. Define a relation S on $\mathcal{T}$ such that $(T_1, T_2) \in S$ if and only if the two triangles are similar. It is easy to check that S is reflexive, symmetric, and transitive.

Let $\mathcal{L}$ be the set of all the (straight) lines on a plane. Define a relation P on $\mathcal{L}$ according to $(L_1, L_2) \in P$ if and only if L_1 and L_2 are parallel lines. Again, it is obvious that P is reflexive, symmetric, and transitive. ▲

Example 7.2.5 The relation T on $\mathbb{R}^*$ is defined as

$$a\,T\,b \Leftrightarrow \frac{a}{b} \in \mathbb{Q}.$$

- Since $\frac{a}{a} = 1 \in \mathbb{Q}$, the relation T is reflexive; it follows that T is not irreflexive.
- The relation T is symmetric, because if $\frac{a}{b}$ can be written as $\frac{m}{n}$ for some integers m and n, then so is its reciprocal $\frac{b}{a}$, because $\frac{b}{a} = \frac{n}{m}$.
- Since $\sqrt{2}\,T\sqrt{18}$ and $\sqrt{18}\,T\sqrt{2}$, yet $\sqrt{2} \neq \sqrt{18}$, we conclude that T is not antisymmetric.
- If $\frac{a}{b}, \frac{b}{c} \in \mathbb{Q}$, then $\frac{a}{b} = \frac{m}{n}$ and $\frac{b}{c} = \frac{p}{q}$ for some nonzero integers m, n, p, and q. Then $\frac{a}{c} = \frac{a}{b} \cdot \frac{b}{c} = \frac{mp}{nq} \in \mathbb{Q}$. Hence, T is transitive.

Therefore, the relation T is reflexive, symmetric, and transitive. ▲

Hands-On Exercise 7.2.3 Consider the relation T on $\mathbb{N}$ defined by

$$a\,T\,b \Leftrightarrow a \mid b.$$

Determine whether T is reflexive, irreflexive, symmetric, antisymmetric, or transitive.

△

Hands-On Exercise 7.2.4 The relation U on the set $\mathbb{Z}^*$ is defined as

$$a\,U\,b \Leftrightarrow a \mid b.$$

Determine whether U is reflexive, irreflexive, symmetric, antisymmetric, or transitive.

△

Example 7.2.6 The relation U on $\mathbb{Z}$ is defined as

$$a\,U\,b \Leftrightarrow 5 \mid (a+b).$$

- The relation U is not reflexive, because $5 \nmid (1+1)$.
- It is not irreflexive either, because $5 \mid (10+10)$.
- If $5 \mid (a+b)$, it is obvious that $5 \mid (b+a)$ because $a+b=b+a$. Thus, U is symmetric.
- We claim that U is not antisymmetric. For example, $5 \mid (2+3)$ and $5 \mid (3+2)$, yet $2 \neq 3$.
- It is not transitive either. For instance, $5 \mid (1+4)$ and $5 \mid (4+6)$, but $5 \nmid (1+6)$.

The relation U is symmetric. ▲

Hands-On Exercise 7.2.5 Determine whether the following relation V on some universal set $\mathcal{U}$ is reflexive, irreflexive, symmetric, antisymmetric, or transitive:

$$(S,T) \in V \Leftrightarrow S \subseteq T.$$

△

Example 7.2.7 Consider the relation V on the set $A = \{0, 1\}$ is defined according to

$$V = \{(0, 0), (1, 1)\}.$$

- The relation V is reflexive, because $(0, 0) \in V$ and $(1, 1) \in V$. Hence, it is not irreflexive.
- It is clearly symmetric, because $(a, b) \in V$ always implies $(b, a) \in V$.
- Indeed, whenever $(a, b) \in V$, we must also have $a = b$, because V consists of only two ordered pairs, both of them are in the form of (a, a). It follows that V is also antisymmetric.
- A similar argument shows that V is transitive.

The relation is reflexive, symmetric, antisymmetric, and transitive. ▲

Hands-On Exercise 7.2.6 Determine whether the following relation W on a nonempty set of individuals in a community is reflexive, irreflexive, symmetric, antisymmetric, or transitive:

$$a\,W\,b \Leftrightarrow a \text{ and } b \text{ have the same last name.}$$

△

Example 7.2.8 Define the relation W on a nonempty set of individuals in a community as

$$a\,W\,b \Leftrightarrow a \text{ is a child of } b.$$

- Nobody can be a child of himself or herself, hence, W cannot be reflexive. Instead, it is irreflexive.
- It is obvious that W cannot be symmetric.
- It may sound weird from the definition that W is antisymmetric:

$$(a \text{ is a child of } b) \wedge (b \text{ is a child of } a) \Rightarrow a = b, \tag{7.1}$$

  but it is true! The reason is, if a is a child of b, then b cannot be a child of a. This makes conjunction

$$(a \text{ is a child of } b) \wedge (b \text{ is a child of } a)$$

  false, which makes the implication (7.1) true.
- A similar argument holds if b is a child of a, and if neither a is a child of b nor b is a child of a. No matter what happens, the implication (7.1) is always true. Therefore W is antisymmetric.
- It may help if we look at antisymmetry from a different angle. The contrapositive of the original definition asserts that when $a \neq b$, three things could happen:

  (i) a and b are incomparable ($\overline{a\,W\,b}$ and $\overline{b\,W\,a}$), that is, a and b are unrelated;

  and if a and b are related, then either

  (ii) $a\,W\,b$ but $\overline{b\,W\,a}$, or
  (iii) $b\,W\,a$ but $\overline{a\,W\,b}$.

Using this observation, it is easy to see why W is antisymmetric.

- It is clear that W is not transitive.

The relation is irreflexive and antisymmetric. ▲

Instead of using two rows of vertices in the digraph that represents a relation on a set A, we can use just one set of vertices to represent the elements of A. A directed line connects vertex a to vertex b if and only if the element a is related to the element b. If b is also related to a, the two vertices will be joined by two directed lines, one in each direction. If a is related to itself, there is a loop around the vertex representing a. See Problem 10 in Exercises 7.1.

From the graphical representation, we determine that the relation R is

- Reflexive if there is a loop at every vertex of G.
- Irreflexive if G is loopless.
- Symmetric if every pair of vertices is connected by none or exactly two directed lines in opposite directions.
- Antisymmetric if every pair of vertices is connected by none or exactly one directed line.
- Transitive if for every unidirectional path joining three vertices a, b, c, in that order, there is also a directed line joining a to c.

The incidence matrix $M = (m_{ij})$ for a relation on A is a square matrix. We find that R is

- Reflexive if every entry on the main diagonal of M is 1.
- Irreflexive if every entry on the main diagonal of M is 0.
- Symmetric if M is symmetric, that is, $m_{ij} = m_{ji}$ whenever $i \neq j$.
- Antisymmetric if $i \neq j$ implies that at least one of m_{ij} and m_{ji} is zero, that is, $m_{ij}m_{ji} = 0$.
- Transitive if $(M^2)_{ij} > 0$ implies $m_{ij} > 0$ whenever $i \neq j$.

For instance, the incidence matrix for the identity relation consists of 1s on the main diagonal, and 0s everywhere else. This is called the identity matrix. If a relation R on A is both symmetric and antisymmetric, its off-diagonal entries are all zeros, so it is a subset of the identity relation.

It is an interesting exercise to prove the test for transitivity. Apply it to Example 7.2.2 to see how it works.

Summary and Review

- A relation from a set A to itself is called a relation on A.
- Given any relation R on a set A, we are interested in five properties that R may or may not have.
- The relation R is said to be reflexive if every element is related to itself, that is, if $x\,R\,x$ for every $x \in A$.
- The relation R is said to be irreflexive if no element is related to itself, that is, if $x \not R\, x$ for every $x \in A$.
- The reflexive property and the irreflexive property are mutually exclusive, and it is possible for a relation to be neither reflexive nor irreflexive.
- The relation R is said to be symmetric if the relation can go in both directions, that is, if $x\,R\,y$ implies $y\,R\,x$ for any $x, y \in A$.
- The relation R is said to be antisymmetric if given any two *distinct* elements x and y, either (i) x and y are not related in any way, or (ii) if x and y are related, they can only be related in one direction.
- A compact way to define antisymmetry is: if $x\,R\,y$ and $y\,R\,x$, then we must have $x = y$.
- Finally, a relation is said to be transitive if we can pass along the relation and relate two elements if they are related via a third element.
- More precisely, R is transitive if $x\,R\,y$ and $y\,R\,z$ implies that $x\,R\,z$.

Exercises 7.2

1. For each relation in Problem 1 in Exercises 7.1, determine which of the five properties are satisfied.

2. For each relation in Problem 3 in Exercises 7.1, determine which of the five properties are satisfied.

3. For the relation in Problem 6 in Exercises 7.1, determine which of the five properties are satisfied.

4. For the relation in Problem 7 in Exercises 7.1, determine which of the five properties are satisfied.

5. For the relation in Problem 8 in Exercises 7.1, determine which of the five properties are satisfied.

6. For the relation in Problem 9 in Exercises 7.1, determine which of the five properties are satisfied.

7. Let S be a nonempty set and define the relation A on $\wp(S)$ by

$$(X, Y) \in A \Leftrightarrow X \cap Y = \emptyset.$$

   It is clear that A is symmetric.

   (a) Explain why A is not reflexive.
   (b) Explain why A is not irreflexive.
   (c) Is A transitive?
   (d) Let $S = \{a, b, c\}$. Draw the directed graph for A, and find the incidence matrix that represents A.

8. For each of these relations on $\mathbb{N} - \{1\}$, determine which of the five properties are satisfied.

   (a) $A_1 = \{(x, y) \mid x \text{ and } y \text{ are relatively prime}\}$
   (b) $A_2 = \{(x, y) \mid x \text{ and } y \text{ are not relatively prime}\}$

9. For each of the following relations on $\mathbb{N}$, determine which of the five properties are satisfied.

   (a) $R_1 = \{(x, y) \mid x \text{ divides } y\}$
   (b) $R_2 = \{(x, y) \mid x + y \text{ is even}\}$
   (c) $R_3 = \{(x, y) \mid xy \text{ is even}\}$

10. For each of the following relations on $\mathbb{N}$, determine which of the five properties are satisfied.

    (a) $S_1 = \{(x, y) \mid y \text{ divides } x\}$
    (b) $S_2 = \{(x, y) \mid x + y \text{ is odd}\}$
    (c) $S_3 = \{(x, y) \mid xy \text{ is odd}\}$

11. For each of the following relations on $\mathbb{Z}$, determine which of the five properties are satisfied.

    (a) $U_1 = \{(x, y) \mid x \leq y\}$
    (b) $U_2 = \{(x, y) \mid x - y \text{ is odd}\}$
    (c) $U_3 = \{(x, y) \mid 3 \text{ divides } x + 2y\}$

12. For each of the following relations on $\mathbb{Z}$, determine which of the five properties are satisfied.

    (a) $V_1 = \{(x, y) \mid xy > 0\}$
    (b) $V_2 = \{(x, y) \mid x - y \text{ is even}\}$
    (c) $V_3 = \{(x, y) \mid x \text{ is a multiple of } y\}$

7.3 Equivalence Relations

Definition. A relation on a set A is an ***equivalence relation*** if it is reflexive, symmetric, and transitive. We often use the tilde notation $a \sim b$ to denote an equivalence relation. ♢

Example 7.3.1 The relations in Examples 7.2.4, 7.2.5, and 7.2.7, are equivalence relations, so are those in Hands-On Exercises 7.2.2 and 7.2.6. ▲

Example 7.3.2 Define a relation $\sim$ on $\mathbb{Z}$ by

$$a \sim b \Leftrightarrow a \equiv b \pmod 4.$$

Verify that $\sim$ is an equivalence relation.

Solution: We need to check three properties:

- It is obvious $a \equiv a \pmod 4$, hence $a \sim a$. The relation $\sim$ is reflexive.
- If $a \sim b$, then $a \equiv b \pmod 4$. It is clear that we also have $b \equiv a \pmod 4$. Hence, $\sim$ is symmetric.
- If $a \sim b$ and $b \sim c$, then

  $$a \equiv b \pmod 4, \qquad \text{and} \qquad b \equiv c \pmod 4.$$

  It follows that $a \equiv c \pmod 4$. Thus $a \sim c$. This shows that $\sim$ is transitive.

Therefore, $\sim$ is an equivalence relation. ▲

Hands-On Exercise 7.3.1 Define a relation $\sim$ on $\mathbb{Z}$ by

$$a \sim b \Leftrightarrow a \equiv b \pmod 6.$$

Verify that $\sim$ is an equivalence relation.

△

Hands-On Exercise 7.3.2 Let $n \geq 2$ be a positive integer. Define a relation $\sim$ on $\mathbb{Z}$ by

$$a \sim b \Leftrightarrow a \equiv b \pmod n.$$

Verify that $\sim$ is an equivalence relation.

△

Take a closer look at Example 7.3.2. All the integers having the same remainder when divided by 4 are related to each other. Define the sets

$$\begin{array}{rcrcr} [0] & = & \{n \in \mathbb{Z} \mid n \bmod 4 = 0\} & = & 4\mathbb{Z}, \\ [1] & = & \{n \in \mathbb{Z} \mid n \bmod 4 = 1\} & = & 1 + 4\mathbb{Z}, \\ [2] & = & \{n \in \mathbb{Z} \mid n \bmod 4 = 2\} & = & 2 + 4\mathbb{Z}, \\ [3] & = & \{n \in \mathbb{Z} \mid n \bmod 4 = 3\} & = & 3 + 4\mathbb{Z}. \end{array}$$

It is clear that every integer belongs to exactly one of these four sets. Hence,

$$\mathbb{Z} = [0] \cup [1] \cup [2] \cup [3].$$

These four sets are pairwise disjoint, so $\mathbb{Z}$ is a *disjoint union* of these four sets. We say that $\{[0], [1], [2], [3]\}$ is a partition of $\mathbb{Z}$.

Definition. A collection $\{S_1, S_2, \ldots, S_n\}$ of nonempty subsets of S is said to be a ***partition*** of S if the subsets $S_1, S_2, \ldots, S_n$ are pairwise disjoint ($S_i \cap S_j = \emptyset$ whenever $i \neq j$), and

$$S_1 \cup S_2 \cup \cdots \cup S_n = S.$$

The subsets $S_1, S_2, \ldots, S_n$ are called the ***parts*** or ***components*** of the partition. ♢

Because of transitivity and symmetry, all the elements related to a fixed element must be related to each other. Thus, if we know one element in the group, we essentially know all its "relatives."

If you know one member in the gang, you know them all!

Definition. Let $\sim$ be an equivalence relation on A. The set

$$[a] = \{x \in A \mid x \sim a\}.$$

is called the ***equivalence class*** of a. ♢

Example 7.3.3 In Example 7.2.4, each equivalence class of the relation S consists of all the triangles that are similar. Note that no triangle can belong to two different equivalence classes. This means that the equivalence classes are pairwise disjoint.

In the same example, each equivalence class of the relation P consists of all the lines that are parallel. Again, take note that no line can belong to two different equivalence classes. Thus, the equivalence classes are pairwise disjoint. ▲

Example 7.3.4 For the relation $\sim$ on $\mathbb{Z}$ defined by $a \sim b \Leftrightarrow a \equiv b \pmod 4$, there are four equivalence classes $[0], [1], [2]$ and $[3]$, and the set $\{[0], [1], [2], [3]\}$ forms a partition of $\mathbb{Z}$. Therefore,

$$\mathbb{Z} = [0] \cup [1] \cup [2] \cup [3],$$

and the four components [0], [1], [2] and [3] are pairwise disjoint. ▲

Hands-On Exercise 7.3.3 What are the equivalence classes of the relation $\sim$ in Hands-On Exercise 7.3.1?

△

Hands-On Exercise 7.3.4 What are the equivalence classes of the relation $\sim$ in in Hands-On Exercise 7.3.2?

△

Hands-On Exercise 7.3.5 For each of the equivalence relations mentioned in Example 7.3.1, determine its equivalence classes.

△

All the elements in the same equivalence class are related to each other. Therefore, the elements in $[a]$ all share the same property that a enjoys, from the viewpoint of the relation $\sim$. In Example 7.3.4, the equivalence class $[0]$ consists of elements that are multiples of 4. The equivalence class $[1]$ consists of elements that, when divided by 4, leave 1 as the remainder, and similarly for the equivalence classes $[2]$ and $[3]$. Because of the common bond between the elements in an equivalence class $[a]$, all these elements can be represented by *any* member within the equivalence class. This is the spirit behind the next theorem.

Theorem 7.3.1 *If $\sim$ is an equivalence relation on A, then $a \sim b \Leftrightarrow [a] = [b]$.*

Proof: We leave the proof as an exercise. ■

Every element in an equivalence class can be its representative.

One may regard equivalence classes as objects with many aliases. Every element in an equivalence class can serve as its representative. So we have to take extra care when we deal with equivalence classes. Do not be fooled by the representatives, and consider two apparently different equivalence classes to be distinct when in reality they may be identical.

Example 7.3.5 Define $\sim$ on a set of individuals in a community according to

$$a \sim b \Leftrightarrow a \text{ and } b \text{ have the same last name.}$$

We have seen that $\sim$ is an equivalence relation. Each equivalence class consists of all the individuals with the same last name in the community. Hence, for example, James Smith, Lucy Smith, and Peter Smith all belong to the same equivalence class. Any Smith can serve as its representative, so we can denote it as, for example, [Peter Smith]. ▲

Example 7.3.6 Define $\sim$ on $\mathbb{R}^+$ according to

$$x \sim y \Leftrightarrow x - y \in \mathbb{Z}.$$

Hence, two real numbers are related if and only if they have the same decimal parts. It is easy to verify that $\sim$ is an equivalence relation, and each equivalence class $[x]$ consists of all the positive real numbers having the same decimal parts as x has. Notice that

$$\mathbb{R}^+ = \bigcup_{x \in (0,1]} [x],$$

which means that the equivalence classes $[x]$, where $x \in (0, 1]$, form a partition of $\mathbb{R}$. ▲

Hands-On Exercise 7.3.6 Prove that the relation $\sim$ in Example 7.3.6 is indeed an equivalence relation.

△

Hands-On Exercise 7.3.7 Define $\sim$ on $\mathbb{R}$ according to

$$x \sim y \Leftrightarrow x - y \in \mathbb{Z}.$$

Show that $\sim$ is an equivalence class. True or false: $-2.14 \in [5, 14]$? Explain.

△

What makes equivalence relations so important is the following ***Fundamental Theorem on Equivalence Relations***.

Theorem 7.3.2 *Given any equivalence relation on a nonempty set A, the set of equivalence classes forms a partition of A. Conversely, any partition $\{A_1, A_2, \ldots, A_n\}$ of a nonempty set A into a finite number of nonempty subsets induces an equivalence relation $\sim$ on A, where $a \sim b$ if and only if $a, b \in A_i$ for some i (thus a and b belong to the same component).*

Proof: It is clear that A is the union of the equivalence classes induced by $\sim$, so it remains to show that these equivalence classes are pairwise disjoint. Assume $[a] \cap [b] \neq \emptyset$. Let $x \in [a] \cap [b]$. Then $x \in [a]$ and $x \in [b]$. Having $x \in [a]$ means $x \sim a$, and $x \in [b]$ implies that $x \sim b$. Symmetry and transitivity imply that $a \sim b$. Theorem 7.3.1 assures that $[a] = [b]$. Therefore, if $[a] \neq [b]$, then $[a] \cap [b] = \emptyset$. This proves that the equivalence classes form a partition of A.

Let $A = A_1 \cup A_2 \cup \cdots \cup A_n$ be a partition of A, define the relation $\sim$ on A according to

$$x \sim y \Leftrightarrow x, y \in A_i \text{ for some } i.$$

It follows immediately from the definition that $x \sim x$, so the relation is reflexive. It is also clear that $x \sim y$ implies $y \sim x$, hence, the relation is symmetric. Finally, if $x \sim y$ and $y \sim z$, then $x, y \in A_i$ for some i, and $y, z \in A_j$ for some j. Since the A_is form a partition of A, the element y cannot belong to two components. This means $i = j$, hence, $x, z \in A_i$. This proves that $\sim$ is transitive. Consequently, $\sim$ is an equivalence relation. ■

The idea behind the theorem is rather simple. Each equivalence class consists of all the "relatives" from the same family, so obviously the set A can be divided into families (equivalence classes). These families do not share any common elements (hence pairwise disjoint), because Theorem 7.3.1 states that any two equivalence classes sharing some common elements must be identical. Therefore, the families form a partition of A. Conversely, given a partition $\mathcal{P}$, we could define a relation that relates all members in the same component. This relation turns out to be an equivalence relation, with each component forming an equivalence class. This equivalence relation is referred to as the ***equivalence relation induced by*** $\mathcal{P}$.

Example 7.3.7 In Example 7.2.4, the relation S is an equivalence relation, and the equivalence classes are the sets of similar triangles, which form a partition of the set $\mathcal{T}$. This means any triangle belongs to one and only one equivalence class. In other words, we can classify the triangles on a plane according to their three interior angles.

The relation P in the same example is also an equivalence relation. Its equivalence classes are the sets of lines that are parallel. Every line on the plane belongs to exactly one equivalence class. Consequently, we can classify the lines on a plane by their slopes. ▲

Example 7.3.8 Over $\mathbb{Z}^*$, define

$$R_3 = \{(m, n) \mid m, n \in \mathbb{Z}^* \text{ and } mn > 0\}.$$

It is not difficult to verify that R_3 is an equivalence relation. There are only two equivalence classes: $[1]$ and $[-1]$, where $[1]$ contains all the positive integers, and $[-1]$ all the negative integers. It is obvious that $\mathbb{Z}^* = [1] \cup [-1]$. ▲

As a line is the union of infinitely many points, a plane is the union of infinitely many parallel lines.

Example 7.3.9 For each $b \in \mathbb{R}$, define L_b to be the line in $\mathbb{R}^2$ (which is also called the xy-plane) with equation $y = 2x + b$. Then $\mathcal{L} = \{L_b \mid b \in \mathbb{R}\}$ is a partition of $\mathbb{R}^2$ because given any point on $\mathbb{R}^2$, there is only one straight line with slope 2 that can pass through it. Such a partition induces an equivalence relation $\sim$ defined by

$$(p, q) \sim (s, t) \Leftrightarrow \text{both } (p, q) \text{ and } (s, t) \text{ lie on } L_b \text{ for some } b.$$

Thus, $(p, q) \sim (s, t)$ if and only if the two points (p, q) and (s, t) lie on the same straight line of slope 2. This means $\frac{q-t}{p-s} = 2$. Therefore, we can restate the definition as

$$(p, q) \sim (s, t) \Leftrightarrow q - t = 2(p - s).$$

For example, $(1, 5) \sim (0, 3)$. In fact, $[(1, 5)]$ corresponds to the line $y = 2x + 3$ or L_3. Similarly, $[(1, 1.25)]$ corresponds to the line $y = 2x - 0.75$ or $L_{-0.75}$. In general, $L_b = [(0, b)]$. ▲

Hands-On Exercise 7.3.8 Consider the partition of $\mathbb{R}^2$ (the xy-plane)

$$\mathbb{R}^2 = \bigcup_{b \in \mathbb{R}} L_b,$$

where L_b is the line satisfying the equation $y = 5x + b$. Determine the equivalence relation induced by this partition.

△

We have studied modular arithmetic extensively. In Hands-On Exercise 7.3.2, you have already proved the following result.

Theorem 7.3.3 *For any positive integer $n \geq 2$, the relation congruence modulo n is an equivalence relation on $\mathbb{Z}$.*

We can now provide a more rigorous definition of $\mathbb{Z}_n$.

Definition. Let $n \geq 2$ be an integer. The equivalence classes $[0], [1], \ldots, [n-1]$ of the relation congruence modulo n are called the ***residue classes modulo*** n. The set

$$\mathbb{Z}_n = \big\{[0], [1], \ldots, [n-1]\big\}$$

is called the set of residue classes modulo n. ♢

Remark. We define two operations $\oplus$ and $\odot$ on the elements of $\mathbb{Z}_n$ according to

$$[a] \oplus [b] = [a + b], \qquad \text{and} \qquad [a] \odot [b] = [ab].$$

We will not go into the details, but we would like to remark that $\langle \mathbb{Z}_n, \oplus, \odot \rangle$ forms an algebraic structure called ring. In practice, we seldom write $\mathbb{Z}_n = \big\{[0], [1], \ldots, [n-1]\big\}$ because it is too

cumbersome. Instead, we just write $\mathbb{Z}_n = \{0, 1, 2, \ldots, n-1\}$. However, what we really work with in $\mathbb{Z}_n$ are the residue classes represented by the integers 0 through $n-1$. $\diamondsuit$

The incidence matrix of an equivalence relation exhibits a beautiful pattern. Conversely, by examining the incidence matrix of a relation, we can tell whether the relation is an equivalence relation.

If we can rearrange the rows and columns of an incidence matrix so that the modified incidence matrix can be divided into blocks of submatrices containing entirely 1s or entirely 0s, such that the 1-submatrices lie on the diagonal, then the underlying relation R is an equivalence relation. Here is the reason. Since the entries in each 1-submatrix are all 1s, this means the corresponding elements are all related to each other. This is the notion of transitivity. Obviously, every element is related to itself. Since the 1-submatrices lie on the diagonal, the matrix, hence the relation, is symmetric. This proves that the underlying relation is an equivalence relation. Each equivalence class consists of all the elements that correspond to the row and columns in the same 1-matrix.

Example 7.3.10 Let $A = \{1, 2, 3, 4, 5\}$ and define the relation R_1 on A by

$$R_1 = \{(1,1), (1,2), (1,3), (2,1), (2,2), (2,3), (3,1), (3,2), (3,3), (4,4), (4,4), (5,4), (5,5)\}.$$

It is clear from the incidence matrix (we add lines to make the 0- and 1-submatrices more outstanding)

$$\begin{array}{c} \\ 1 \\ 2 \\ 3 \\ 4 \\ 5 \end{array} \begin{array}{c} \begin{array}{ccccc} 1 & 2 & 3 & 4 & 5 \end{array} \\ \left(\begin{array}{ccc|cc} 1 & 1 & 1 & 0 & 0 \\ 1 & 1 & 1 & 0 & 0 \\ 1 & 1 & 1 & 0 & 0 \\ \hline 0 & 0 & 0 & 1 & 1 \\ 0 & 0 & 0 & 1 & 1 \end{array}\right) \end{array}$$

that R_1 is an equivalence relation and that it has two equivalence classes: $[1] = [2] = [3] = \{1, 2, 3\}$, and $[4] = [5] = \{4, 5\}$, such that $A = [1] \cup [4]$. ▲

Example 7.3.11 Let $A = \{a, b, c, d\}$. Define the relation R_2 on A by

$$R_2 = \{(a,a), (a,c), (b,b), (b,d), (c,a), (c,c), (d,b), (d,d)\}.$$

After rewriting the incidence matrix

$$\begin{array}{c} \\ a \\ b \\ c \\ d \end{array} \begin{array}{c} \begin{array}{cccc} a & b & c & d \end{array} \\ \begin{pmatrix} 1 & 0 & 1 & 0 \\ 0 & 1 & 0 & 1 \\ 1 & 0 & 1 & 0 \\ 0 & 1 & 0 & 1 \end{pmatrix} \end{array} \rightsquigarrow \begin{array}{c} \\ a \\ c \\ b \\ d \end{array} \begin{array}{c} \begin{array}{cccc} a & c & b & d \end{array} \\ \left(\begin{array}{cc|cc} 1 & 1 & 0 & 0 \\ 1 & 1 & 0 & 0 \\ \hline 0 & 0 & 1 & 1 \\ 0 & 0 & 1 & 1 \end{array}\right) \end{array}$$

it becomes clear that R_2 is an equivalence relation, with $[a] = [c] = \{a, c\}$, and $[b] = [d] = \{b, d\}$, such that $A = [a] \cup [b]$. ▲

Hands-On Exercise 7.3.9 The relation S defined on the set $\{1, 2, 3, 4, 5, 6\}$ is known to be

$$\begin{aligned} S = \{&(1,1), (1,4), (2,2), (2,5), (2,6), (3,3), \\ &(4,1), (4,4), (5,2), (5,5), (5,6), (6,2), (6,5), (6,6)\}. \end{aligned}$$

Show that S is an equivalence relation by studying its incidence matrix, and rewriting it if necessary. Determine the contents of its equivalence classes.

△

Example 7.3.12 Find the equivalence relation R induced by the partition

$$\mathcal{P} = \big\{\{1\}, \{3\}, \{2, 4, 5, 6\}\big\}$$

of $A = \{1, 2, 3, 4, 5, 6\}$.

Solution: From the two 1-element equivalence classes $\{1\}$ and $\{3\}$, we find two ordered pairs $(1, 1)$ and $(3, 3)$ that belong to R. From the equivalence class $\{2, 4, 5, 6\}$, any pair of elements produce an ordered pair that belongs to R. Therefore,

$$\begin{aligned} R \;=\; & \{(1,1), (3,3), (2,2), (2,4), (2,5), (2,6), (4,2), (4,4), (4,5), (4,6), \\ & (5,2), (5,4), (5,5), (5,6), (6,2), (6,4), (6,5), (6,6)\}. \end{aligned}$$

Alternatively, we can construct the incidence matrix

$$\begin{array}{c} \\ 1 \\ 3 \\ 2 \\ 4 \\ 5 \\ 6 \end{array} \begin{array}{c} \begin{array}{cccccc} 1 & 3 & 2 & 4 & 5 & 6 \end{array} \\ \left(\begin{array}{c|c|cccc} 1 & 0 & 0 & 0 & 0 & 0 \\ \hline 0 & 1 & 0 & 0 & 0 & 0 \\ \hline 0 & 0 & 1 & 1 & 1 & 1 \\ 0 & 0 & 1 & 1 & 1 & 1 \\ 0 & 0 & 1 & 1 & 1 & 1 \\ 0 & 0 & 1 & 1 & 1 & 1 \end{array}\right) \end{array}$$

from which the ordered pairs in R can be easily obtained. ▲

Hands-On Exercise 7.3.10 Find the equivalence relation R induced by the partition

$$\mathcal{P} = \big\{\{a, d\}, \{b, c, g\}, \{e, f\}\big\}$$

of $A = \{a, b, c, d, e, f, g\}$ by listing all its ordered pairs (the roster method).

△

Summary and Review

- A relation R on a set A is an equivalence relation if it is reflexive, symmetric, and transitive.
- If R is an equivalence relation on the set A, its equivalence classes form a partition of A.
- In each equivalence class, all the elements are related and every element in A belongs to one and only one equivalence class.
- The relation R determines the membership in each equivalence class, and every element in the equivalence class can be used to represent that equivalence class.
- In a sense, if you know one member within an equivalence class, you also know all the other elements in the equivalence class because they are all related according to R.
- Conversely, given a partition of A, we can use it to define an equivalence relation by declaring two elements to be related if they belong to the same component in the partition.

Exercises 7.3

1. Show that each of the following relations $\sim$ on $\mathbb{Z}$ is an equivalence relation, and find its equivalence classes.

   (a) $m \sim n \Leftrightarrow |m-3| = |n-3|$
   (b) $m \sim n \Leftrightarrow m+n$ is even

2. Show that each of the following relations $\sim$ on $\mathbb{Z}$ is an equivalence relation, and find its equivalence classes.

   (a) $m \sim n \Leftrightarrow 3 \mid (m+2n)$
   (b) $m \sim n \Leftrightarrow 5 \mid (2m+3n)$

3. Let T be a fixed subset of a nonempty set S. Define the relation $\sim$ on $\wp(S)$ by

$$X \sim Y \Leftrightarrow X \cap T = Y \cap T,$$

   Show that $\sim$ is an equivalence relation. In particular, let $S = \{1,2,3,4\}$ and $T = \{1,3\}$.

   (a) True or false: $\{1,2,4\} \sim \{1,4,5\}$?
   (b) How about $\{1,2,4\} \sim \{1,3,4\}$?
   (c) Find $[\{1,5\}]$
   (d) Describe $[X]$ for any $X \in \wp(S)$.

4. For each of the following relations $\sim$ on $\mathbb{R} \times \mathbb{R}$, determine whether it is an equivalence relation. For those that are, describe geometrically the equivalence class $[(a,b)]$.

   (a) $(x_1,y_1) \sim (x_2,y_2) \Leftrightarrow y_1 - x_1^2 = y_2 - x_2^2$.
   (b) $(x_1,y_1) \sim (x_2,y_2) \Leftrightarrow (x_1-1)^2 + y_1^2 = (x_2-1)^2 + y_2^2$

5. For each of the following relations $\sim$ on $\mathbb{R} \times \mathbb{R}$, determine whether it is an equivalence relation. For those that are, describe geometrically the equivalence class $[(a,b)]$.

   (a) $(x_1,y_1) \sim (x_2,y_2) \Leftrightarrow x_1 + y_2 = x_2 + y_1$
   (b) $(x_1,y_1) \sim (x_2,y_2) \Leftrightarrow (x_1-x_2)(y_1-y_2) = 0$

6. For each of the following relations $\sim$ on $\mathbb{R} \times \mathbb{R}$, determine whether it is an equivalence relation. For those that are, describe geometrically the equivalence class $[(a,b)]$.

   (a) $(x_1,y_1) \sim (x_2,y_2) \Leftrightarrow |x_1| + |y_1| = |x_2| + |y_2|$
   (b) $(x_1,y_1) \sim (x_2,y_2) \Leftrightarrow x_1y_1 = x_2y_2$

7. Define the relation $\sim$ on $\mathbb{Q}$ by

$$x \sim y \Leftrightarrow 2(x-y) \in \mathbb{Z}.$$

   Show that $\sim$ is an equivalence relation. Describe the equivalence classes $[0]$ and $\left[\frac{1}{4}\right]$.

8. Define the relation $\sim$ on $\mathbb{Q}$ by

$$x \sim y \Leftrightarrow \frac{x-y}{2} \in \mathbb{Z}.$$

   Show that $\sim$ is an equivalence relation. Describe the equivalence classes $[0]$, $[1]$ and $\left[\frac{1}{2}\right]$.

9. Consider the following relation on $\{a,b,c,d,e\}$:

$$\begin{aligned} R = \{&(a,a),(a,c),(a,e),(b,b),(b,d),(c,a),(c,c),(c,e),\\ &(d,b),(d,d),(e,a),(e,c),(e,e)\}. \end{aligned}$$

   Show that it is an equivalence relation, and describe its equivalence classes.

   Hint: Use the matrix representation of the relation.

10. Each part below gives a partition of $A = \{a, b, c, d, e, f, g\}$. Find the equivalence relation on A induced by the partition.

    (a) $\mathcal{P}_1 = \big\{\{a, b\}, \{c, d\}, \{e, f\}, \{g\}\big\}$ (b) $\mathcal{P}_2 = \big\{\{a, c, e, g\}, \{b, d, f\}\big\}$
    (c) $\mathcal{P}_3 = \big\{\{a, b, d, e, f\}, \{c, g\}\big\}$ (d) $\mathcal{P}_4 = \big\{\{a, b, c, d, e, f, g\}\big\}$

11. Let $\sim$ be an equivalence relation on A. Prove that if $a \sim b$, then $[a] = [b]$.

12. Let $\sim$ be an equivalence relation on A. Prove that if $[a] = [b]$, then $a \sim b$.

7.4 Partial and Total Ordering

Two special relations occur frequently in mathematics. Both have to do with some sort of ordering of the elements in a set. A branch of mathematics is devoted to their study. As you can tell from the brief discussion in this section, they cover many familiar concepts.

Definition. A relation on a nonempty set A is called a ***partial ordering*** or a ***partial-order relation*** if it is reflexive, antisymmetric, and transitive. We often use $\preceq$ to denote a partial ordering, and called $(A, \preceq)$ a ***partially ordered set*** or a ***poset***.

Example 7.4.1 The usual "less than or equal to" relation on R, denoted $\leq$, is a perfect example of partial ordering. In fact, this is the reason why we adopt the notation $\preceq$, as it reflects the similarities between the two symbols. ▲

Example 7.4.2 Another classic example of partial ordering is the subset relation, denoted $\subseteq$, on $\wp(S)$, where S is any set of elements. Observe that S can be empty, in which case $\wp(\emptyset) = \{\emptyset\}$, and $(\wp(\emptyset), \subseteq)$ is obviously a partially ordered set. ▲

Example 7.4.3 Another standard example of poset is $(\mathbb{N}, |)$. It is easy to verify that the "divides" relation over the natural numbers is a partial ordering. Can you explain why $(\mathbb{Z}^*, |)$ is not a poset? ▲

Hands-On Exercise 7.4.1 Find a counterexample to illustrate why the "divides" relation, denoted $|$, over $\mathbb{Z}^*$ is not antisymmetric. Is the "divides" relation reflexive over $\mathbb{Z}$? How about transitivity?

△

Hands-On Exercise 7.4.2 Define the relation $\sqsubseteq$ on $\wp(\{a, b, c, d\})$ according to

$$S \sqsubseteq T \Leftrightarrow S \subseteq T \cup \{a\}.$$

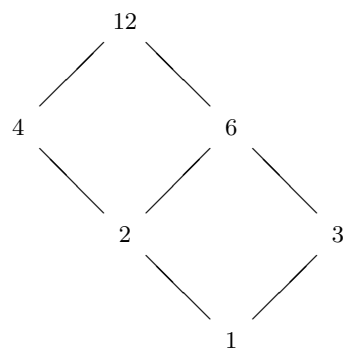
Is $(\wp(\{a, b, c, d\}, \sqsubseteq)$ a poset? Which properties it does not possess? Explain.

△

Obviously, if $a \preceq b$ but $a \neq b$, then we can write $a \prec b$. We sometimes say a ***precedes*** b, or b ***succeeds*** a. We also say a is the ***predecessor*** of b, or b is the ***successor*** of a.

The digraph for a poset can be simplified. Since a is always related to a itself, it is redundant to draw a loop around every vertex. Since $a \preceq b$ and $b \preceq c$ always imply that $a \preceq c$, there is no need to include the arc (directed edge) from a to c. So we follow the convention that we only draw an arc from a to b if $a \prec b$ *and* there does not exist another element t such that $a \prec t$ and $t \prec b$. Lastly, if $a \prec b$, we can place b above a so that all the arcs are pointing upward. This suggests that we can use undirected lines to make the graph easier to read. All these modifications lead to a much simpler graphical representation called a ***Hasse diagram***.

Example 7.4.4 It is clear that $(\{1, 2, 3, 4, 6, 12\}, |)$ is a poset. Its Hasse diagram is displayed below.



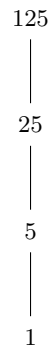
In this convention of using undirected line, the $\preceq$ relation (hence, the ordering of the elements) is read from the bottom up. ▲

Hands-On Exercise 7.4.3 Draw the Hasse diagram for the poset $(\{1, 2, 3, 4, 6, 9, 12, 18, 36\}, |)$.

△

The definition of a poset does not require every pair of distinct elements to be comparable. This means there may exist $a \neq b$ such that $a \not\preceq b$ and $b \not\preceq a$. An example can be found in the numbers 2 and 3 in Example 7.4.4. If a partial ordering has the additional property that for any two distinct elements a and b, either $a \prec b$ or $b \prec a$ (hence, any pair of distinct elements are comparable), we call the relation a ***total ordering***.

Example 7.4.5 The poset $(\mathbb{N}, \leq)$ is a totally ordered set. The poset $(\{1, 5, 25, 125\}, |)$ is also a totally ordered set. Its Hasse diagram is shown below.



It is clear that the Hasse diagram of any totally ordered set will look like the one displayed above. Consequently, a total ordering is also called a ***linear ordering***. A totally ordered set is also called a ***chain***. ▲

Hands-On Exercise 7.4.4 Construct the Hasse diagram for the poset $(\{1, 2, 4, 18, 16\}, |)$. Is it a totally ordered set?

△

Hands-On Exercise 7.4.5 Construct the Hasse diagram for the poset $(\wp(\{a, b, c\}), \subseteq)$.

△

Summary and Review

- A relation that is reflexive, antisymmetric, and transitive is called a partial ordering.
- A set with a partial ordering is called a partially ordered set or a poset.
- A poset with every pair of distinct elements comparable is called a totally ordered set.
- A total ordering is also called a linear ordering, and a totally ordered set is also called a chain.

Exercises 7.4

1. Let A be the set of natural numbers that are divisors of 30. Construct the Hasse diagram of $(A, |)$.

2. Let $S = \big\{\{a\}, \{b\}, \{c\}, \{a, b\}, \{a, c\}, \{b, c\}\big\}$. Construct the Hasse diagram for $(S, \subseteq)$.

3. Let $(A, \preceq)$ be a poset, and B a nonempty subset of A. Show that $(B, \preceq)$ is also a poset. Naturally, we call $(B, \preceq)$ a ***subposet*** of $(A, \preceq)$.

4. Define the relation $\preceq$ on $\mathbb{Z}$ according to

$$a \preceq b \Leftrightarrow a = b \text{ or } a \bmod 3 < b \bmod 3.$$

   (a) Show that $(\mathbb{Z}, \preceq)$ is a poset.

   (b) Let $B = \{-3, -2, -1, 0, 1, 2, -3\}$. Construct the Hasse diagram for the subposet $(B, \preceq)$.

5. Define the relation $\preceq$ on $\mathbb{Z}$ according to

$$a \preceq b \Leftrightarrow a = b \text{ or } |a| < |b|.$$

   (a) Show that $(\mathbb{Z}, \preceq)$ is a poset.

   (b) Construct the Hasse diagram for the subposet $(B, \preceq)$, where $B = \{-2, -1, 0, 1, 2\}$.

6. Define the relation $\preceq$ on $\mathbb{Z} \times \mathbb{Z}$ according to

$$(a, b) \preceq (c, d) \Leftrightarrow (a, b) = (c, d) \text{ or } a^2 + b^2 < c^2 + d^2.$$

   (a) Show that $(\mathbb{Z} \times \mathbb{Z}, \preceq)$ is a poset.

   (b) Construct the Hasse diagram for the subposet $(B, \preceq)$, where $B = \{0, 1, 2\} \times \{0, 1, 2\}$.

7. Construct an example of a subset B of $\wp(\{a, b, c, d\})$ such that $(B, \subseteq)$ is a totally ordered set.

8. Let

$$A = \{(m, n) \mid m, n \in \mathbb{N} \text{ and } \gcd(m, n) = 1\},$$

   and define the relation $\preceq$ on A according to

$$(a, b) \preceq (c, d) \Leftrightarrow ad \leq bc.$$

   Prove that $(A, \preceq)$ is a totally ordered set.

Chapter 6

DMOI 1: Counting

Chapter 1

Counting

One of the first things you learn in mathematics is how to count. Now we want to count large collections of things quickly and precisely. For example:

- In a group of 10 people, if everyone shakes hands with everyone else exactly once, how many handshakes took place?
- How many ways can you distribute 10 girl scout cookies to 7 boy scouts?
- How many anagrams are there of "anagram"?

Before tackling questions like these, let's look at the basics of counting.

1.1 Additive and Multiplicative Principles

Investigate!

1. A restaurant offers 8 appetizers and 14 entrées. How many choices do you have if:
   (a) you will eat one dish, either an appetizer or an entrée?
   (b) you are extra hungry and want to eat both an appetizer and an entrée?
2. Think about the methods you used to solve question 1. Write down the rules for these methods.
3. Do your rules work? A standard deck of playing cards has 26 red cards and 12 face cards.
   (a) How many ways can you select a card which is either red or a face card?
   (b) How many ways can you select a card which is both red and a face card?
   (c) How many ways can you select two cards so that the first one is red and the second one is a face card?

Attempt the above activity before proceeding

Consider this rather simple counting problem: at Red Dogs and Donuts, there are 14 varieties of donuts, and 16 types of hot dogs. If you want either a donut or a dog, how many options do you have? This isn't too hard, just add 14 and 16. Will that always work? What is important here?

Additive Principle

The **additive principle** states that if event A can occur in m ways, and event B can occur in n *disjoint* ways, then the event "A or B" can occur in $m + n$ ways.

It is important that the events be **disjoint**: i.e., that there is no way for A and B to both happen at the same time. For example, a standard deck of 52 cards contains 26 red cards and 12 face cards. However, the number of ways to select a card which is either red or a face card is not $26 + 12 = 38$. This is because there are 6 cards which are both red and face cards.

Example 1.1.1
How many two letter "words" start with either A or B? (A **word** is just a string of letters; it doesn't have to be English, or even pronounceable.)

Solution. First, how many two letter words start with A? We just need to select the second letter, which can be accomplished in 26 ways. So there are 26 words starting with A. There are also 26 words that start with B. To select a word which starts with either A or B, we can pick the word from the first 26 or the second 26, for a total of 52 words.

The additive principle also works with more than two events. Say, in addition to your 14 choices for donuts and 16 for dogs, you would also consider eating one of 15 waffles? How many choices do you have now? You would have $14 + 16 + 15 = 45$ options.

Example 1.1.2
How many two letter words start with one of the 5 vowels?

Solution. There are 26 two letter words starting with A, another 26 starting with E, and so on. We will have 5 groups of 26. So we add 26 to itself 5 times. Of course it would be easier to just multiply $5 \cdot 26$. We are really using the additive principle again, just using multiplication as a shortcut.

Example 1.1.3
Suppose you are going for some fro-yo. You can pick one of 6 yogurt choices, and one of 4 toppings. How many choices do you have?

Solution. Break your choices up into disjoint events: A are the choices with the first topping, B the choices featuring the second topping, and so on. There are four events; each can occur in 6 ways (one for each yogurt flavor). The events are disjoint, so the total number of choices is $6 + 6 + 6 + 6 = 24$.

Note that in both of the previous examples, when using the additive principle on a bunch of events all the same size, it is quicker to multiply. This really is the same, and not just because $6 + 6 + 6 + 6 = 4 \cdot 6$. We can first select the topping in 4 ways (that is, we first select which of the disjoint events we will take). For each of those first 4 choices, we now have 6 choices of yogurt. We have:

Multiplicative Principle

The **multiplicative principle** states that if event A can occur in m ways, and each possibility for A allows for exactly n ways for event B, then the event "A and B" can occur in $m \cdot n$ ways.

The multiplicative principle generalizes to more than two events as well.

Example 1.1.4
How many license plates can you make out of three letters followed by three numerical digits?

Solution. Here we have six events: the first letter, the second letter, the third letter, the first digit, the second digit, and the third digit. The first three events can each happen in 26 ways; the last three can each happen in 10 ways. So the total number of license plates will be $26 \cdot 26 \cdot 26 \cdot 10 \cdot 10 \cdot 10$, using the multiplicative principle.

Does this make sense? Think about how we would pick a license plate. How many choices we would have? First, we need to pick the first letter. There are 26 choices. Now for each of those, there are 26 choices for the second letter: 26 second letters with first letter A, 26 second letters with first letter B, and so on. We add 26 to itself 26 times. Or quicker: there are $26 \cdot 26$ choices for the first two letters.

Now for each choice of the first two letters, we have 26 choices for the third letter. That is, 26 third letters for the first two letters AA, 26 choices for the third letter after starting AB, and so on. There are $26 \cdot 26$ of these 26 third letter choices, for a total of $(26 \cdot 26) \cdot 26$ choices for the first three letters. And for each of these $26 \cdot 26 \cdot 26$ choices of letters, we have a bunch of choices for the remaining digits.

In fact, there are going to be exactly 1000 choices for the numbers. We can see this because there are 1000 three-digit numbers (000 through 999). This is 10 choices for the first digit, 10 for the second, and 10 for the third. The multiplicative principle says we multiply: $10 \cdot 10 \cdot 10 = 1000$.

All together, there were 26^3 choices for the three letters, and 10^3 choices for the numbers, so we have a total of $26^3 \cdot 10^3$ choices of license plates.

Careful: "and" doesn't mean "times." For example, how many playing cards are both red and a face card? Not $26 \cdot 12$. The answer is 6, and we needed to know something about cards to answer that question.

Another caution: how many ways can you select two cards, so that the first one is a red card and the second one is a face card? This looks more like the multiplicative principle (you are counting two separate events) but the answer is not $26 \cdot 12$ here either. The problem is that while there are 26 ways for the first card to be selected, it is not the case that *for each* of those there are 12 ways to select the second card. If the first card was both red and a face card, then there would be only 11 choices for the second card. [1]

Example 1.1.5: Counting functions
How many functions $f : \{1, 2, 3, 4, 5\} \to \{a, b, c, d\}$ are there?

Solution. Remember that a function sends each element of the domain to exactly one element of the codomain. To determine a function, we just need to specify the image of each element in the domain. Where can we send 1? There are 4 choices. Where can we send 2? Again, 4 choices. What we have here is 5 "events" (picking the image of an element in the domain) each of which can happen in 4 ways (the choices for that image). Thus there are $4 \cdot 4 \cdot 4 \cdot 4 \cdot 4 = 4^5$ functions.

This is more than just an example of how we can use the multiplicative principle in a particular counting question. What we have here is a general interpretation of certain applications of the multiplicative principle using rigorously defined mathematical objects: functions. Whenever we have a counting question that asks for the the number of outcomes of a repeated event, we can interpret that as asking for the number of functions from $\{1, 2, \ldots, n\}$ (where n is the number of times the event is repeated) to $\{1, 2, \ldots, k\}$ (where k is the number of ways that event can occur).

Counting With Sets

Do you believe the additive and multiplicative principles? How would you convince someone they are correct? This is surprisingly difficult. They seem so simple, so obvious. But why do they work?

To make things clearer, and more mathematically rigorous, we will use sets. Do not skip this section! It might seem like we are just trying to give a proof of these principles, but we are doing a lot more. If we understand the additive and multiplicative principles rigorously, we will be better at applying them, and knowing when and when not to apply them at all.

We will look at the additive and multiplicative principles in a slightly different way. Instead of thinking about event A and event B, we want to think of a set A and a set B. The sets will contain all the different ways the event can happen. (It will be helpful to be able to switch back and forth between these two models when checking that we have counted correctly.) Here's what we mean:

Example 1.1.6
Suppose you own 9 shirts and 5 pairs of pants.

1. How many outfits can you make?
2. If today is half-naked-day, and you will wear only a shirt or only a pair of pants, how many choices do you have?

Solution. By now you should agree that the answer to the first question is $9 \cdot 5 = 45$ and the answer to the second question is $9 + 5 = 14$. These are the multiplicative and additive principles. There are

[1] To solve this problem, you could break it into two cases. First, count how many ways there are to select the two cards when the first card is a red non-face card. Second, count how many ways when the first card is a red face card. Doing so makes the events in each separate case independent, so the multiplicative principle can be applied.

two events: picking a shirt and picking a pair of pants. The first event can happen in 9 ways and the second event can happen in 5 ways. To get both a shirt and a pair of pants, you multiply. To get just one article of clothing, you add.

Now look at this using sets. There are two sets, call them S and P. The set S contains all 9 shirts so $|S| = 9$ while $|P| = 5$, since there are 5 elements in the set P (namely your 5 pairs of pants). What are we asking in terms of these sets? Well in question 2, we really want $|S \cup P|$, the number of elements in the union of shirts and pants. This is just $|S| + |P|$ (since there is no overlap; $|S \cap P| = 0$). Question 1 is slightly more complicated. Your first guess might be to find $|S \cap P|$, but this is not right (there is nothing in the intersection). We are not asking for how many clothing items are both a shirt and a pair of pants. Instead, we want one of each. We could think of this as asking how many pairs (x, y) there are, where x is a shirt and y is a pair of pants. As we will soon verify, this number is $|S| \cdot |P|$.

From this example we can see right away how to rephrase our additive principle in terms of sets:

Additive Principle (with sets)

Given two sets A and B, if $A \cap B = \emptyset$ (that is, if there is no element in common to both A and B), then

$$|A \cup B| = |A| + |B|.$$

This hardly needs a proof. To find $A \cup B$, you take everything in A and throw in everything in B. Since there is no element in both sets already, you will have $|A|$ things and add $|B|$ new things to it. This is what adding does! Of course, we can easily extend this to any number of disjoint sets.

From the example above, we see that in order to investigate the multiplicative principle carefully, we need to consider ordered pairs. We should define this carefully:

Cartesian Product

Given sets A and B, we can form the *set* $A \times B = \{(x, y) : x \in A \wedge y \in B\}$ to be the set of all ordered pairs (x, y) where x is an element of A and y is an element of B. We call $A \times B$ the **Cartesian product** of A and B.

Example 1.1.7
Let $A = \{1, 2\}$ and $B = \{3, 4, 5\}$. Find $A \times B$.

Solution. We want to find ordered pairs (a, b) where a can be either 1 or 2 and b can be either 3, 4, or 5. $A \times B$ is the set of all of these pairs:

$$A \times B = \{(1, 3), (1, 4), (1, 5), (2, 3), (2, 4), (2, 5)\}$$

The question is, what is $|A \times B|$? To figure this out, write out $A \times B$. Let $A = \{a_1, a_2, a_3, \ldots, a_m\}$ and $B = \{b_1, b_2, b_3, \ldots, b_n\}$ (so $|A| = m$ and $|B| = n$). The set $A \times B$ contains all pairs with the first half of the pair being some $a_i \in A$ and the second being one of the $b_j \in B$. In other words:

$$\begin{aligned} A \times B = \{&(a_1, b_1), (a_1, b_2), (a_1, b_3), \ldots (a_1, b_n), \\ &(a_2, b_1), (a_2, b_2), (a_2, b_3), \ldots, (a_2, b_n), \\ &(a_3, b_1), (a_3, b_2), (a_3, b_3), \ldots, (a_3, b_n), \\ &\vdots \\ &(a_m, b_1), (a_m, b_2), (a_m, b_3), \ldots, (a_m, b_n)\}. \end{aligned}$$

Notice what we have done here: we made m rows of n pairs, for a total of $m \cdot n$ pairs.

Each row above is really $\{a_i\} \times B$ for some $a_i \in A$. That is, we fixed the A-element. Broken up this way, we have

$$A \times B = (\{a_1\} \times B) \cup (\{a_2\} \times B) \cup (\{a_3\} \times B) \cup \cdots \cup (\{a_m\} \times B).$$

So $A \times B$ is really the union of m disjoint sets. Each of those sets has n elements in them. The total (using the additive principle) is $n + n + n + \cdots + n = m \cdot n$.

To summarize:

Multiplicative Principle (with sets)

Given two sets A and B, we have $|A \times B| = |A| \cdot |B|$.

Again, we can easily extend this to any number of sets.

Principle of Inclusion/Exclusion

Investigate!

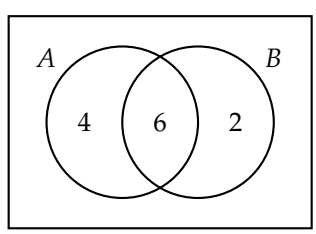
A recent buzz marketing campaign for *Village Inn* surveyed patrons on their pie preferences. People were asked whether they enjoyed (A) Apple, (B) Blueberry or (C) Cherry pie (respondents answered yes or no to each type of pie, and could say yes to more than one type). The following table shows the results of the survey.

Pies enjoyed:	A	B	C	AB	AC	BC	ABC
Number of people:	20	13	26	9	15	7	5

How many of those asked enjoy at least one of the kinds of pie? Also, explain why the answer is not 95.

Attempt the above activity before proceeding

While we are thinking about sets, consider what happens to the additive principle when the sets are NOT disjoint. Suppose we want to find $|A \cup B|$ and know that $|A| = 10$ and $|B| = 8$. This is not enough information though. We do not know how many of the 8 elements in B are also elements of A. However, if we also know that $|A \cap B| = 6$, then we can say exactly how many elements are in A, and, of those, how many are in B and how many are not (6 of the 10 elements are in B, so 4 are in A but not in B). We could fill in a Venn diagram as follows:



This says there are 6 elements in $A \cap B$, 4 elements in $A \setminus B$ and 2 elements in $B \setminus A$. Now *these* three sets *are* disjoint, so we can use the additive principle to find the number of elements in $A \cup B$. It is $6 + 4 + 2 = 12$.

This will always work, but drawing a Venn diagram is more than we need to do. In fact, it would be nice to relate this problem to the case where A and B are disjoint. Is there one rule we can make that works in either case?

Here is another way to get the answer to the problem above. Start by just adding $|A| + |B|$. This is $10 + 8 = 18$, which would be the answer if $|A \cap B| = 0$. We see that we are off by exactly 6, which just so happens to be $|A \cap B|$. So perhaps we guess,

$$|A \cup B| = |A| + |B| - |A \cap B|.$$

This works for this one example. Will it always work? Think about what we are doing here. We want to know how many things are either in A or B (or both). We can throw in everything in A, and everything in B. This would give $|A| + |B|$ many elements. But of course when you actually take the union, you do not repeat elements that are in both. So far we have counted every element in $A \cap B$ exactly twice: once when we put in the elements from A and once when we included the elements from B. We correct by subtracting out the number of elements we have counted twice. So we added them in twice, subtracted once, leaving them counted only one time.

In other words, we have:

Cardinality of a union (2 sets)

For any finite sets A and B,

$$|A \cup B| = |A| + |B| - |A \cap B|.$$

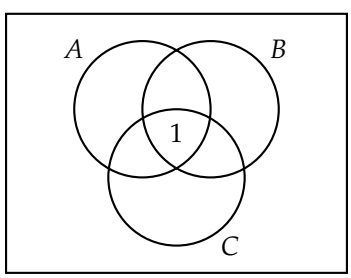
We can do something similar with three sets.

Example 1.1.8
An examination in three subjects, Algebra, Biology, and Chemistry, was taken by 41 students. The following table shows how many students failed in each single subject and in their various combinations:

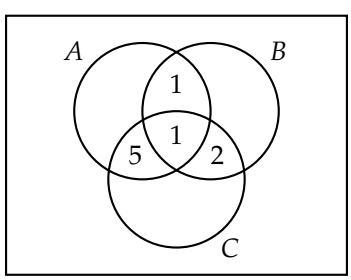
Subject:	A	B	C	AB	AC	BC	ABC
Failed:	12	5	8	2	6	3	1

How many students failed at least one subject?

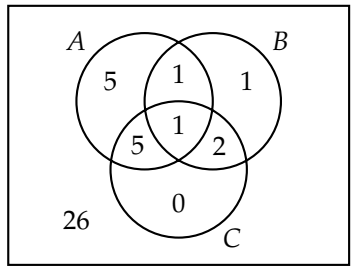
Solution. The answer is not 37, even though the sum of the numbers above is 37. For example, while 12 students failed Algebra, 2 of those students also failed Biology, 6 also failed Chemestry, and 1 of those failed all three subjects. In fact, that 1 student who failed all three subjects is counted a total of 7 times in the total 37. To clarify things, let us think of the students who failed Algebra as the elements of the set A, and similarly for sets B and C. The one student who failed all three subjects is the lone element of the set $A \cap B \cap C$. Thus, in Venn diagrams:



Now let's fill in the other intersections. We know $A \cap B$ contains 2 elements, but 1 element has already been counted. So we should put a 1 in the region where A and B intersect (but C does not). Similarly, we calculate the cardinality of $(A \cap C) \setminus B$, and $(B \cap C) \setminus A$:



Next, we determine the numbers which should go in the remaining regions, including outside of all three circles. This last number is the number of students who did not fail any subject:



We found 5 goes in the "A only" region because the entire circle for A needed to have a total of 12, and 7 were already accounted for. Similarly, we calculate the "B only" region to contain only 1 student and the "C only" region to contain no students.

Thus the number of students who failed at least one class is 15 (the sum of the numbers in each of the eight disjoint regions). The number of students who passed all three classes is 26: the total number of students, 41, less the 15 who failed at least one class.

Note that we can also answer other questions. For example, now many students failed just Chemistry? None. How many passed Algebra but failed both Biology and Chemistry? This corresponds to the region inside both B and C but outside of A, containing 2 students.

Could we have solved the problem above in an algebraic way? While the additive principle generalizes to any number of sets, when we add a third set here, we must be careful. With two sets, we needed to know the cardinalities of A, B, and $A \cap B$ in order to find the cardinality of $A \cup B$. With three sets we need more information. There are more ways the sets can combine. Not surprisingly then, the formula for cardinality of the union of three non-disjoint sets is more complicated:

Cardinality of a union (3 sets)

For any finite sets A, B, and C,

$$|A \cup B \cup C| = |A| + |B| + |C| - |A \cap B| - |A \cap C| - |B \cap C| + |A \cap B \cap C|$$

To determine how many elements are in at least one of A, B, or C we add up all the elements in each of those sets. However, when we do that, any element in both A and B is counted twice. Also, each element in both A and C is counted twice, as are elements in B and C, so we take each of those out of our sum once. But now what about the elements which are in $A \cap B \cap C$ (in all three sets)? We added them in three times, but also removed them three times. They have not yet been counted. Thus we add those elements back in at the end.

Returning to our example above, we have $|A| = 12$, $|B| = 5$, $|C| = 8$. We also have $|A \cap B| = 2$, $|A \cap C| = 6$, $|B \cap C| = 3$, and $|A \cap B \cap C| = 1$. Therefore:

$$|A \cup B \cup C| = 12 + 5 + 8 - 2 - 6 - 3 + 1 = 15$$

This is what we got when we solved the problem using Venn diagrams.

This process of adding in, then taking out, then adding back in, and so on is called the *Principle of Inclusion/Exclusion*, or simply PIE. We will return to this counting technique later to solve for more complicated problems (involving more than 3 sets).

Exercises

1. Your wardrobe consists of 5 shirts, 3 pairs of pants, and 17 bow ties. How many different outfits can you make?

2. For your college interview, you must wear a tie. You own 3 regular (boring) ties and 5 (cool) bow ties.

(a) How many choices do you have for your neck-wear?

(b) You realize that the interview is for clown college, so you should probably wear both a regular tie and a bow tie. How many choices do you have now?

(c) For the rest of your outfit, you have 5 shirts, 4 skirts, 3 pants, and 7 dresses. You want to select either a shirt to wear with a skirt or pants, or just a dress. How many outfits do you have to choose from?

3. Your Blu-ray collection consists of 9 comedies and 7 horror movies. Give an example of a question for which the answer is:

(a) 16.

(b) 63.

4. We usually write numbers in decimal form (or base 10), meaning numbers are composed using 10 different "digits" $\{0, 1, \ldots, 9\}$. Sometimes though it is useful to write numbers **hexadecimal** or base 16. Now there are 16 distinct digits that can be used to form numbers: $\{0, 1, \ldots, 9, \mathrm{A}, \mathrm{B}, \mathrm{C}, \mathrm{D}, \mathrm{E}, \mathrm{F}\}$. So for example, a 3 digit hexadecimal number might be 2B8.

(a) How many 2-digit hexadecimals are there in which the first digit is E or F? Explain your answer in terms of the additive principle (using either events or sets).

(b) Explain why your answer to the previous part is correct in terms of the multiplicative principle (using either events or sets). Why do both the additive and multiplicative principles give you the same answer?

(c) How many 3-digit hexadecimals start with a letter (A-F) and end with a numeral (0-9)? Explain.

(d) How many 3-digit hexadecimals start with a letter (A-F) or end with a numeral (0-9) (or both)? Explain.

5. Suppose you have sets A and B with $|A| = 10$ and $|B| = 15$.

(a) What is the largest possible value for $|A \cap B|$?

(b) What is the smallest possible value for $|A \cap B|$?

(c) What are the possible values for $|A \cup B|$?

6. If $|A| = 8$ and $|B| = 5$, what is $|A \cup B| + |A \cap B|$?

7. A group of college students were asked about their TV watching habits. Of those surveyed, 28 students watch *The Walking Dead*, 19 watch *The Blacklist*, and 24 watch *Game of Thrones*. Additionally, 16 watch *The Walking Dead* and *The Blacklist*, 14 watch *The Walking Dead* and *Game of Thrones*, and 10 watch *The Blacklist* and *Game of Thrones*. There are 8 students who watch all three shows. How many students surveyed watched at least one of the shows?

8. In a recent survey, 30 students reported whether they liked their potatoes Mashed, French-fried, or Twice-baked. 15 liked them mashed, 20 liked French fries, and 9 liked twice baked potatoes. Additionally, 12 students liked both mashed and fried potatoes, 5 liked French fries and twice baked potatoes, 6 liked mashed and baked, and 3 liked all three styles. How many students *hate* potatoes? Explain why your answer is correct.

9. For how many $n \in \{1, 2, \ldots, 500\}$ is n a multiple of one or more of 5, 6, or 7?

Hint. To find out how many numbers are divisible by 6 and 7, for example, take 500/42 and round down.

10. Let A, B, and C be sets.

(a) Find $|(A \cup C) \setminus B|$ provided $|A| = 50$, $|B| = 45$, $|C| = 40$, $|A \cap B| = 20$, $|A \cap C| = 15$, $|B \cap C| = 23$, and $|A \cap B \cap C| = 12$.

(b) Describe a set in terms of A, B, and C with cardinality 26.

11. Consider all 5 letter "words" made from the letters a through h. (Recall, words are just strings of letters, not necessarily actual English words.)

(a) How many of these words are there total?

(b) How many of these words contain no repeated letters?

(c) How many of these words start with the sub-word "aha"?

(d) How many of these words either start with "aha" or end with "bah" or both?

(e) How many of the words containing no repeats also do not contain the sub-word "bad"?

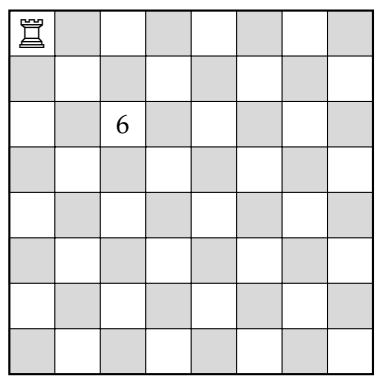
12. For how many three digit numbers (100 to 999) is the *sum of the digits* even? (For example, 343 has an even sum of digits: $3 + 4 + 3 = 10$ which is even.) Find the answer and explain why it is correct in at least two *different* ways.

13. The number 735000 factors as $2^3 \cdot 3 \cdot 5^4 \cdot 7^2$. How many divisors does it have? Explain your answer using the multiplicative principle.

1.2 Binomial Coefficients

Investigate!

In chess, a rook can move only in straight lines (not diagonally). Fill in each square of the chess board below with the number of different shortest paths the rook, in the upper left corner, can take to get to that square. For example, one square is already filled in. There are six different paths from the rook to the square: DDRR (down down right right), DRDR, DRRD, RDDR, RDRD and RRDD.



Attempt the above activity before proceeding

Here are some apparently different discrete objects we can count: subsets, bit strings, lattice paths, and binomial coefficients. We will give an example of each type of counting problem (and say what these things even are). As we will see, these counting problems are surprisingly similar.

Subsets

Subsets should be familiar, otherwise read over Section 0.3 again. Suppose we look at the set $A = \{1, 2, 3, 4, 5\}$. How many subsets of A contain exactly 3 elements?

First, a simpler question: How many subsets of A are there total? In other words, what is $|\mathcal{P}(A)|$ (the cardinality of the power set of A)? Think about how we would build a subset. We need to decide, for each of the elements of A, whether or not to include the element in our subset. So we need to decide "yes" or "no" for the element 1. And for each choice we make, we need to decide "yes" or "no" for the element 2. And so on. For each of the 5 elements, we have 2 choices. Therefore the number of subsets is simply $2 \cdot 2 \cdot 2 \cdot 2 \cdot 2 = 2^5$ (by the multiplicative principle).

Of those 32 subsets, how many have 3 elements? This is not obvious. Note that we cannot just use the multiplicative principle. Maybe we want to say we have 2 choices (yes/no) for the first element, 2 choices for the second, 2 choices for the third, and then only 1 choice for the other two. But what if we said "no" to one of the first three elements? Then we would have two choices for the 4th element. What a mess!

Another (bad) idea: we need to pick three elements to be in our subset. There are 5 elements to choose from. So there are 5 choices for the first element, and for each of those 4 choices for the second, and then 3 for the third (last) element. The multiplicative principle would say then that there are a total of $5 \cdot 4 \cdot 3 = 60$ ways to select the 3 element subset. But this cannot be correct ($60 > 32$ for one thing). One of the outcomes we would get from these choices would be the set $\{3, 2, 5\}$, by choosing the element 3 first, then the element 2, then the element 5. Another outcome would be $\{5, 2, 3\}$ by choosing the element 5 first, then the element 2, then the element 3. But these are the same set! We can correct this by dividing: for each set of three elements, there are 6 outcomes counted amoung our 60 (since there are 3 choices for which element we list first, 2 for which we list second, and 1 for which we list last). So we expect there to be 10 3-element subsets of A.

Is this right? Well, we could list out all 10 of them, being very systematic in doing so, to make sure we don't miss any or list any twice. Or we could try to count how many subsets of A *don't* have 3 elements in them. How many have no elements? Just 1 (the empty set). How many have 5? Again, just 1. These are the cases in which we say "no" to all elements, or "yes" to all elements. Okay, what about the subsets which contain a single element? There are 5 of these. We must say "yes" to exactly one element, and there are 5 to choose from. This is also the number of subsets containing 4 elements. Those are the ones for which we must say "no" to exactly one element.

So far we have counted 12 of the 32 subsets. We have not yet counted the subsets with cardinality 2 and with cardinality 3. There are a total of 20 subsets left to split up between these two groups. But the number of each must be the same! If we say "yes" to exactly two elements, that can be accomplished in exactly the same number of ways as the number of ways we can say "no" to exactly two elements. So the number of 2-element subsets is equal to the number of 3-element subsets. Together there are 20 of these subsets, so 10 each.

Number of elements:	0	1	2	3	4	5
Number of subsets:	1	5	10	10	5	1

Bit Strings

"Bit" is short for "binary digit," so a **bit string** is a string of binary digits. The **binary digits** are simply the numbers 0 and 1. All of the following are bit strings:

1001 0 1111 1010101010

The number of bits (0's or 1's) in the string is the **length** of the string; the strings above have lengths 4, 1, 4, and 10 respectively. We also can ask how many of the bits are 1's. The number of 1's in a bit string is the **weight** of the string; the weights of the above strings are 2, 0, 4, and 5 respectively.

Bit Strings

- An **n-bit string** is a bit string of length n. That is, it is a string containing n symbols, each of which is a bit, either 0 or 1.
- The **weight** of a bit string is the number of 1's in it.
- $\mathbf{B}^n$ is the *set* of all n-bit strings.
- $\mathbf{B}^n_k$ is the set of all n-bit strings of weight k.

For example, the elements of the set $\mathbf{B}^3_2$ are the bit strings 011, 101, and 110. Those are the only strings containing three bits exactly two of which are 1's.

The counting questions: How many bit strings have length 5? How many of those have weight 3? In other words, we are asking for the cardinalities $|\mathbf{B}^5|$ and $|\mathbf{B}^5_3|$.

To find the number of 5-bit strings is straight forward. We have 5 bits, and each can either be a 0 or a 1. So there are 2 choices for the first bit, 2 choices for the second, and so on. By the multiplicative principle, there are $2 \cdot 2 \cdot 2 \cdot 2 \cdot 2 = 2^5 = 32$ such strings.

Finding the number of 5-bit strings of weight 3 is harder. Think about how such a string could start. The first bit must be either a 0 or a 1. In the first case (the string starts with a 0), we must then decide on four more bits. To have a total of three 1's, among those four remaining bits there must be three 1's. To count all of these strings, we must include all 4-bit strings of weight 3. In the second case (the string starts with a 1), we still have four bits to choose, but now only two of them can be 1's, so we should look at all the 4-bit strings of weight 2. So the strings in $\mathbf{B}_3^5$ all have the form $1\mathbf{B}_2^4$ (that is, a 1 followed by a string from $\mathbf{B}_2^4$) or $0\mathbf{B}_3^4$. These two sets are disjoint, so we can use the additive principle:

$$|\mathbf{B}_3^5| = |\mathbf{B}_2^4| + |\mathbf{B}_3^4|.$$

This is an example of a **recurrence relation**. We represented one instance of our counting problem in terms of two simpler instances of the problem. If only we knew the cardinalities of $\mathbf{B}_2^4$ and $\mathbf{B}_3^4$. Repeating the same reasoning,

$$|\mathbf{B}_2^4| = |\mathbf{B}_1^3| + |\mathbf{B}_2^3| \quad \text{and} \quad |\mathbf{B}_3^4| = |\mathbf{B}_2^3| + |\mathbf{B}_3^3|.$$

We can keep going down, but this should be good enough. Both $\mathbf{B}_1^3$ and $\mathbf{B}_2^3$ contain 3 bit strings: we must pick one of the three bits to be a 1 (three ways to do that) or one of the three bits to be a 0 (three ways to do that). Also, $\mathbf{B}_3^3$ contains just one string: 111. Thus $|\mathbf{B}_2^4| = 6$ and $|\mathbf{B}_3^4| = 4$, which puts $\mathbf{B}_3^5$ at a total of 10 strings.

But wait —32 and 10 were the answers to the counting questions about subsets. Coincidence? Not at all. Each bit string can be thought of as a *code* for a subset. For the set $A = \{1, 2, 3, 4, 5\}$, we would use 5-bit strings, one bit for each element of A. Each bit in the string is a 0 if its corresponding element of A is not in the subset, and a 1 if the element of A is in the subset. Remember, deciding the subset amounted to a sequence of five yes/no votes for the elements of A. Instead of yes, we put a 1; instead of no, we put a 0.

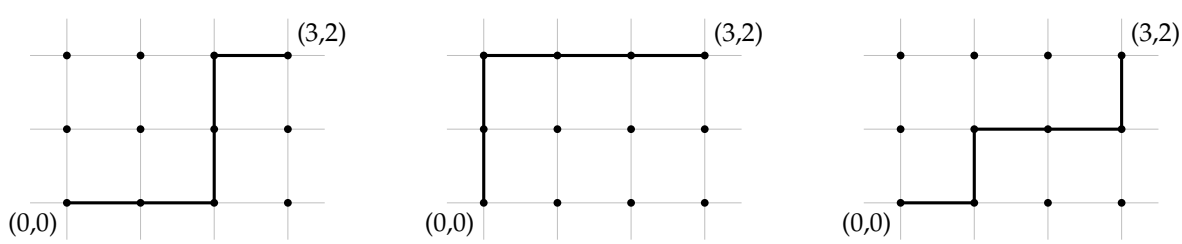
For example, the bit string 11001 represents the subset $\{1, 2, 5\}$ since the first, second and fifth bits are 1's. The subset $\{3, 5\}$ would be coded by the string 00101. What we really have here is a bijection from $\mathcal{P}(A)$ to $\mathbf{B}^5$.

Now for a subset to contain exactly three elements, the corresponding bit string must contain exactly three 1's. In other words, the weight must be 3. Thus counting the number of 3-element subsets of A is the same as counting the number 5-bit strings of weight 3.

Lattice Paths

The **integer lattice** is the set of all points in the Cartesian plane for which both the x and y coordinates are integers. If you like to draw graphs on graph paper, the lattice is the set of all the intersections of the grid lines.

A **lattice path** is one of the shortest possible paths connecting two points on the lattice, moving only horizontally and vertically. For example, here are three possible lattice paths from the points $(0, 0)$ to $(3, 2)$:

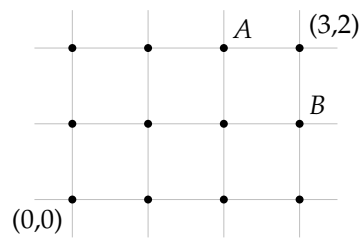


Notice to ensure the path is the *shortest* possible, each move must be either to the right or up. Additionally, in this case, note that no matter what path we take, we must make three steps right and two steps up. No matter what order we make these steps, there will always be 5 steps. Thus each path has *length* 5.

The counting question: how many lattice paths are there between $(0, 0)$ and $(3, 2)$? We could try to draw all of these, or instead of drawing them, maybe just list which direction we travel on each of the 5 steps. One path might be RRUUR, or maybe UURRR, or perhaps RURRU (those correspond to the three paths drawn above). So how many such strings of R's and U's are there?

Notice that each of these strings must contain 5 symbols. Exactly 3 of them must be R's (since our destination is 3 units to the right). This seems awfully familiar. In fact, what if we used 1's instead of R's and 0's instead of U's? Then we would just have 5-bit strings of weight 3. There are 10 of those, so there are 10 lattice paths from (0,0) to (3,2).

The correspondence between bit strings and lattice paths does not stop there. Here is another way to count lattice paths. Consider the lattice shown below:



Any lattice path from (0,0) to (3,2) must pass through exactly one of A and B. The point A is 4 steps away from (0,0) and two of them are towards the right. The number of lattice paths to A is the same as the number of 4-bit strings of weight 2, namely 6. The point B is 4 steps away from (0,0), but now 3 of them are towards the right. So the number of paths to point B is the same as the number of 4-bit strings of weight 3, namely 4. So the total number of paths to (3,2) is just $6 + 4$. This is the same way we calculated the number of 5-bit strings of weight 3. The point: the exact same recurrence relation exists for bit strings and for lattice paths.

Binomial Coefficients

Binomial coefficients are the coefficients in the expanded version of a binomial, such as $(x + y)^5$. What happens when we multiply such a binomial out? We will expand $(x + y)^n$ for various values of n. Each of these are done by multiplying everything out (i.e., FOIL-ing) and then collecting like terms.

$$(x + y)^1 = x + y$$

$$(x + y)^2 = x^2 + 2xy + y^2$$

$$(x + y)^3 = x^3 + 3x^2y + 3xy^2 + y^3$$

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

In fact, there is a quicker way to expand the above binomials. For example, consider the next one, $(x + y)^5$. What we are really doing is multiplying out,

$$(x + y)(x + y)(x + y)(x + y)(x + y).$$

If that looks daunting, go back to the case of $(x + y)^3 = (x + y)(x + y)(x + y)$. Why do we only have one x^3 and y^3 but three x^2y and xy^2 terms? Every time we distribute over an $(x + y)$ we create two copies of what is left, one multiplied by x, the other multiplied by y. To get x^3, we need to pick the "multiplied by x" side every time (we don't have any y's in the term). This will only happen once. On the other hand, to get x^2y we need to select the x side twice and the y side once. In other words, we need to pick one of the three $(x + y)$ terms to "contribute" their y.

Similarly, in the expansion of $(x + y)^5$, there will be only one x^5 term and one y^5 term. This is because to get an x^5, we need to use the x term in each of the copies of the binomial $(x + y)$, and similarly for y^5. What about x^4y? To get terms like this, we need to use four x's and one y, so we need exactly one of the five binomials to contribute a y. There are 5 choices for this, so there are 5 ways to get x^4y, so the coefficient of x^4y is 5. This is also the coefficient for xy^4 for the same (but opposite) reason: there are 5 ways to pick which of the 5 binomials contribute the single x. So far we have

$$(x + y)^5 = x^5 + 5x^4y + \underline{\ ?\ }\, x^3y^2 + \underline{\ ?\ }\, x^2y^3 + 5xy^4 + y^5.$$

We still need the coefficients of x^3y^2 and x^2y^3. In both cases, we need to pick exactly 3 of the 5 binomials to contribute one variable, the other two to contribute the other. Wait. This sounds familiar. We have 5

things, each can be one of two things, and we need a total of 3 of one of them. That's just like taking 5 bits and making sure exactly 3 of them are 1's. So the coefficient of x^3y^2 (and also x^2y^3) will be exactly the same as the number of bit strings of length 5 and weight 3, which we found earlier to be 10. So we have:

$$(x+y)^5 = x^5 + 5x^4y + 10x^3y^2 + 10x^2y^3 + 5xy^4 + y^5.$$

These numbers we keep seeing over and over again. They are the number of subsets of a particular size, the number of bit strings of a particular weight, the number of lattice paths, and the coefficients of these binomial products. We will call them **binomial coefficients**. We even have a special symbol for them: $\binom{n}{k}$.

Binomial Coefficients

For each integer $n \geq 0$ and integer k with $0 \leq k \leq n$ there is a number

$$\binom{n}{k}$$

read "n choose k." We have:

- $\binom{n}{k} = |\mathbf{B}_k^n|$, the number of n-bit strings of weight k.
- $\binom{n}{k}$ is the number of subsets of a set of size n each with cardinality k.
- $\binom{n}{k}$ is the number of lattice paths of length n containing k steps to the right.
- $\binom{n}{k}$ is the coefficient of $x^k y^{n-k}$ in the expansion of $(x+y)^n$.
- $\binom{n}{k}$ is the number of ways to select k objects from a total of n objects.

The last bullet point is usually taken as the definition of $\binom{n}{k}$. Out of n objects we must choose k of them, so there are n choose k ways of doing this. Each of our counting problems above can be viewed in this way:

- How many subsets of $\{1,2,3,4,5\}$ contain exactly 3 elements? We must choose 3 of the 5 elements to be in our subset. There are $\binom{5}{3}$ ways to do this, so there are $\binom{5}{3}$ such subsets.
- How many bit strings have length 5 and weight 3? We must choose 3 of the 5 bits to be 1's. There are $\binom{5}{3}$ ways to do this, so there are $\binom{5}{3}$ such bit strings.
- How many lattice paths are there from (0,0) to (3,2)? We must choose 3 of the 5 steps to be towards the right. There are $\binom{5}{3}$ ways to do this, so there are $\binom{5}{3}$ such lattice paths.
- What is the coefficient of x^3y^2 in the expansion of $(x+y)^5$? We must choose 3 of the 5 copies of the binomial to contribute an x. There are $\binom{5}{3}$ ways to do this, so the coefficient is $\binom{5}{3}$.

It should be clear that in each case above, we have the right answer. All we had to do is phrase the question correctly and it became obvious that $\binom{5}{3}$ is correct. However, this does not tell us that the answer is in fact 10 in each case. We will eventually find a formula for $\binom{n}{k}$, but for now, look back at how we arrived at the answer 10 in our counting problems above. It all came down to bit strings, and we have a recurrence relation for bit strings:

$$|\mathbf{B}_k^n| = |\mathbf{B}_{k-1}^{n-1}| + |\mathbf{B}_k^{n-1}|.$$

Remember, this is because we can start the bit string with either a 1 or a 0. In both cases, we have $n-1$ more bits to pick. The strings starting with 1 must contain $k-1$ more 1's, while the strings starting with 0 still need k more 1's.

Since $|\mathbf{B}_k^n| = \binom{n}{k}$, the same recurrence relation holds for binomial coefficients:

Recurrence relation for $\binom{n}{k}$

$$\binom{n}{k} = \binom{n-1}{k-1} + \binom{n-1}{k}$$

Pascal's Triangle

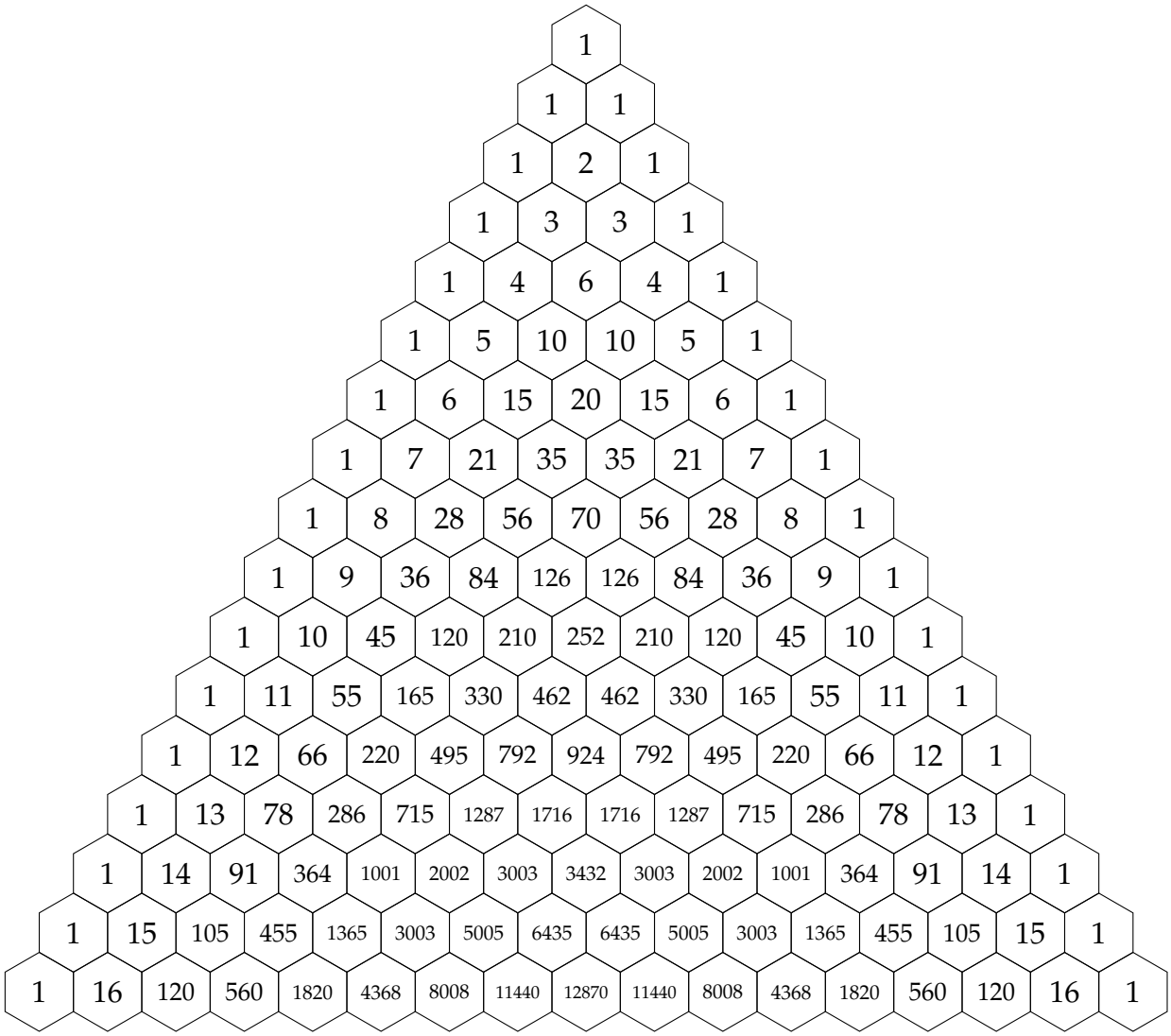
Let's arrange the binomial coefficients $\binom{n}{k}$ into a triangle like follows:

$$\begin{array}{ccccccccc}
 & & & & \binom{0}{0} & & & & \\
 & & & \binom{1}{0} & & \binom{1}{1} & & & \\
 & & \binom{2}{0} & & \binom{2}{1} & & \binom{2}{2} & & \\
 & \binom{3}{0} & & \binom{3}{1} & & \binom{3}{2} & & \binom{3}{3} & \\
\binom{4}{0} & & \binom{4}{1} & & \binom{4}{2} & & \binom{4}{3} & & \binom{4}{4}
\end{array}$$

This can continue as far down as we like. The recurrence relation for $\binom{n}{k}$ tells us that each entry in the triangle is the sum of the two entries above it. The entries on the sides of the triangle are always 1. This is because $\binom{n}{0} = 1$ for all n since there is only one way to pick 0 of n objects and $\binom{n}{n} = 1$ since there is one way to select all n out of n objects. Using the recurrence relation, and the fact that the sides of the triangle are 1's, we can easily replace all the entries above with the correct values of $\binom{n}{k}$. Doing so gives us **Pascal's triangle**.

We can use Pascal's triangle to calculate binomial coefficients. For example, using the triangle below, we can find $\binom{12}{6} = 924$.

Pascal's Triangle



Exercises

1. Let $S = \{1, 2, 3, 4, 5, 6\}$

(a) How many subsets are there total?

(b) How many subsets have $\{2, 3, 5\}$ as a subset?

(c) How many subsets contain at least one odd number?

(d) How many subsets contain exactly one even number?

2. Let $S = \{1, 2, 3, 4, 5, 6\}$

(a) How many subsets are there of cardinality 4?

(b) How many subsets of cardinality 4 have $\{2, 3, 5\}$ as a subset?

(c) How many subsets of cardinality 4 contain at least one odd number?

(d) How many subsets of cardinality 4 contain exactly one even number?

3. Let $A = \{1, 2, 3, \ldots, 9\}$.

(a) How many subsets of A are there? That is, find $|\mathcal{P}(A)|$. Explain.
(b) How many subsets of A contain exactly 5 elements? Explain.
(c) How many subsets of A contain only even numbers? Explain.
(d) How many subsets of A contain an even number of elements? Explain.

4. How many 9-bit strings (that is, bit strings of length 9) are there which:

(a) Start with the sub-string 101? Explain.
(b) Have weight 5 (i.e., contain exactly five 1's) and start with the sub-string 101? Explain.
(c) Either start with 101 or end with 11 (or both)? Explain.
(d) Have weight 5 and either start with 101 or end with 11 (or both)? Explain.

5. You break your piggy-bank to discover lots of pennies and nickels. You start arranging these in rows of 6 coins.

(a) You find yourself making rows containing an equal number of pennies and nickels. For fun, you decide to lay out every possible such row. How many coins will you need?
(b) How many coins would you need to make all possible rows of 6 coins (not necessarily with equal number of pennies and nickels)?

6. How many 10-bit strings contain 6 or more 1's?

7. How many subsets of $\{0, 1, \ldots, 9\}$ have cardinality 6 or more?

Hint. Break the question into five cases.

8. What is the coefficient of x^{12} in $(x + 2)^{15}$?

9. What is the coefficient of x^9 in the expansion of $(x + 1)^{14} + x^3(x + 2)^{15}$?

10. How many shortest lattice paths start at (3,3) and

(a) end at (10,10)?
(b) end at (10,10) and pass through (5,7)?
(c) end at (10,10) and avoid (5,7)?

11. Gridtown USA, besides having excellent donut shoppes, is known for its precisely laid out grid of streets and avenues. Streets run east-west, and avenues north-south, for the entire stretch of the town, never curving and never interrupted by parks or schools or the like.
Suppose you live on the corner of 1st and 1st and work on the corner of 12th and 12th. Thus you must travel 22 blocks to get to work as quickly as possible.

(a) How many different routes can you take to work, assuming you want to get there as quickly as possible?
(b) Now suppose you want to stop and get a donut on the way to work, from your favorite donut shoppe on the corner of 8th st and 10th ave. How many routes to work, via the donut shoppe, can you take (again, ensuring the shortest possible route)?
(c) Disaster Strikes Gridtown: there is a pothole on 4th avenue between 5th and 6th street. How many routes to work can you take avoiding that unsightly (and dangerous) stretch of road?
(d) How many routes are there both avoiding the pothole and visiting the donut shoppe?

12. Suppose you are ordering a large pizza from *D.P. Dough*. You want 3 distinct toppings, chosen from their list of 11 vegetarian toppings.

(a) How many choices do you have for your pizza?
(b) How many choices do you have for your pizza if you refuse to have pineapple as one of your toppings?
(c) How many choices do you have for your pizza if you *insist* on having pineapple as one of your toppings?
(d) How do the three questions above relate to each other?

13. Explain why the coefficient of x^5y^3 the same as the coefficient of x^3y^5 in the expansion of $(x + y)^8$?

1.3 Combinations and Permutations

Investigate!

You have a bunch of chips which come in five different colors: red, blue, green, purple and yellow.

1. How many different two-chip stacks can you make if the bottom chip must be red or blue? Explain your answer using both the additive and multiplicative principles.
2. How many different three-chip stacks can you make if the bottom chip must be red or blue and the top chip must be green, purple or yellow? How does this problem relate to the previous one?
3. How many different three-chip stacks are there in which no color is repeated? What about four-chip stacks?
4. Suppose you wanted to take three different colored chips and put them in your pocket. How many different choices do you have? What if you wanted four different colored chips? How do these problems relate to the previous one?

Attempt the above activity before proceeding

A **permutation** is a (possible) rearrangement of objects. For example, there are 6 permutations of the letters a, b, c:

$$abc, \quad acb, \quad bac, \quad bca, \quad cab, \quad cba.$$

We know that we have them all listed above —there are 3 choices for which letter we put first, then 2 choices for which letter comes next, which leaves only 1 choice for the last letter. The multiplicative principle says we multiply $3 \cdot 2 \cdot 1$.

Example 1.3.1
How many permutations are there of the letters a, b, c, d, e, f?

Solution. We do NOT want to try to list all of these out. However, if we did, we would need to pick a letter to write down first. There are 6 choices for that letter. For each choice of first letter, there are 5 choices for the second letter (we cannot repeat the first letter; we are rearranging letters and only have one of each), and for each of those, there are 4 choices for the third, 3 choices for the fourth, 2 choices for the fifth and finally only 1 choice for the last letter. So there are $6 \cdot 5 \cdot 4 \cdot 3 \cdot 2 \cdot 1 = 720$ permutations of the 6 letters.

A piece of notation is helpful here: $n!$, read "n factorial", is the product of all positive integers less than or equal to n (for reasons of convenience, we also define $0!$ to be 1). So the number of permutation of 6 letters, as seen in the previous example is $6! = 6 \cdot 5 \cdot 4 \cdot 3 \cdot 2 \cdot 1$. This generalizes:

Permutations of n elements

There are $n! = n \cdot (n-1) \cdot (n-2) \cdot \cdots \cdot 2 \cdot 1$ permutations of n (distinct) elements.

Example 1.3.2: Counting Bijective Functions
How many functions $f : \{1, 2, \ldots, 8\} \to \{1, 2, \ldots, 8\}$ are *bijective*?

Solution. Remember what it means for a function to be bijective: each element in the codomain must be the image of exactly one element of the domain. Using two-line notation, we could write one of these bijections as

$$f = \begin{pmatrix} 1 & 2 & 3 & 4 & 5 & 6 & 7 & 8 \\ 3 & 1 & 5 & 8 & 7 & 6 & 2 & 4 \end{pmatrix}$$

What we are really doing is just rearranging the elements of the codomain, so we are creating a permutation of 8 elements. In fact, "permutation" is another term used to describe bijective functions from a finite set to itself.

If you believe this, then you see the answer must be $8! = 8 \cdot 7 \cdot \cdots \cdot 1 = 40320$. You can see this directly as well: for each element of the domain, we must pick a distinct element of the codomain to map to. There are 8 choices for where to send 1, then 7 choices for where to send 2, and so on. We multiply using the multiplicative principle.

Sometimes we do not want to permute all of the letters/numbers/elements we are given.

Example 1.3.3
How many 4 letter "words" can you make from the letters a through f, with no repeated letters?

Solution. This is just like the problem of permuting 4 letters, only now we have more choices for each letter. For the first letter, there are 6 choices. For each of those, there are 5 choices for the second letter. Then there are 4 choices for the third letter, and 3 choices for the last letter. The total number of words is $6 \cdot 5 \cdot 4 \cdot 3 = 360$. This is not 6! because we never multiplied by 2 and 1. We could start with 6! and then cancel the 2 and 1, and thus write $\frac{6!}{2!}$.

In general, we can ask how many permutations exist of k objects choosing those objects from a larger collection of n objects. (In the example above, $k = 4$, and $n = 6$.) We write this number $P(n, k)$ and sometimes call it a **k-permutation of n elements**. From the example above, we see that to compute $P(n, k)$ we must apply the multiplicative principle to k numbers, starting with n and counting backwards. For example

$$P(10, 4) = 10 \cdot 9 \cdot 8 \cdot 7.$$

Notice again that $P(10, 4)$ starts out looking like 10!, but we stop after 7. We can formally account for this "stopping" by dividing away the part of the factorial we do not want:

$$P(10, 4) = \frac{10 \cdot 9 \cdot 8 \cdot 7 \cdot 6 \cdot 5 \cdot 4 \cdot 3 \cdot 2 \cdot 1}{6 \cdot 5 \cdot 4 \cdot 3 \cdot 2 \cdot 1} = \frac{10!}{6!}.$$

Careful: The factorial in the denominator is not 4! but rather $(10 - 4)!$.

k-permutations of n elements

$P(n, k)$ is the number of **k-permutations of n elements**, the number of ways to *arrange* k objects chosen from n distinct objects.

$$P(n, k) = \frac{n!}{(n - k)!}.$$

Note that when $n = k$, we have $P(n, n) = \frac{n!}{(n-n)!} = n!$ (since we defined 0! to be 1). This makes sense —we already know $n!$ gives the number of permutations of all n objects.

Example 1.3.4: Counting injective functions
How many functions $f : \{1, 2, 3\} \to \{1, 2, 3, 4, 5, 6, 7, 8\}$ are *injective*?

Solution. Note that it doesn't make sense to ask for the number of *bijections* here, as there are none (because the codomain is larger than the domain, there are no surjections). But for a function to be injective, we just can't use an element of the codomain more than once.

We need to pick an element from the codomain to be the image of 1. There are 8 choices. Then we need to pick one of the remaining 7 elements to be the image of 2. Finally, one of the remaining 6 elements must be the image of 3. So the total number of functions is $8 \cdot 7 \cdot 6 = P(8, 3)$.

What this demonstrates in general is that the number of injections $f : A \to B$, where $|A| = k$ and $|B| = n$, is $P(n, k)$.

Here is another way to find the number of k-permutations of n elements: first select which k elements will be in the permutation, then count how many ways there are to arrange them. Once you have selected the k objects, we know there are $k!$ ways to arrange (permute) them. But how do you select k objects from the n? You have n objects, and you need to *choose* k of them. You can do that in $\binom{n}{k}$ ways. Then for each choice of those k elements, we can permute *them* in $k!$ ways. Using the multiplicative principle, we get another formula for $P(n,k)$:

$$P(n,k) = \binom{n}{k} \cdot k!.$$

Now since we have a closed formula for $P(n,k)$ already, we can substitute that in:

$$\frac{n!}{(n-k)!} = \binom{n}{k} \cdot k!.$$

If we divide both sides by $k!$ we get a closed formula for $\binom{n}{k}$.

Closed formula for $\binom{n}{k}$

$$\binom{n}{k} = \frac{n!}{(n-k)!k!}$$

We say $P(n,k)$ counts *permutations*, and $\binom{n}{k}$ counts *combinations*. The formulas for each are very similar, there is just an extra $k!$ in the denominator of $\binom{n}{k}$. That extra $k!$ accounts for the fact that $\binom{n}{k}$ does not distinguish between the different orders that the k objects can appear in. We are just selecting (or choosing) the k objects, not arranging them. Perhaps "combination" is a misleading label. We don't mean it like a combination lock (where the order would definitely matter). Perhaps a better metaphor is a combination of flavors — you just need to decide which flavors to combine, not the order in which to combine them.

To further illustrate the connection between combinations and permutations, we close with an example.

Example 1.3.5
You decide to have a dinner party. Even though you are incredibly popular and have 14 different friends, you only have enough chairs to invite 6 of them.

1. How many choices do you have for which 6 friends to invite?
2. What if you need to decide not only which friends to invite but also where to seat them along your long table? How many choices do you have then?

Solution.

1. You must simply choose 6 friends from a group of 14. This can be done in $\binom{14}{6}$ ways. We can find this number either by using Pascal's triangle or the closed formula: $\frac{14!}{8!\cdot 6!} = 3003$.
2. Here you must count all the ways you can permute 6 friends chosen from a group of 14. So the answer is $P(14,6)$, which can be calculated as $\frac{14!}{8!} = 2192190$.

   Notice that we can think of this counting problem as a question about counting functions: how many injective functions are there from your set of 6 chairs to your set of 14 friends (the functions are injective because you can't have a single chair go to two of your friends).

How are these numbers related? Notice that $P(14,6)$ is *much* larger than $\binom{14}{6}$. This makes sense. $\binom{14}{6}$ picks 6 friends, but $P(14,6)$ arranges the 6 friends as well as picks them. In fact, we can say exactly how much larger $P(14,6)$ is. In both counting problems we choose 6 out of 14 friends. For the first one, we stop there, at 3003 ways. But for the second counting problem, each of those 3003 choices of 6 friends can be arranged in exactly 6! ways. So now we have $3003 \cdot 6!$ choices and that is exactly 2192190.

Alternatively, look at the first problem another way. We want to select 6 out of 14 friends, but we do not care about the order they are selected in. To select 6 out of 14 friends, we might try this:

$$14 \cdot 13 \cdot 12 \cdot 11 \cdot 10 \cdot 9.$$

This is a reasonable guess, since we have 14 choices for the first guest, then 13 for the second, and so on. But the guess is wrong (in fact, that product is exactly $2192190 = P(14, 6)$). It distinguishes between the different orders in which we could invite the guests. To correct for this, we could divide by the number of different arrangements of the 6 guests (so that all of these would count as just one outcome). There are precisely 6! ways to arrange 6 guests, so the correct answer to the first question is

$$\frac{14 \cdot 13 \cdot 12 \cdot 11 \cdot 10 \cdot 9}{6!}.$$

Note that another way to write this is

$$\frac{14!}{8! \cdot 6!}.$$

which is what we had originally.

Exercises

1. A pizza parlor offers 10 toppings.

(a) How many 3-topping pizzas could they put on their menu? Assume double toppings are not allowed.

(b) How many total pizzas are possible, with between zero and ten toppings (but not double toppings) allowed?

(c) The pizza parlor will list the 10 toppings in two equal-sized columns on their menu. How many ways can they arrange the toppings in the left column?

2. A combination lock consists of a dial with 40 numbers on it. To open the lock, you turn the dial to the right until you reach a first number, then to the left until you get to second number, then to the right again to the third number. The numbers must be distinct. How many different combinations are possible?

3. Using the digits 2 through 8, find the number of different 5-digit numbers such that:

(a) Digits can be used more than once.

(b) Digits cannot be repeated, but can come in any order.

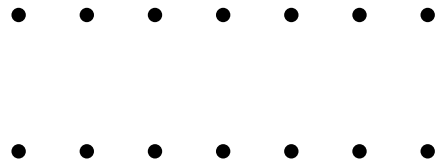
(c) Digits cannot be repeated and must be written in increasing order.

(d) Which of the above counting questions is a combination and which is a permutation? Explain why this makes sense.

4. How many triangles are there with vertices from the points shown below? Note, we are not allowing degenerate triangles - ones with all three vertices on the same line, but we do allow non-right triangles. Explain why your answer is correct.

Hint. You need exactly two points on either the x- or y-axis, but don't over-count the right triangles.

5. How many quadrilaterals can you draw using the dots below as vertices (corners)?



6. How many of the quadrilaterals possible in the previous problem are:

(a) Squares?

(b) Rectangles?

(c) Parallelograms?

(d) Trapezoids?[2]

(e) Trapezoids that are not parallelograms?

7. An *anagram* of a word is just a rearrangement of its letters. How many different anagrams of "uncopyrightable" are there? (This happens to be the longest common English word without any repeated letters.)

8. How many anagrams are there of the word "assesses" that start with the letter "a"?

9. How many anagrams are there of "anagram"?

10. On a business retreat, your company of 20 businessmen and businesswomen go golfing.

(a) You need to divide up into foursomes (groups of 4 people): a first foursome, a second foursome, and so on. How many ways can you do this?

(b) After all your hard work, you realize that in fact, you want each foursome to include one of the five Board members. How many ways can you do this?

11. How many different seating arrangements are possible for King Arthur and his 9 knights around their round table?

12. Consider sets A and B with $|A| = 10$ and $|B| = 17$.

(a) How many functions $f : A \to B$ are there?

(b) How many functions $f : A \to B$ are injective?

13. Consider functions $f : \{1, 2, 3, 4\} \to \{1, 2, 3, 4, 5, 6\}$.

(a) How many functions are there total?

(b) How many functions are injective?

(c) How many of the injective functions are *increasing*? To be increasing means that if $a < b$ then $f(a) < f(b)$, or in other words, the outputs get larger as the inputs get larger.

14. We have seen that the formula for $P(n, k)$ is $\dfrac{n!}{(n-k)!}$. Your task here is to explain *why* this is the right formula.

(a) Suppose you have 12 chips, each a different color. How many different stacks of 5 chips can you make? Explain your answer and why it is the same as using the formula for $P(12, 5)$.

(b) Using the scenario of the 12 chips again, what does 12! count? What does 7! count? Explain.

(c) Explain why it makes sense to divide 12! by 7! when computing $P(12, 5)$ (in terms of the chips).

(d) Does your explanation work for numbers other than 12 and 5? Explain the formula $P(n, k) = \frac{n!}{(n-k)!}$ using the variables n and k.

[2]Here, as in calculus, a trapezoid is defined as a quadrilateral with *at least* one pair of parallel sides. In particular, parallelograms are trapezoids.

1.4 Combinatorial Proofs

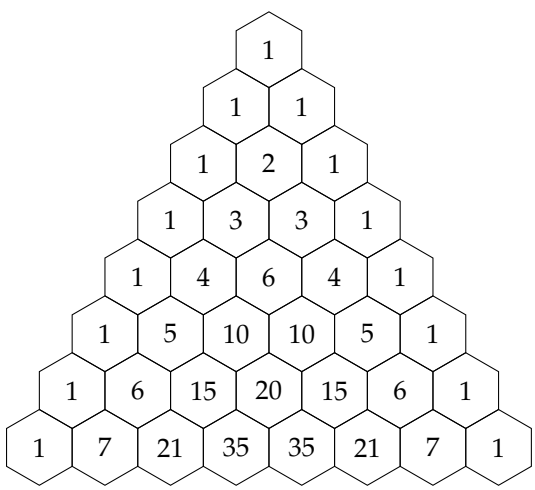
Investigate!

1. The Stanley Cup is decided in a best of 7 tournament between two teams. In how many ways can your team win? Let's answer this question two ways:
   (a) How many of the 7 games does your team need to win? How many ways can this happen?
   (b) What if the tournament goes all 7 games? So you win the last game. How many ways can the first 6 games go down?
   (c) What if the tournament goes just 6 games? How many ways can this happen? What about 5 games? 4 games?
   (d) What are the two different ways to compute the number of ways your team can win? Write down an equation involving binomial coefficients (that is, $\binom{n}{k}$'s). What pattern in Pascal's triangle is this an example of?
2. Generalize. What if the rules changed and you played a best of 9 tournament (5 wins required)? What if you played an n game tournament with k wins required to be named champion?

Attempt the above activity before proceeding

Patterns in Pascal's Triangle

Have a look again at Pascal's triangle. Forget for a moment where it comes from. Just look at it as a mathematical object. What do you notice?



There are lots of patterns hidden away in the triangle, enough to fill a reasonably sized book. Here are just a few of the most obvious ones:

1. The entries on the border of the triangle are all 1.
2. Any entry not on the border is the sum of the two entries above it.
3. The triangle is symmetric. In any row, entries on the left side are mirrored on the right side.
4. The sum of all entries on a given row is a power of 2. (You should check this!)

We would like to state these observations in a more precise way, and then prove that they are correct. Now each entry in Pascal's triangle is in fact a binomial coefficient. The 1 on the very top of the triangle is $\binom{0}{0}$.

The next row (which we will call row 1, even though it is not the top-most row) consists of $\binom{1}{0}$ and $\binom{1}{1}$. Row 4 (the row 1, 4, 6, 4, 1) consists of the binomial coefficients

$$\binom{4}{0}\ \binom{4}{1}\ \binom{4}{2}\ \binom{4}{3}\ \binom{4}{4}.$$

Given this description of the elements in Pascal's triangle, we can rewrite the above observations as follows:

1. $\binom{n}{0} = 1$ and $\binom{n}{n} = 1$.
2. $\binom{n}{k} = \binom{n-1}{k-1} + \binom{n-1}{k}$.
3. $\binom{n}{k} = \binom{n}{n-k}$.
4. $\binom{n}{0} + \binom{n}{1} + \binom{n}{2} + \cdots + \binom{n}{n} = 2^n$.

Each of these is an example of a **binomial identity**: an identity (i.e., equation) involving binomial coefficients.

Our goal is to establish these identities. We wish to prove that they hold for all values of n and k. These proofs can be done in many ways. One option would be to give algebraic proofs, using the formula for $\binom{n}{k}$:

$$\binom{n}{k} = \frac{n!}{(n-k)!\,k!}.$$

Here's how you might do that for the second identity above.

Example 1.4.1
Give an algebraic proof for the binomial identity

$$\binom{n}{k} = \binom{n-1}{k-1} + \binom{n-1}{k}.$$

Solution.

Proof. By the definition of $\binom{n}{k}$, we have

$$\binom{n-1}{k-1} = \frac{(n-1)!}{(n-1-(k-1))!(k-1)!} = \frac{(n-1)!}{(n-k)!(k-1)!}$$

and

$$\binom{n-1}{k} = \frac{(n-1)!}{(n-1-k)!k!}.$$

Thus, starting with the right-hand side of the equation:

$$\begin{aligned}
\binom{n-1}{k-1} + \binom{n-1}{k} &= \frac{(n-1)!}{(n-k)!(k-1)!} + \frac{(n-1)!}{(n-1-k)!\,k!} \\
&= \frac{(n-1)!k}{(n-k)!\,k!} + \frac{(n-1)!(n-k)}{(n-k)!\,k!} \\
&= \frac{(n-1)!(k+n-k)}{(n-k)!\,k!} \\
&= \frac{n!}{(n-k)!\,k!} \\
&= \binom{n}{k}.
\end{aligned}$$

The second line (where the common denominator is found) works because $k(k-1)! = k!$ and $(n-k)(n-k-1)! = (n-k)!$. QED

This is certainly a valid proof, but also is entirely useless. Even if you understand the proof perfectly, it does not tell you *why* the identity is true. A better approach would be to explain what $\binom{n}{k}$ *means* and then say why that is also what $\binom{n-1}{k-1} + \binom{n-1}{k}$ means. Let's see how this works for the four identities we observed above.

Example 1.4.2
Explain why $\binom{n}{0} = 1$ and $\binom{n}{n} = 1$.

Solution. What do these binomial coefficients tell us? Well, $\binom{n}{0}$ gives the number of ways to select 0 objects from a collection of n objects. There is only one way to do this, namely to not select any of the objects. Thus $\binom{n}{0} = 1$. Similarly, $\binom{n}{n}$ gives the number of ways to select n objects from a collection of n objects. There is only one way to do this: select all n objects. Thus $\binom{n}{n} = 1$.

Alternatively, we know that $\binom{n}{0}$ is the number of n-bit strings with weight 0. There is only one such string, the string of all 0's. So $\binom{n}{0} = 1$. Similarly $\binom{n}{n}$ is the number of n-bit strings with weight n. There is only one string with this property, the string of all 1's.

Another way: $\binom{n}{0}$ gives the number of subsets of a set of size n containing 0 elements. There is only one such subset, the empty set. $\binom{n}{n}$ gives the number of subsets containing n elements. The only such subset is the original set (of all elements).

Example 1.4.3
Explain why $\binom{n}{k} = \binom{n-1}{k-1} + \binom{n-1}{k}$.

Solution. The easiest way to see this is to consider bit strings. $\binom{n}{k}$ is the number of bit strings of length n containing k 1's. Of all of these strings, some start with a 1 and the rest start with a 0. First consider all the bit strings which start with a 1. After the 1, there must be $n-1$ more bits (to get the total length up to n) and exactly $k-1$ of them must be 1's (as we already have one, and we need k total). How many strings are there like that? There are exactly $\binom{n-1}{k-1}$ such bit strings, so of all the length n bit strings containing k 1's, $\binom{n-1}{k-1}$ of them start with a 1. Similarly, there are $\binom{n-1}{k}$ which start with a 0 (we still need $n-1$ bits and now k of them must be 1's). Since there are $\binom{n-1}{k}$ bit strings containing $n-1$ bits with k 1's, that is the number of length n bit strings with k 1's which start with a 0. Therefore $\binom{n}{k} = \binom{n-1}{k-1} + \binom{n-1}{k}$.

Another way: consider the question, how many ways can you select k pizza toppings from a menu containing n choices? One way to do this is just $\binom{n}{k}$. Another way to answer the same question is to first decide whether or not you want anchovies. If you do want anchovies, you still need to pick $k-1$ toppings, now from just $n-1$ choices. That can be done in $\binom{n-1}{k-1}$ ways. If you do not want anchovies, then you still need to select k toppings from $n-1$ choices (the anchovies are out). You can do that in $\binom{n-1}{k}$ ways. Since the choices with anchovies are disjoint from the choices without anchovies, the total choices are $\binom{n-1}{k-1} + \binom{n-1}{k}$. But wait. We answered the same question in two different ways, so the two answers must be the same. Thus $\binom{n}{k} = \binom{n-1}{k-1} + \binom{n-1}{k}$.

You can also explain (prove) this identity by counting subsets, or even lattice paths.

Example 1.4.4
Prove the binomial identity $\binom{n}{k} = \binom{n}{n-k}$.

Solution. Why is this true? $\binom{n}{k}$ counts the number of ways to select k things from n choices. On the other hand, $\binom{n}{n-k}$ counts the number of ways to select $n-k$ things from n choices. Are these really the same? Well, what if instead of selecting the $n-k$ things you choose to exclude them. How many ways are there to choose $n-k$ things to exclude from n choices. Clearly this is $\binom{n}{n-k}$ as well (it doesn't matter whether you include or exclude the things once you have chosen them). And if you exclude $n-k$ things, then you are including the other k things. So the set of outcomes should be the same.

Let's try the pizza counting example like we did above. How many ways are there to pick k toppings from a list of n choices? On the one hand, the answer is simply $\binom{n}{k}$. Alternatively, you

could make a list of all the toppings you don't want. To end up with a pizza containing exactly k toppings, you need to pick $n-k$ toppings to not put on the pizza. You have $\binom{n}{n-k}$ choices for the toppings you don't want. Both of these ways give you a pizza with k toppings, in fact all the ways to get a pizza with k toppings. Thus these two answers must be the same: $\binom{n}{k} = \binom{n}{n-k}$.

You can also prove (explain) this identity using bit strings, subsets, or lattice paths. The bit string argument is nice: $\binom{n}{k}$ counts the number of bit strings of length n with k 1's. This is also the number of bit string of length n with k 0's (just replace each 1 with a 0 and each 0 with a 1). But if a string of length n has k 0's, it must have $n-k$ 1's. And there are exactly $\binom{n}{n-k}$ strings of length n with $n-k$ 1's.

Example 1.4.5
Prove the binomial identity $\binom{n}{0} + \binom{n}{1} + \binom{n}{2} + \cdots + \binom{n}{n} = 2^n$.

Solution. Let's do a "pizza proof" again. We need to find a question about pizza toppings which has 2^n as the answer. How about this: If a pizza joint offers n toppings, how many pizzas can you build using any number of toppings from no toppings to all toppings, using each topping at most once?

On one hand, the answer is 2^n. For each topping you can say "yes" or "no," so you have two choices for each topping.

On the other hand, divide the possible pizzas into disjoint groups: the pizzas with no toppings, the pizzas with one topping, the pizzas with two toppings, etc. If we want no toppings, there is only one pizza like that (the empty pizza, if you will) but it would be better to think of that number as $\binom{n}{0}$ since we choose 0 of the n toppings. How many pizzas have 1 topping? We need to choose 1 of the n toppings, so $\binom{n}{1}$. We have:

- Pizzas with 0 toppings: $\binom{n}{0}$
- Pizzas with 1 topping: $\binom{n}{1}$
- Pizzas with 2 toppings: $\binom{n}{2}$
- $\vdots$
- Pizzas with n toppings: $\binom{n}{n}$.

The total number of possible pizzas will be the sum of these, which is exactly the left-hand side of the identity we are trying to prove.

Again, we could have proved the identity using subsets, bit strings, or lattice paths (although the lattice path argument is a little tricky).

Hopefully this gives some idea of how explanatory proofs of binomial identities can go. It is worth pointing out that more traditional proofs can also be beautiful. [3] For example, consider the following rather slick proof of the last identity.

Expand the binomial $(x+y)^n$:

$$(x+y)^n = \binom{n}{0}x^n + \binom{n}{1}x^{n-1}y + \binom{n}{2}x^{n-2}y^2 + \cdots + \binom{n}{n-1}x \cdot y^n + \binom{n}{n}y^n.$$

Let $x = 1$ and $y = 1$. We get:

$$(1+1)^n = \binom{n}{0}1^n + \binom{n}{1}1^{n-1}1 + \binom{n}{2}1^{n-2}1^2 + \cdots + \binom{n}{n-1}1 \cdot 1^n + \binom{n}{n}1^n.$$

Of course this simplifies to:

$$(2)^n = \binom{n}{0} + \binom{n}{1} + \binom{n}{2} + \cdots + \binom{n}{n-1} + \binom{n}{n}.$$

[3] Most every binomial identity can be proved using mathematical induction, using the recursive definition for $\binom{n}{k}$. We will discuss induction in Section 2.5.

Something fun to try: Let $x = 1$ and $y = 2$. Neat huh?

More Proofs

The explanatory proofs given in the above examples are typically called **combinatorial proofs**. In general, to give a combinatorial proof for a binomial identity, say $A = B$ you do the following:

1. Find a counting problem you will be able to answer in two ways.
2. Explain why one answer to the counting problem is A.
3. Explain why the other answer to the counting problem is B.

Since both A and B are the answers to the same question, we must have $A = B$.

The tricky thing is coming up with the question. This is not always obvious, but it gets easier the more counting problems you solve. You will start to recognize types of answers as the answers to types of questions. More often what will happen is you will be solving a counting problem and happen to think up two different ways of finding the answer. Now you have a binomial identity and the proof is right there. The proof *is* the problem you just solved together with your two solutions.

For example, consider this counting question:

How many 10-letter words use exactly four A's, three B's, two C's and one D?

Let's try to solve this problem. We have 10 spots for letters to go. Four of those need to be A's. We can pick the four A-spots in $\binom{10}{4}$ ways. Now where can we put the B's? Well there are only 6 spots left, we need to pick 3 of them. This can be done in $\binom{6}{3}$ ways. The two C's need to go in two of the 3 remaining spots, so we have $\binom{3}{2}$ ways of doing that. That leaves just one spot of the D, but we could write that 1 choice as $\binom{1}{1}$. Thus the answer is:

$$\binom{10}{4}\binom{6}{3}\binom{3}{2}\binom{1}{1}.$$

But why stop there? We can find the answer another way too. First let's decide where to put the one D: we have 10 spots, we need to choose 1 of them, so this can be done in $\binom{10}{1}$ ways. Next, choose one of the $\binom{9}{2}$ ways to place the two C's. We now have 7 spots left, and three of them need to be filled with B's. There are $\binom{7}{3}$ ways to do this. Finally the A's can be placed in $\binom{4}{4}$ (that is, only one) ways. So another answer to the question is

$$\binom{10}{1}\binom{9}{2}\binom{7}{3}\binom{4}{4}.$$

Interesting. This gives us the binomial identity:

$$\binom{10}{4}\binom{6}{3}\binom{3}{2}\binom{1}{1} = \binom{10}{1}\binom{9}{2}\binom{7}{3}\binom{4}{4}.$$

Here are a couple of other binomial identities with combinatorial proofs.

Example 1.4.6
Prove the identity

$$1n + 2(n-1) + 3(n-2) + \cdots + (n-1)2 + n1 = \binom{n+2}{3}.$$

Solution. To give a combinatorial proof we need to think up a question we can answer in two ways: one way needs to give the left-hand-side of the identity, the other way needs to be the right-hand-side of the identity. Our clue to what question to ask comes from the right-hand side: $\binom{n+2}{3}$ counts the number of ways to select 3 things from a group of $n + 2$ things. Let's name those things $1, 2, 3, \ldots, n + 2$. In other words, we want to find 3-element subsets of those numbers (since order should not matter, subsets are exactly the right thing to think about). We will have to be a bit clever to explain why the left-hand-side also gives the number of these subsets. Here's the proof.

Proof. Consider the question "How many 3-element subsets are there of the set $\{1, 2, 3, \ldots, n+2\}$?" We answer this in two ways:

Answer 1: We must select 3 elements from the collection of $n+2$ elements. This can be done in $\binom{n+2}{3}$ ways.

Answer 2: Break this problem up into cases by what the middle number in the subset is. Say each subset is $\{a, b, c\}$ written in increasing order. We count the number of subsets for each distinct value of b. The smallest possible value of b is 2, and the largest is $n+1$.

When $b = 2$, there are $1 \cdot n$ subsets: 1 choice for a and n choices (3 through $n+2$) for c.

When $b = 3$, there are $2 \cdot (n-1)$ subsets: 2 choices for a and $n-1$ choices for c.

When $b = 4$, there are $3 \cdot (n-2)$ subsets: 3 choices for a and $n-2$ choices for c.

And so on. When $b = n+1$, there are n choices for a and only 1 choice for c, so $n \cdot 1$ subsets.

Therefore the total number of subsets is

$$1n + 2(n-1) + 3(n-2) + \cdots + (n-1)2 + n1.$$

Since Answer 1 and Answer 2 are answers to the same question, they must be equal. Therefore

$$1n + 2(n-1) + 3(n-2) + \cdots + (n-1)2 + n1 = \binom{n+2}{3}.$$

QED

Example 1.4.7

Prove the binomial identity

$$\binom{n}{0}^2 + \binom{n}{1}^2 + \binom{n}{2}^2 + \cdots + \binom{n}{n}^2 = \binom{2n}{n}.$$

Solution. We will give two different proofs of this fact. The first will be very similar to the previous example (counting subsets). The second proof is a little slicker, using lattice paths.

Proof. Consider the question: "How many pizzas can you make using n toppings when there are $2n$ toppings to choose from?"

Answer 1: There are $2n$ toppings, from which you must choose n. This can be done in $\binom{2n}{n}$ ways.

Answer 2: Divide the toppings into two groups of n toppings (perhaps n meats and n veggies). Any choice of n toppings must include some number from the first group and some number from the second group. Consider each possible number of meat toppings separately:

0 meats: $\binom{n}{0}\binom{n}{n}$, since you need to choose 0 of the n meats and n of the n veggies.

1 meat: $\binom{n}{1}\binom{n}{n-1}$, since you need 1 of n meats so $n-1$ of n veggies.

2 meats: $\binom{n}{2}\binom{n}{n-2}$. Choose 2 meats and the remaining $n-2$ toppings from the n veggies.

And so on. The last case is n meats, which can be done in $\binom{n}{n}\binom{n}{0}$ ways.

Thus the total number of pizzas possible is

$$\binom{n}{0}\binom{n}{n} + \binom{n}{1}\binom{n}{n-1} + \binom{n}{2}\binom{n}{n-2} + \cdots + \binom{n}{n}\binom{n}{0}.$$

This is not quite the left-hand side ... yet. Notice that $\binom{n}{n} = \binom{n}{0}$ and $\binom{n}{n-1} = \binom{n}{1}$ and so on, by the identity in Example 1.4.4. Thus we do indeed get

$$\binom{n}{0}^2 + \binom{n}{1}^2 + \binom{n}{2}^2 + \cdots + \binom{n}{n}^2.$$

Since these two answers are answers to the same question, they must be equal, and thus

$$\binom{n}{0}^2 + \binom{n}{1}^2 + \binom{n}{2}^2 + \cdots + \binom{n}{n}^2 = \binom{2n}{n}.$$

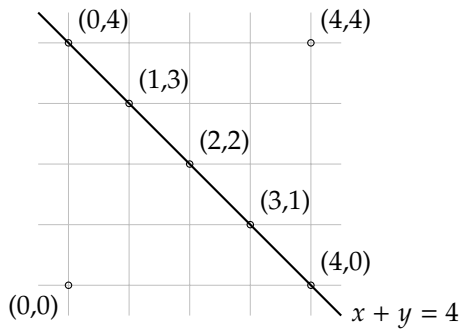
QED

For an alternative proof, we use lattice paths. This is reasonable to consider because the right-hand side of the identity reminds us of the number of paths from $(0,0)$ to (n,n).

Proof. Consider the question: How many lattice paths are there from $(0,0)$ to (n,n)?

Answer 1: We must travel $2n$ steps, and n of them must be in the up direction. Thus there are $\binom{2n}{n}$ paths.

Answer 2: Note that any path from $(0,0)$ to (n,n) must cross the line $x+y=n$. That is, any path must pass through exactly one of the points: $(0,n)$, $(1,n-1)$, $(2,n-2)$, $\ldots$, $(n,0)$. For example, this is what happens in the case $n=4$:



How many paths pass through $(0,n)$? To get to that point, you must travel n units, and 0 of them are to the right, so there are $\binom{n}{0}$ ways to get to $(0,n)$. From $(0,n)$ to (n,n) takes n steps, and 0 of them are up. So there are $\binom{n}{0}$ ways to get from $(0,n)$ to (n,n). Therefore there are $\binom{n}{0}\binom{n}{0}$ paths from $(0,0)$ to (n,n) through the point $(0,n)$.

What about through $(1,n-1)$. There are $\binom{n}{1}$ paths to get there (n steps, 1 to the right) and $\binom{n}{1}$ paths to complete the journey to (n,n) (n steps, 1 up). So there are $\binom{n}{1}\binom{n}{1}$ paths from $(0,0)$ to (n,n) through $(1,n-1)$.

In general, to get to (n,n) through the point $(k,n-k)$ we have $\binom{n}{k}$ paths to the midpoint and then $\binom{n}{k}$ paths from the midpoint to (n,n). So there are $\binom{n}{k}\binom{n}{k}$ paths from $(0,0)$ to (n,n) through $(k,n-k)$.

All together then the total paths from $(0,0)$ to (n,n) passing through exactly one of these midpoints is

$$\binom{n}{0}^2+\binom{n}{1}^2+\binom{n}{2}^2+\cdots+\binom{n}{n}^2.$$

Since these two answers are answers to the same question, they must be equal, and thus

$$\binom{n}{0}^2+\binom{n}{1}^2+\binom{n}{2}^2+\cdots+\binom{n}{n}^2=\binom{2n}{n}.$$

QED

Exercises

1. Prove the identity $\binom{n}{k}=\binom{n-1}{k-1}+\binom{n-1}{k}$ using a question about subsets.

2. Give a combinatorial proof of the identity $2+2+2=3\cdot 2$.

3. Give a combinatorial proof for the identity $1+2+3+\cdots+n=\binom{n+1}{2}$.

4. A woman is getting married. She has 15 best friends but can only select 6 of them to be her bridesmaids, one of which needs to be her maid of honor. How many ways can she do this?

(a) What if she first selects the 6 bridesmaids, and then selects one of them to be the maid of honor?

(b) What if she first selects her maid of honor, and then 5 other bridemaids?

(c) Explain why $6\binom{15}{6}=15\binom{14}{5}$.

5. Give a combinatorial proof of the identity $\binom{n}{2}\binom{n-2}{k-2} = \binom{n}{k}\binom{k}{2}$.

6. Consider the bit strings in $\mathbf{B}_2^6$ (bit strings of length 6 and weight 2).

(a) How many of those bit strings start with 1?

(b) How many of those bit strings start with 01?

(c) How many of those bit strings start with 001?

(d) Are there any other strings we have not counted yet? Which ones, and how many are there?

(e) How many bit strings are there total in $\mathbf{B}_2^6$?

(f) What binomial identity have you just given a combinatorial proof for?

7. Let's count **ternary** digit strings, that is, strings in which each digit can be 0, 1, or 2.

(a) How many ternary digit strings contain exactly n digits?

(b) How many ternary digit strings contain exactly n digits and n 2's.

(c) How many ternary digit strings contain exactly n digits and $n-1$ 2's. (Hint: where can you put the non-2 digit, and then what could it be?)

(d) How many ternary digit strings contain exactly n digits and $n-2$ 2's. (Hint: see previous hint)

(e) How many ternary digit strings contain exactly n digits and $n-k$ 2's.

(f) How many ternary digit strings contain exactly n digits and no 2's. (Hint: what kind of a string is this?)

(g) Use the above parts to give a combinatorial proof for the identity

$$\binom{n}{0} + 2\binom{n}{1} + 2^2\binom{n}{2} + 2^3\binom{n}{3} + \cdots + 2^n\binom{n}{n} = 3^n.$$

8. How many ways are there to rearrange the letters in the word "rearrange"? Answer this question in at least two different ways to establish a binomial identity.

9. Give a combinatorial proof for the identity $P(n,k) = \binom{n}{k}k!$

10. Establish the identity below using a combinatorial proof.

$$\binom{2}{2}\binom{n}{2} + \binom{3}{2}\binom{n-1}{2} + \binom{4}{2}\binom{n-2}{2} + \cdots + \binom{n}{2}\binom{2}{2} = \binom{n+3}{5}.$$

1.5 Stars and Bars

Investigate!

Suppose you have some number of identical Rubik's cubes to distribute to your friends. Imagine you start with a single row of the cubes.

1. Find the number of different ways you can distribute the cubes provided:

   (a) You have 3 cubes to give to 2 people.

   (b) You have 4 cubes to give to 2 people.

   (c) You have 5 cubes to give to 2 people.

   (d) You have 3 cubes to give to 3 people.

   (e) You have 4 cubes to give to 3 people.

   (f) You have 5 cubes to give to 3 people.

2. Make a conjecture about how many different ways you could distribute 7 cubes to 4 people. Explain.

3. What if each person were required to get *at least one* cube? How would your answers change?

Attempt the above activity before proceeding

Consider the following counting problem:

You have 7 cookies to give to 4 kids. How many ways can you do this?

Take a moment to think about how you might solve this problem. You may assume that it is acceptable to give a kid no cookies. Also, the cookies are all identical and the order in which you give out the cookies does not matter.

Before solving the problem, here is a wrong answer: You might guess that the answer should be 4^7 because for each of the 7 cookies, there are 4 choices of kids to which you can give the cookie. This is reasonable, but wrong. To see why, consider a few possible outcomes: we could assign the first six cookies to kid A, and the seventh cookie to kid B. Another outcome would assign the first cookie to kid B and the six remaining cookies to kid A. Both outcomes are included in the 4^7 answer. But for our counting problem, both outcomes are really the same – kid A gets six cookies and kid B gets one cookie.

What do outcomes actually look like? How can we represent them? One approach would be to write an outcome as a string of four numbers like this:

$$3112,$$

which represent the outcome in which the first kid gets 3 cookies, the second and third kid each get 1 cookie, and the fourth kid gets 2 cookies. Represented this way, the order in which the numbers occur matters. 1312 is a different outcome, because the first kid gets a one cookie instead of 3. Each number in the string can be any integer between 0 and 7. But the answer is not 7^4. We need the *sum* of the numbers to be 7.

Another way we might represent outcomes is to write a string of seven letters:

$$\text{ABAADCD},$$

which represents that the first cookie goes to kid A, the second cookie goes to kid B, the third and fourth cookies go to kid A, and so on. In fact, this outcome is identical to the previous one—A gets 3 cookies, B and C get 1 each and D gets 2. Each of the seven letters in the string can be any of the 4 possible letters (one for each kid), but the number of such strings is not 4^7, because here order does *not* matter. In fact, another way to write the same outcome is

$$\text{AAABCDD}.$$

This will be the preferred representation of the outcome. Since we can write the letters in any order, we might as well write them in *alphabetical* order for the purposes of counting. So we will write all the A's first, then all the B's, and so on.

Now think about how you could specify such an outcome. All we really need to do is say when to switch from one letter to the next. In terms of cookies, we need to say after how many cookies do we stop giving cookies to the first kid and start giving cookies to the second kid. And then after how many do we switch to the third kid? And after how many do we switch to the fourth? So yet another way to represent an outcome is like this:

$$***|*|*|**$$

Three cookies go to the first kid, then we switch and give one cookie to the second kid, then switch, one to the third kid, switch, two to the fourth kid. Notice that we need 7 stars and 3 bars – one star for each cookie, and one bar for each switch between kids, so one fewer bars than there are kids (we don't need to switch after the last kid – we are done).

Why have we done all of this? Simple: to count the number of ways to distribute 7 cookies to 4 kids, all we need to do is count how many *stars and bars* charts there are. But a **stars and bars chart** is just a string of symbols, some stars and some bars. If instead of stars and bars we would use 0's and 1's, it would just be a bit string. We know how to count those.

Before we get too excited, we should make sure that really *any* string of (in our case) 7 stars and 3 bars corresponds to a different way to distribute cookies to kids. In particular consider a string like this:

$$|***||****$$

Does that correspond to a cookie distribution? Yes. It represents the distribution in which kid A gets 0 cookies (because we switch to kid B before any stars), kid B gets three cookies (three stars before the next bar), kid C gets 0 cookies (no stars before the next bar) and kid D gets the remaining 4 cookies. No matter how the stars and bars are arranged, we can distribute cookies in that way. Also, given any way to distribute cookies, we can represent that with a stars and bars chart. For example, the distribution in which kid A gets 6 cookies and kid B gets 1 cookie has the following chart:

$$******|*||$$

After all that work we are finally ready to count. Each way to distribute cookies corresponds to a stars and bars chart with 7 stars and 3 bars. So there are 10 symbols, and we must choose 3 of them to be bars. Thus:

$$\text{There are } \binom{10}{3} \text{ ways to distribute 7 cookies to 4 kids.}$$

While we are at it, we can also answer a related question: how many ways are there to distribute 7 cookies to 4 kids so that each kid gets at least one cookie? What can you say about the corresponding stars and bars charts? The charts must start and end with at least one star (so that kids A and D) get cookies, and also no two bars can be adjacent (so that kids B and C are not skipped). One way to assure this is to only place bars in the spaces *between* the stars. With 7 stars, there are 6 spots between the stars, so we must choose 3 of those 6 spots to fill with bars. Thus there are $\binom{6}{3}$ ways to distribute 7 cookies to 4 kids giving at least one cookie to each kid.

Another (and more general) way to approach this modified problem is to first give each kid one cookie. Now the remaining 3 cookies can be distributed to the 4 kids without restrictions. So we have 3 stars and 3 bars for a total of 6 symbols, 3 of which must be bars. So again we see that there are $\binom{6}{3}$ ways to distribute the cookies.

Stars and bars can be used in counting problems other than kids and cookies. Here are a few examples:

Example 1.5.1
Your favorite mathematical pizza chain offers 10 toppings. How many pizzas can you make if you are allowed 6 toppings? The order of toppings does not matter but now you are allowed repeats. So one possible pizza is triple sausage, double pineapple, and onions.

Solution. We get 6 toppings (counting possible repeats). Represent each of these toppings as a star. Think of going down the menu one topping at a time: you see anchovies first, and skip to the next, sausage. You say yes to sausage 3 times (use 3 stars), then switch to the next topping on the list. You keep skipping until you get to pineapple, which you say yes to twice. Another switch and you are at onions. You say yes once. Then you keep switching until you get to the last topping, never saying yes again (since you already have said yes 6 times. There are 10 toppings to choose from, so we must switch from considering one topping to the next 9 times. These are the bars.

Now that we are confident that we have the right number of stars and bars, we answer the question simply: there are 6 stars and 9 bars, so 15 symbols. We need to pick 9 of them to be bars, so there number of pizzas possible is

$$\binom{15}{9}.$$

Example 1.5.2
How many 7 digit phone numbers are there in which the digits are non-increasing? That is, every digit is less than or equal to the previous one.

Solution. We need to decide on 7 digits so we will use 7 stars. The bars will represent a switch from each possible single digit number down the next smaller one. So the phone number 866-5221 is represented by the stars and bars chart

$$|*||**|*|||**|*|$$

There are 10 choices for each digit (0-9) so we must switch between choices 9 times. We have 7 stars and 9 bars, so the total number of phone numbers is

$$\binom{16}{9}.$$

Example 1.5.3
How many integer solutions are there to the equation

$$x_1 + x_2 + x_3 + x_4 + x_5 = 13.$$

(An **integer solution** to an equation is a solution in which the unknown must have an integer value.)

1. where $x_i \geq 0$ for each x_i?
2. where $x_i > 0$ for each x_i?
3. where $x_i \geq 2$ for each x_i?

Solution. This problem is just like giving 13 cookies to 5 kids. We need to say how many of the 13 units go to each of the 5 variables. In other words, we have 13 stars and 4 bars (the bars are like the "+" signs in the equation).

1. If x_i can be 0 or greater, we are in the standard case with no restrictions. So 13 stars and 4 bars can be arranged in $\binom{17}{4}$ ways.
2. Now each variable must be at least 1. So give one unit to each variable to satisfy that restriction. Now there are 8 stars left, and still 4 bars, so the number of solutions is $\binom{12}{4}$.
3. Now each variable must be 2 or greater. So before any counting, give each variable 2 units. We now have 3 remaining stars and 4 bars, so there are $\binom{7}{4}$ solutions.

Counting with Functions

Many of the counting problems in this section might at first appear to be examples of counting *functions*. After all, when we try to count the number of ways to distribute cookies to kids, we are assigning each cookie to a kid, just like you assign elements of the domain of a function to elements in the codomain. However, the number of ways to assign 7 cookies to 4 kids is $\binom{10}{7} = 120$, while the number of functions $f : \{1, 2, 3, 4, 5, 6, 7\} \to \{a, b, c, d\}$ is $4^7 = 16384$. What is going on here?

When we count functions, we consider the following two functions, for example, to be different:

$$f = \begin{pmatrix} 1 & 2 & 3 & 4 & 5 & 6 & 7 \\ a & b & c & c & c & c & c \end{pmatrix} \qquad g = \begin{pmatrix} 1 & 2 & 3 & 4 & 5 & 6 & 7 \\ b & a & c & c & c & c & c \end{pmatrix}.$$

But these two functions would correspond to the *same* cookie distribution: kids a and b each get one cookie, kid c gets the rest (and none for kid d).

The point: elements of the domain are distinguished, cookies are indistinguishable. This is analagous to the distinction between permutations (like counting functions) and combinations (not).

Exercises

1. A **multiset** is a collection of objects, just like a set, but can contain an object more than once (the order of the elements still doesn't matter). For example, $\{1, 1, 2, 5, 5, 7\}$ is a multiset of size 6.

(a) How many *sets* of size 5 can be made using the 10 numeric digits 0 through 9?

(b) How many *multi*sets of size 5 can be made using the 10 numeric digits 0 through 9?

2. Each of the counting problems below can be solved with stars and bars. For each, say what outcome the diagram

$$* * *| * || * *|$$

represents, if there are the correct number of stars and bars for the problem. Otherwise, say why the diagram does not represent any outcome, and what a correct diagram would look like.

(a) How many ways are there to select a handful of 6 jellybeans from a jar that contains 5 different flavors?

(b) How many ways can you distribute 5 identical lollipops to 6 kids?

(c) How many 6-letter words can you make using the 5 vowels?

(d) How many solutions are there to the equation $x_1 + x_2 + x_3 + x_4 = 6$.

3. After gym class you are tasked with putting the 14 identical dodgeballs away into 5 bins.

(a) How many ways can you do this if there are no restrictions?

(b) How many ways can you do this if each bin must contain at least one dodgeball?

4. How many integer solutions are there to the equation $x + y + z = 8$ for which

(a) x, y, and z are all positive?

(b) x, y, and z are all non-negative?

(c) x, y, and z are all greater than -3.

5. Using the digits 2 through 8, find the number of different 5-digit numbers such that:

(a) Digits cannot be repeated and must be written in increasing order. For example, 23678 is okay, but 32678 is not.

(b) Digits *can* be repeated and must be written in *non-decreasing* order. For example, 24448 is okay, but 24484 is not.

6. When playing Yahtzee, you roll five regular 6-sided dice. How many different outcomes are possible from a single roll? The order of the dice does not matter.

7. Your friend tells you she has 7 coins in her hand (just pennies, nickels, dimes and quarters). If you guess how many of each kind of coin she has, she will give them to you. If you guess randomly, what is the probability that you will be correct?

8. How many integer solutions to $x_1 + x_2 + x_3 + x_4 = 25$ are there for which $x_1 \geq 1$, $x_2 \geq 2$, $x_3 \geq 3$ and $x_4 \geq 4$?

9. Solve the three counting problems below. Then say why it makes sense that they all have the same answer. That is, say how you can interpret them as each other.

(a) How many ways are there to distribute 8 cookies to 3 kids?

(b) How many solutions in non-negative integers are there to $x + y + z = 8$?

(c) How many different packs of 8 crayons can you make using crayons that come in red, blue and yellow?

10. Consider functions $f : \{1, 2, 3, 4, 5\} \to \{0, 1, 2, \ldots, 9\}$.

(a) How many of these functions are strictly increasing? Explain. (A function is strictly increasing provided if $a < b$, then $f(a) < f(b)$.)

(b) How many of the functions are non-decreasing? Explain. (A function is non-decreasing provided if $a < b$, then $f(a) \leq f(b)$.)

11. *Conic*, your favorite math themed fast food drive-in offers 20 flavors which can be added to your soda. You have enough money to buy a large soda with 4 added flavors. How many different soda concoctions can you order if:

(a) You refuse to use any of the flavors more than once?

(b) You refuse repeats but care about the order the flavors are added?

(c) You allow yourself multiple shots of the same flavor?

(d) You allow yourself multiple shots, and care about the order the flavors are added?

1.6 Advanced Counting Using PIE

Investigate!

You have 11 identical mini key-lime pies to give to 4 children. However, you don't want any kid to get more than 3 pies. How many ways can you distribute the pies?

1. How many ways are there to distribute the pies without any restriction?
2. Let's get rid of the ways that one or more kid gets too many pies. How many ways are there to distribute the pies if Al gets too many pies? What if Bruce gets too many? Or Cat? Or Dent?
3. What if two kids get too many pies? How many ways can this happen? Does it matter which two kids you pick to overfeed?
4. Is it possible that three kids get too many pies? If so, how many ways can this happen?
5. How should you combine all the numbers you found above to answer the original question?

Suppose now you have 13 pies and 7 children. No child can have more than 2 pies. How many ways can you distribute the pies?

Attempt the above activity before proceeding

Stars and bars allows us to count the number of ways to distribute 10 cookies to 3 kids and natural number solutions to $x + y + z = 11$, for example. A relatively easy modification allows us to put a *lower bound* restriction on these problems: perhaps each kid must get at least two cookies or $x, y, z \geq 2$. This was done by first assigning each kid (or variable) 2 cookies (or units) and then distributing the rest using stars and bars.

What if we wanted an *upper bound* restriction? For example, we might insist that no kid gets more than 4 cookies or that $x, y, z \leq 4$. It turns out this is considerably harder, but still possible. The idea is to count all the distributions and then remove those that violate the condition. In other words, we must count the number of ways to distribute 11 cookies to 3 kids in which *one or more* of the kids gets more than 4 cookies. For any particular kid, this is not a problem; we do this using stars and bars. But how to combine the number of ways for kid A, or B or C? We must use the PIE.

The Principle of Inclusion/Exclusion (PIE) gives a method for finding the cardinality of the union of not necessarily disjoint sets. We saw in Subsection how this works with three sets. To find how many things are in *one or more* of the sets A, B, and C, we should just add up the number of things in each of these sets. However, if there is any overlap among the sets, those elements are counted multiple times. So we subtract the things in each intersection of a pair of sets. But doing this removes elements which are in all three sets once too often, so we need to add it back in. In terms of cardinality of sets, we have

$$|A \cup B \cup C| = |A| + |B| + |C| - |A \cap B| - |A \cap C| - |B \cap C| + |A \cap B \cap C|.$$

Example 1.6.1

Three kids, Alberto, Bernadette, and Carlos, decide to share 11 cookies. They wonder how many ways they could split the cookies up provided that none of them receive more than 4 cookies (someone receiving no cookies is for some reason acceptable to these kids).

Solution. Without the "no more than 4" restriction, the answer would be $\binom{13}{2}$, using 11 stars and 2 bars (separating the three kids). Now count the number of ways that one or more of the kids violates the condition, i.e., gets at least 4 cookies.

Let A be the set of outcomes in which Alberto gets more than 4 cookies. Let B be the set of outcomes in which Bernadette gets more than 4 cookies. Let C be the set of outcomes in which Carlos gets more than 4 cookies. We then are looking (for the sake of subtraction) for the size of the set $A \cup B \cup C$. Using PIE, we must find the sizes of $|A|$, $|B|$, $|C|$, $|A \cap B|$ and so on. Here is what we find.

- $|A| = \binom{8}{2}$. First give Alberto 5 cookies, then distribute the remaining 6 to the three kids without restrictions, using 6 stars and 2 bars.
- $|B| = \binom{8}{2}$. Just like above, only now Bernadette gets 5 cookies at the start.
- $|C| = \binom{8}{2}$. Carlos gets 5 cookies first.
- $|A \cap B| = \binom{3}{2}$. Give Alberto and Bernadette 5 cookies each, leaving 1 (star) to distribute to the three kids (2 bars).
- $|A \cap C| = \binom{3}{2}$. Alberto and Carlos get 5 cookies first.
- $|B \cap C| = \binom{3}{2}$. Bernadette and Carlos get 5 cookies first.
- $|A \cap B \cap C| = 0$. It is not possible for all three kids to get 4 or more cookies.

Combining all of these we see

$$|A \cup B \cup C| = \binom{8}{2} + \binom{8}{2} + \binom{8}{2} - \binom{3}{2} - \binom{3}{2} - \binom{3}{2} + 0 = 75.$$

Thus the answer to the original question is $\binom{13}{2} - 75 = 78 - 75 = 3$. This makes sense now that we see it. The only way to ensure that no kid gets more than 4 cookies is to give two kids 4 cookies and one kid 3; there are three choices for which kid that should be. We could have found the answer much quicker through this observation, but the point of the example is to illustrate that PIE works!

For four or more sets, we do not write down a formula for PIE. Instead, we just think of the principle: add up all the elements in single sets, then subtract out things you counted twice (elements in the intersection of a *pair* of sets), then add back in elements you removed too often (elements in the intersection of groups of three sets), then take back out elements you added back in too often (elements in the intersection of groups of four sets), then add back in, take back out, add back in, etc. This would be very difficult if it wasn't for the fact that in these problems, all the cardinalities of the single sets are equal, as are all the cardinalities of the intersections of two sets, and that of three sets, and so on. Thus we can group all of these together and multiply by how many different combinations of $1, 2, 3, \ldots$ sets there are.

Example 1.6.2
How many ways can you distribute 10 cookies to 4 kids so that no kid gets more than 2 cookies?

Solution. There are $\binom{13}{3}$ ways to distribute 10 cookies to 4 kids (using 10 stars and 3 bars). We will subtract all the outcomes in which a kid gets 3 or more cookies. How many outcomes are there like that? We can force kid A to eat 3 or more cookies by giving him 3 cookies before we start. Doing so reduces the problem to one in which we have 7 cookies to give to 4 kids without any restrictions. In that case, we have 7 stars (the 7 remaining cookies) and 3 bars (one less than the number of kids) so we can distribute the cookies in $\binom{10}{3}$ ways. Of course we could choose any one of the 4 kids to give too many cookies, so it would appear that there are $\binom{4}{1}\binom{10}{3}$ ways to distribute the cookies giving too many to one kid. But in fact, we have over counted.

We must get rid of the outcomes in which two kids have too many cookies. There are $\binom{4}{2}$ ways to select 2 kids to give extra cookies. It takes 6 cookies to do this, leaving only 4 cookies. So we have

4 stars and still 3 bars. The remaining 4 cookies can thus be distributed in $\binom{7}{3}$ ways (for each of the $\binom{4}{2}$ choices of which 2 kids to over-feed).

But now we have removed too much. We must add back in all the ways to give too many cookies to three kids. This uses 9 cookies, leaving only 1 to distribute to the 4 kids using stars and bars, which can be done in $\binom{4}{3}$ ways. We must consider this outcome for every possible choice of which three kids we over-feed, and there are $\binom{4}{3}$ ways of selecting that set of 3 kids.

Next we would subtract all the ways to give four kids too many cookies, but in this case, that number is 0.

All together we get that the number of ways to distribute 10 cookies to 4 kids without giving any kid more than 2 cookies is:

$$\binom{13}{3} - \left[\binom{4}{1}\binom{10}{3} - \binom{4}{2}\binom{7}{3} + \binom{4}{3}\binom{4}{3}\right]$$

which is

$$286 - [480 - 210 + 16] = 0.$$

This makes sense: there is NO way to distribute 10 cookies to 4 kids and make sure that nobody gets more than 2. It is slightly surprising that

$$\binom{13}{3} = \left[\binom{4}{1}\binom{10}{3} - \binom{4}{2}\binom{7}{3} + \binom{4}{3}\binom{4}{3}\right]$$

but since PIE works, this equality must hold.

Just so you don't think that these problems always have easier solutions, consider the following example.

Example 1.6.3
Earlier (Example 1.5.3) we counted the number of solutions to the equation

$$x_1 + x_2 + x_3 + x_4 + x_5 = 13$$

where $x_i \geq 0$ for each x_i.

How many of those solutions have $0 \leq x_i \leq 3$ for each x_i?

Solution. We must subtract off the number of solutions in which one or more of the variables has a value greater than 3. We will need to use PIE because counting the number of solutions for which each of the five variables separately are greater than 3 counts solutions multiple times. Here is what we get:

- Total solutions: $\binom{17}{4}$.
- Solutions where $x_1 > 3$: $\binom{13}{4}$. Give x_1 4 units first, then distribute the remaining 9 units to the 5 variables.
- Solutions where $x_1 > 3$ and $x_2 > 3$: $\binom{9}{4}$. After you give 4 units to x_1 and another 4 to x_2, you only have 5 units left to distribute.
- Solutions where $x_1 > 3$, $x_2 > 3$ and $x_3 > 3$: $\binom{5}{4}$.
- Solutions where $x_1 > 3$, $x_2 > 3$, $x_3 > 3$, and $x_4 > 3$: 0.

We also need to account for the fact that we could choose any of the five variables in the place of x_1 above (so there will be $\binom{5}{1}$ outcomes like this), any pair of variables in the place of x_1 and x_2 ($\binom{5}{2}$ outcomes) and so on. It is because of this that the double counting occurs, so we need to use PIE. All together we have that the number of solutions with $0 \leq x_i \leq 3$ is

$$\binom{17}{4} - \left[\binom{5}{1}\binom{13}{4} - \binom{5}{2}\binom{9}{4} + \binom{5}{3}\binom{5}{4}\right] = 15.$$

Counting Derangements

Investigate!

For your senior prank, you decide to switch the nameplates on your favorite 5 professors' doors. So that none of them feel left out, you want to make sure that all of the nameplates end up on the wrong door. How many ways can this be accomplished?

Attempt the above activity before proceeding

The advanced use of PIE has applications beyond stars and bars. A **derangement** of n elements $\{1, 2, 3, \ldots, n\}$ is a permutation in which no element is fixed. For example, there are 6 permutations of the three elements $\{1, 2, 3\}$:

$$123 \quad 132 \quad 213 \quad 231 \quad 312 \quad 321.$$

but most of these have one or more elements fixed: 123 has all three elements fixed since all three elements are in their original positions, 132 has the first element fixed (1 is in its original first position), and so on. In fact, the only derangements of three elements are

$$231 \text{ and } 312.$$

If we go up to 4 elements, there are 24 permutations (because we have 4 choices for the first element, 3 choices for the second, 2 choices for the third leaving only 1 choice for the last). How many of these are derangements? If you list out all 24 permutations and eliminate those which are not derangements, you will be left with just 9 derangements. Let's see how we can get that number using PIE.

Example 1.6.4
How many derangements are there of 4 elements?

Solution. We count all permutations, and subtract those which are not derangements. There are $4! = 24$ permutations of 4 elements. Now for a permutation to *not* be a derangement, at least one of the 4 elements must be fixed. There are $\binom{4}{1}$ choices for which single element we fix. Once fixed, we need to find a permutation of the other three elements. There are $3!$ permutations on 3 elements. But now we have counted too many non-derangements, so we must subtract those permutations which fix two elements. There are $\binom{4}{2}$ choices for which two elements we fix, and then for each pair, $2!$ permutations of the remaining elements. But this subtracts too many, so add back in permutations which fix 3 elements, all $\binom{4}{3}1!$ of them. Finally subtract the $\binom{4}{4}0!$ permutations (recall $0! = 1$) which fix all four elements. All together we get that the number of derangements of 4 elements is:

$$4! - \left[\binom{4}{1}3! - \binom{4}{2}2! + \binom{4}{3}1! - \binom{4}{4}0!\right] = 24 - 15 = 9.$$

Of course we can use a similar formula to count the derangements of any number of elements. However, the more elements we have, the longer the formula gets. Here is another example:

Example 1.6.5
Five gentlemen attend a party, leaving their hats at the door. At the end of the party, they hastily grab hats on their way out. How many different ways could this happen so that none of the gentlemen leave with their own hat?

Solution. We are counting derangements on 5 elements. There are $5!$ ways for the gentlemen to grab hats in any order—but many of these permutations will result in someone getting their own hat. So we subtract all the ways in which one or more of the men get their own hat. In other words,

we subtract the non-derangements. Doing so requires PIE. Thus the answer is:

$$5! - \left[\binom{5}{1}4! - \binom{5}{2}3! + \binom{5}{3}2! - \binom{5}{4}1! + \binom{5}{5}0!\right].$$

Counting Functions

Investigate!

- Consider all functions $f : \{1,2,3,4,5\} \to \{1,2,3,4,5\}$. How many functions are there all together? How many of those are injective? Remember, a function is an injection if every input goes to a different output.
- Consider all functions $f : \{1,2,3,4,5\} \to \{1,2,3,4,5\}$. How many of the *injections* have the property that $f(x) \neq x$ for any $x \in \{1,2,3,4,5\}$?

  Your friend claims that the answer is:

  $$5! - \left[\binom{5}{1}4! - \binom{5}{2}3! + \binom{5}{3}2! - \binom{5}{4}1! + \binom{5}{5}0!\right].$$

  Explain why this is correct.
- Recall that a *surjection* is a function for which every element of the codomain is in the range. How many of the functions $f : \{1,2,3,4,5\} \to \{1,2,3,4,5\}$ are surjective? Use PIE!

Attempt the above activity before proceeding

We have seen throughout this chapter that many counting questions can be rephrased as questions about counting functions with certain properties. This is reasonable since many counting questions can be thought of as counting the number of ways to assign elements from one set to elements of another.

Example 1.6.6
You decide to give away your video game collection so to better spend your time studying advance mathematics. How many ways can you do this, provided:

1. You want to distribute your 3 different PS4 games among 5 friends, so that no friend gets more than one game?
2. You want to distribute your 8 different 3DS games among 5 friends?
3. You want to distribute your 8 different SNES games among 5 friends, so that each friend gets at least one game?

In each case, model the counting question as a function counting question.

Solution.

1. We must use the three games (call them 1, 2, 3) as the domain and the 5 friends (a,b,c,d,e) as the codomain (otherwise the function would not be defined for the whole domain when a friend didn't get any game). So how many functions are there with domain $\{1,2,3\}$ and codomain $\{a,b,c,d,e\}$? The answer to this is $5^3 = 125$, since we can assign any of 5 elements to be the image of 1, any of 5 elements to be the image of 2 and any of 5 elements to be the image of 3.

   But this is not the correct answer to our counting problem, because one of these functions is $f = \begin{pmatrix} 1 & 2 & 3 \\ a & a & a \end{pmatrix}$; one friend can get more than one game. What we really need to do is count

injective functions. This gives $P(5,3) = 60$ functions, which is the answer to our counting question.

2. Again, we need to use the 8 games as the domain and the 5 friends as the codomain. We are counting all functions, so the number of ways to distribute the games is 5^8.

3. This question is harder. Use the games as the domain and friends as the codomain (otherwise an element of the domain would have more than one image, which is impossible). To ensure that every friend gets at least one game means that every element of the codomain is in the range. In other words, we are looking for *surjective* functions. How do you count those??

In Example 1.1.5 we saw how to count all functions (using the multiplicative principle) and in Example 1.3.4 we learned how to count injective functions (using permutations). Surjective functions are not as easily counted (unless the size of the domain is smaller than the codomain, in which case there are none).

The idea is to count the functions which are *not* surjective, and then subtract that from the total number of functions. This works very well when the codomain has two elements in it:

Example 1.6.7
How many functions $f : \{1,2,3,4,5\} \to \{a,b\}$ are surjective?

Solution. There are 2^5 functions all together, two choices for where to send each of the 5 elements of the domain. Now of these, the functions which are *not* surjective must exclude one or more elements of the codomain from the range. So first, consider functions for which a is not in the range. This can only happen one way: everything gets sent to b. Alternatively, we could exclude b from the range. Then everything gets sent to a, so there is only one function like this. These are the only ways in which a function could not be surjective (no function excludes both a and b from the range) so there are exactly $2^5 - 2$ surjective functions.

When there are three elements in the codomain, there are now three choices for a single element to exclude from the range. Additionally, we could pick pairs of two elements to exclude from the range, and we must make sure we don't over count these. It's PIE time!

Example 1.6.8
How many functions $f : \{1,2,3,4,5\} \to \{a,b,c\}$ are surjective?

Solution. Again start with the total number of functions: 3^5 (as each of the five elements of the domain can go to any of three elements of the codomain). Now we count the functions which are *not* surjective.

Start by excluding a from the range. Then we have two choices (b or c) for where to send each of the five elements of the domain. Thus there are 2^5 functions which exclude a from the range. Similarly, there are 2^5 functions which exclude b, and another 2^5 which exclude c. Now have we counted all functions which are not surjective? Yes, but in fact, we have counted some multiple times. For example, the function which sends everything to c was one of the 2^5 functions we counted when we excluded a from the range, and also one of the 2^5 functions we counted when we excluded b from the range. We must subtract out all the functions which specifically exclude two elements from the range. There is 1 function when we exclude a and b (everything goes to c), one function when we exclude a and c, and one function when we exclude b and c.

We are using PIE: to count the functions which are not surjective, we added up the functions which exclude a, b, and c separately, then subtracted the functions which exclude pairs of elements. We would then add back in the functions which exclude groups of three elements, except that there are no such functions. We find that the number of functions which are *not* surjective is

$$2^5 + 2^5 + 2^5 - 1 - 1 - 1 + 0.$$

Perhaps a more descriptive way to write this is

$$\binom{3}{1}2^5 - \binom{3}{2}1^5 + \binom{3}{3}0^5.$$

since each of the 2^5's was the result of choosing 1 of the 3 elements of the codomain to exclude from the range, each of the three 1^5's was the result of choosing 2 of the 3 elements of the codomain to exclude. Writing 1^5 instead of 1 makes sense too: we have 1 choice of were to send each of the 5 elements of the domain.

Now we can finally count the number of surjective functions:

$$3^5 - \left[\binom{3}{1}2^5 - \binom{3}{2}1^5\right] = 150.$$

You might worry that to count surjective functions when the codomain is larger than 3 elements would be too tedious. We need to use PIE but with more than 3 sets the formula for PIE is very long. However, we have lucked out. As we saw in the example above, the number of functions which exclude a single element from the range is the same no matter which single element is excluded. Similarly, the number of functions which exclude a pair of elements will be the same for every pair. With larger codomains, we will see the same behavior with groups of 3, 4, and more elements excluded. So instead of adding/subtracting each of these, we can simply add or subtract all of them at once, if you know how many there are. This works just like it did in for the other types of counting questions in this section, only now the size of the various combinations of sets is a number raised to a power, as opposed to a binomial coefficient or factorial. Here's what happens with 4 and 5 elements in the codomain.

Example 1.6.9

1. How many functions $f : \{1, 2, 3, 4, 5\} \to \{a, b, c, d\}$ are surjective?
2. How many functions $f : \{1, 2, 3, 4, 5\} \to \{a, b, c, d, e\}$ are surjective?

Solution.

1. There are 4^5 functions all together; we will subtract the functions which are not surjective. We could exclude any one of the four elements of the codomain, and doing so will leave us with 3^5 functions for each excluded element. This counts too many so we subtract the functions which exclude two of the four elements of the codomain, each pair giving 2^5 functions. But this excludes too many, so we add back in the functions which exclude three of the four elements of the codomain, each triple giving 1^5 function. There are $\binom{4}{1}$ groups of functions excluding a single element, $\binom{4}{2}$ groups of functions excluding a pair of elements, and $\binom{4}{3}$ groups of functions excluding a triple of elements. This means that the number of functions which are *not* surjective is:

$$\binom{4}{1}3^5 - \binom{4}{2}2^5 + \binom{4}{3}1^5.$$

   We can now say that the number of functions which are surjective is:

$$4^5 - \left[\binom{4}{1}3^5 - \binom{4}{2}2^5 + \binom{4}{3}1^5\right].$$

2. The number of surjective functions is:

$$5^5 - \left[\binom{5}{1}4^5 - \binom{5}{2}3^5 + \binom{5}{3}2^5 - \binom{5}{4}1^5\right].$$

We took the total number of functions 5^5 and subtracted all that were not surjective. There were $\binom{5}{1}$ ways to select a single element from the codomain to exclude from the range, and for each there were 4^5 functions. But this double counts, so we use PIE and subtract functions excluding two elements from the range: there are $\binom{5}{2}$ choices for the two elements to exclude, and for each pair, 3^5 functions. This takes out too many functions, so we add back in functions which exclude 3 elements from the range: $\binom{5}{3}$ choices for which three to exclude, and then 2^5 functions for each choice of elements. Finally we take back out the 1 function which excludes 4 elements for each of the $\binom{5}{4}$ choices of 4 elements.

If you happen to calculate this number precisely, you will get 120 surjections. That happens to also be the value of 5!. This might seem like an amazing coincidence until you realize that every surjective function $f : X \to Y$ with $|X| = |Y|$ finite must necessarily be a bijection. The number of bijections is always $|X|!$ in this case. What we have here is a *combinatorial proof* of the following identity:

$$n^n - \langle[\binom{n}{1}(n-1)^n - \binom{n}{2}(n-2)^n + \cdots + \binom{n}{n-1}1^n\rangle] = n!.$$

We have seen that counting surjective functions is another nice example of the advanced use of the Principle of Inclusion/Exclusion. Also, counting injective functions turns out to be equivalent to permutations, and counting all functions has a solution akin to those counting problems where order matters but repeats are allowed (like counting the number of words you can make from a given set of letters).

These are not just a few more examples of the techniques we have developed in this chapter. Quite the opposite: everything we have learned in this chapter are examples of *counting functions*!

Example 1.6.10
How many 5-letter words can you make using the eight letters a through h? How many contain no repeated letters?

Solution. By now it should be no surprise that there are 8^5 words, and $P(8,5)$ words without repeated letters. The new piece here is that we are actually counting functions. For the first problem, we are counting all functions from $\{1,2,\ldots,5\}$ to $\{a,b,\ldots,h\}$. The numbers in the domain represent the *position* of the letter in the word, the codomain represents the letter that could be assigned to that position. If we ask for no repeated letters, we are asking for injective functions.

If A and B are *any* sets with $|A| = 5$ and $|B| = 8$, then the number of functions $f : A \to B$ is 8^5 and the number of injections is $P(8,5)$. So if you can represent your counting problem as a function counting problem, most of the work is done.

Example 1.6.11
How many subsets are there of $\{1,2,\ldots,9\}$? How many 9-bit strings are there (of any weight)?

Solution. We saw in Section 1.2 that the answer to both these questions is 2^9, as we can say yes or no (or 0 or 1) to each of the 9 elements in the set (positions in the bit-string). But 2^9 also looks like the answer you get from counting functions. In fact, if you count all functions $f : A \to B$ with $|A| = 9$ and $|B| = 2$, this is exactly what you get.

This makes sense! Let $A = \{1,2,\ldots,9\}$ and $B = \{y,n\}$. We are assigning each element of the set either a yes or a no. Or in the language of bit-strings, we would take the 9 positions in the bit string as our domain and the set $\{0,1\}$ as the codomain.

So far we have not used a function as a model for binomial coefficients (combinations). Think for a moment about the relationship between combinations and permutations, say specifically $\binom{9}{3}$ and $P(9,3)$. We *do* have a function model for $P(9,3)$. This is the number of *injective* functions from a set of size 3 (say $\{1,2,3\}$ to a set of size 9 (say $\{1,2,\ldots,9\}$) since there are 9 choices for where to send the first element of the domain, then only 8 choices for the second, and 7 choices for the third. For example, the function might look like this:

$$f(1) = 5 \qquad f(2) = 8 \qquad f(3) = 4.$$

This is a different function from:

$$f(1) = 4 \qquad f(2) = 5 \qquad f(3) = 8.$$

Now $P(9,3)$ counts these as different outcomes correctly, but $\binom{9}{3}$ will count these (among others) as just one outcome. In fact, in terms of functions $\binom{9}{3}$ just counts the number of different ranges possible of injective functions. This should not be a surprise since binomial coefficients counts subsets, and the range is a possible subset of the codomain.[4]

While it is possible to interpret combinations as functions, perhaps the better advice is to instead use combinations (or stars and bars) when functions are not quite the right way to interpret the counting question.

Exercises

1. The dollar menu at your favorite tax-free fast food restaurant has 7 items. You have \$10 to spend. How many different meals can you buy if you spend all your money and:

(a) Purchase at least one of each item.

(b) Possibly skip some items.

(c) Don't get more than 2 of any particular item.

2. After a late night of math studying, you and your friends decide to go to your favorite tax-free fast food Mexican restaurant, *Burrito Chime*. You decide to order off of the dollar menu, which has 7 items. Your group has \$16 to spend (and will spend all of it).

(a) How many different orders are possible? Explain. (The *order* in which the order is placed does not matter - just which and how many of each item that is ordered.)

(b) How many different orders are possible if you want to get at least one of each item? Explain.

(c) How many different orders are possible if you don't get more than 4 of any one item? Explain.

3. After another gym class you are tasked with putting the 14 identical dodgeballs away into 5 bins. This time, no bin can hold more than 6 balls. How many ways can you clean up?

4. Consider the equation $x_1 + x_2 + x_3 + x_4 = 15$. How many solutions are there with $2 \leq x_i \leq 5$ for all $i \in \{1,2,3,4\}$?

5. Suppose you planned on giving 7 gold stars to some of the 13 star students in your class. Each student can receive at most one star. How many ways can you do this? Use PIE, and also an easier method, and compare your results.

6. Based on the previous question, give a combinatorial proof for the identity:

$$\binom{n}{k} = \binom{n+k-1}{k} - \sum_{j=1}^{n} (-1)^{j+1} \binom{n}{j} \binom{n+k-(2j+1)}{k}.$$

7. Illustrate how the counting of derangements works by writing all permutations of $\{1,2,3,4\}$ and the crossing out those which are not derangements. Keep track of the permutations you cross out more than once, using PIE.

8. How many permutations of $\{1,2,3,4,5\}$ leave exactly 1 element fixed?

9. Ten ladies of a certain age drop off their red hats at the hat check of a museum. As they are leaving, the hat check attendant gives the hats back randomly. In how many ways can exactly six of the ladies receive their own hat (and the other four not)? Explain.

[4] A more mathematically sophisticated interpretation of combinations is that we are defining two injective functions to be *equivalent* if they have the same range, and then counting the number of equivalence classes under this notion of equivalence.

10. The Grinch sneaks into a room with 6 Christmas presents to 6 different people. He proceeds to switch the name-labels on the presents. How many ways could he do this if:

(a) No present is allowed to end up with its original label? Explain what each term in your answer represents.

(b) Exactly 2 presents keep their original labels? Explain.

(c) Exactly 5 presents keep their original labels? Explain.

11. Consider functions $f : \{1,2,3,4\} \to \{a,b,c,d,e,f\}$. How many functions have the property that $f(1) \neq a$ or $f(2) \neq b$, or both?

12. Consider sets A and B with $|A| = 10$ and $|B| = 5$. How many functions $f : A \to B$ are surjective?

13. Let $A = \{1,2,3,4,5\}$. How many injective functions $f : A \to A$ have the property that for each $x \in A$, $f(x) \neq x$?

14. Let d_n be the number of derangements of n objects. For example, using the techniques of this section, we find

$$d_3 = 3! - \left(\binom{3}{1}2! - \binom{3}{2}1! + \binom{3}{3}0!\right)$$

We can use the formula for $\binom{n}{k}$ to write this all in terms of factorials. After simplifying, for d_3 we would get

$$d_3 = 3!\left(1 - \frac{1}{1} + \frac{1}{2} - \frac{1}{6}\right)$$

Generalize this to find a nicer formula for d_n. Bonus: For large n, approximately what fraction of all permutations are derangements? Use your knowledge of Taylor series from calculus.

1.7 Chapter Summary

Investigate!

Suppose you have a huge box of animal crackers containing plenty of each of 10 different animals. For the counting questions below, carefully examine their similarities and differences, and then give an answer. The answers are all one of the following:

$$P(10,6) \qquad \binom{10}{6} \qquad 10^6 \qquad \binom{15}{9}.$$

1. How many animal parades containing 6 crackers can you line up?
2. How many animal parades of 6 crackers can you line up so that the animals appear in alphabetical order?
3. How many ways could you line up 6 different animals in alphabetical order?
4. How many ways could you line up 6 different animals if they can come in any order?
5. How many ways could you give 6 children one animal cracker each?
6. How many ways could you give 6 children one animal cracker each so that no two kids get the same animal?
7. How many ways could you give out 6 giraffes to 10 kids?
8. Write a question about giving animal crackers to kids that has the answer $\binom{10}{6}$.

Attempt the above activity before proceeding

With all the different counting techniques we have mastered in this last chapter, it might be difficult to know when to apply which technique. Indeed, it is very easy to get mixed up and use the wrong counting method for a given problem. You get better with practice. As you practice you start to notice some trends that can help you distinguish between types of counting problems. Here are some suggestions that you might find helpful when deciding how to tackle a counting problem and checking whether your solution is correct.

- Remember that you are counting the number of items in some *list of outcomes*. Write down part of this list. Write down an element in the middle of the list – how are you deciding whether your element really is in the list. Could you get this element more than once using your proposed answer?
- If generating an element on the list involves selecting something (for example, picking a letter or picking a position to put a letter, etc), can the things you select be repeated? Remember, permutations and combinations select objects from a set *without* repeats.
- Does order matter? Be careful here and be sure you know what your answer really means. We usually say that order matters when you get different outcomes when the same objects are selected in different orders. Combinations and "Stars & Bars" are used when order *does not* matter.
- There are four possibilities when it comes to order and repeats. If order matters and repeats are allowed, the answer will look like n^k. If order matters and repeats are not allowed, we have $P(n,k)$. If order doesn't matter and repeats are allowed, use stars and bars. If order doesn't matter and repeats are not allowed, use $\binom{n}{k}$. But be careful: this only applies when you are selecting things, and you should make sure you know exactly what you are selecting before determining which case you are in.
- Think about how you would represent your counting problem in terms of sets or functions. We know how to count different sorts of sets and different types of functions.
- As we saw with combinatorial proofs, you can often solve a counting problem in more than one way. Do that, and compare your numerical answers. If they don't match, something is amiss.

While we have covered many counting techniques, we have really only scratched the surface of the large subject of *enumerative combinatorics*. There are mathematicians doing original research in this area even as you read this. Counting can be really hard.

In the next chapter, we will approach counting questions from a very different direction, and in doing so, answer infinitely many counting questions at the same time. We will create *sequences* of answers to related questions.

Chapter Review

1. You have 9 presents to give to your 4 kids. How many ways can this be done if:

(a) The presents are identical, and each kid gets at least one present?

(b) The presents are identical, and some kids might get no presents?

(c) The presents are unique, and some kids might get no presents?

(d) The presents are unique and each kid gets at least one present?

2. For each of the following counting problems, say whether the answer is $\binom{10}{4}$, $P(10,4)$, or neither. If you answer is "neither," say what the answer should be instead.

(a) How many shortest lattice paths are there from $(0,0)$ to $(10,4)$?

(b) If you have 10 bow ties, and you want to select 4 of them for next week, how many choices do you have?

(c) Suppose you have 10 bow ties and you will wear one on each of the next 4 days. How many choices do you have?

(d) If you want to wear 4 of your 10 bow ties next week (Monday through Sunday), how many ways can this be accomplished?

(e) Out of a group of 10 classmates, how many ways can you rank your top 4 friends?

(f) If 10 students come to their professor's office but only 4 can fit at a time, how different combinations of 4 students can see the prof first?

(g) How many 4 letter words can be made from the first 10 letters of the alphabet?

(h) How many ways can you make the word "cake" from the first 10 letters of the alphabet?

(i) How many ways are there to distribute 10 apples among 4 children?

(j) If you have 10 kids (and live in a shoe) and 4 types of cereal, how many ways can your kids eat breakfast?

(k) How many ways can you arrange exactly 4 ones in a string of 10 binary digits?

(l) You want to select 4 single digit numbers as your lotto picks. How many choices do you have?

(m) 10 kids want ice-cream. You have 4 varieties. How many ways are there to give the kids as much ice-cream as they want?

(n) How many 1-1 functions are there from $\{1, 2, \ldots, 10\}$ to $\{a, b, c, d\}$?

(o) How many surjective functions are there from $\{1, 2, \ldots, 10\}$ to $\{a, b, c, d\}$?

(p) Each of your 10 bow ties match 4 pairs of suspenders. How many outfits can you make?

(q) After the party, the 10 kids each choose one of 4 party-favors. How many outcomes?

(r) How many 6-elements subsets are there of the set $\{1, 2, \ldots, 10\}$

(s) How many ways can you split up 11 kids into 5 teams?

(t) How many solutions are there to $x_1 + x_2 + \cdots + x_5 = 6$ where each x_i is non-negative?

(u) Your band goes on tour. There are 10 cities within driving distance, but only enough time to play 4 of them. How many choices do you have for the cities on your tour?

(v) In how many different ways can you play the 4 cities you choose?

(w) Out of the 10 breakfast cereals available, you want to have 4 bowls. How many ways can you do this?

(x) There are 10 types of cookies available. You want to make a 4 cookie stack. How many different stacks can you make?

(y) From your home at (0,0) you want to go to either the donut shop at (5,4) or the one at (3,6). How many paths could you take?

(z) How many 10-digit numbers do not contain a sub-string of 4 repeated digits?

3. Recall, you own 3 regular ties and 5 bow ties. You realize that it would be okay to wear more than two ties to your clown college interview.

(a) You must select some of your ties to wear. Everything is okay, from no ties up to all ties. How many choices do you have?

(b) If you want to wear at least one regular tie and one bow tie, but are willing to wear up to all your ties, how many choices do you have for which ties to wear?

(c) How many choices do you have if you wear exactly 2 of the 3 regular ties and 3 of the 5 bow ties?

(d) Once you have selected 2 regular and 3 bow ties, in how many orders could you put the ties on, assuming you must have one of the three bow ties on top?

4. Give a counting question where the answer is $8 \cdot 3 \cdot 3 \cdot 5$. Give another question where the answer is $8 + 3 + 3 + 5$.

5. Consider five digit numbers $\alpha = a_1a_2a_3a_4a_5$, with each digit from the set $\{1, 2, 3, 4\}$.

(a) How many such numbers are there?

(b) How many such numbers are there for which the *sum* of the digits is even?

(c) How many such numbers contain more even digits than odd digits?

6. In a recent small survey of airline passengers, 25 said they had flown American in the last year, 30 had flown Jet Blue, and 20 had flown Continental. Of those, 10 reported they had flown on American and Jet Blue, 12 had flown on Jet Blue and Continental, and 7 had flown on American and Continental. 5 passengers had flown on all three airlines.
How many passengers were surveyed? (Assume the results above make up the entire survey.)

7. Recall, by 8-bit strings, we mean strings of binary digits, of length 8.

(a) How many 8-bit strings are there total?

(b) How many 8-bit strings have weight 5?

(c) How many subsets of the set $\{a, b, c, d, e, f, g, h\}$ contain exactly 5 elements?

(d) Explain why your answers to parts (b) and (c) are the same. Why are these questions equivalent?

8. What is the coefficient of x^{10} in the expansion of $(x+1)^{13} + x^2(x+1)^{17}$?

9. How many 8-letter words contain exactly 5 vowels (a,e,i,o,u)? What if repeated letters were not allowed?

10. For each of the following, find the number of shortest lattice paths from $(0, 0)$ to $(8, 8)$ which:

(a) pass through the point $(2, 3)$.

(b) avoid (do not pass through) the point $(7, 5)$.

(c) either pass through $(2, 3)$ or $(5, 7)$ (or both).

11. You live in Grid-Town on the corner of 2nd and 3rd, and work in a building on the corner of 10th and 13th. How many routes are there which take you from home to work and then back home, but by a different route?

12. How many 10-bit strings start with 111 or end with 101 or both?

13. How many 10-bit strings of weight 6 start with 111 or end with 101 or both?

14. How many 6 letter words made from the letters a, b, c, d, e, f without repeats do not contain the subword "bad" in (a) consecutive letters? or (b) not-necessarily consecutive letters (but in order)?

15. Explain using lattice paths why $\sum_{k=0}^{n} \binom{n}{k} = 2^n$.

16. Suppose you have 20 one-dollar bills to give out as prizes to your top 5 discrete math students. How many ways can you do this if:

(a) Each of the 5 students gets at least 1 dollar?

(b) Some students might get nothing?

(c) Each student gets at least 1 dollar but no more than 7 dollars?

Hint. Stars and bars.

17. How many functions $f : \{1, 2, 3, 4, 5\} \to \{a, b, c, d, e\}$ are there satisfying:

(a) $f(1) = a$ or $f(2) = b$ (or both)?

(b) $f(1) \neq a$ or $f(2) \neq b$ (or both)?

(c) $f(1) \neq a$ *and* $f(2) \neq b$, and f is injective?

(d) f is surjective, but $f(1) \neq a$, $f(2) \neq b$, $f(3) \neq c$, $f(4) \neq d$ and $f(5) \neq e$?

18. How many functions map $\{1, 2, 3, 4, 5, 6\}$ *onto* $\{a, b, c, d\}$ (i.e., how many *surjections* are there)?

19. To thank your math professor for doing such an amazing job all semester, you decide to bake Oscar cookies. You know how to make 10 different types of cookies.

(a) If you want to give your professor 4 different types of cookies, how many different combinations of cookie type can you select? Explain your answer.

(b) To keep things interesting, you decide to make a different number of each type of cookie. If again you want to select 4 cookie types, how many ways can you select the cookie types and decide for which there will be the most, second most, etc. Explain your answer.

(c) You change your mind again. This time you decide you will make a total of 12 cookies. Each cookie could be any one of the 10 types of cookies you know how to bake (and it's okay if you leave some types out). How many choices do you have? Explain.

(d) You realize that the previous plan did not account for presentation. This time, you once again want to make 12 cookies, each one could be any one of the 10 types of cookies. However, now you plan to shape the cookies into the numerals $1, 2, \ldots, 12$ (and probably arrange them to make a giant clock, but you haven't decided on that yet). How many choices do you have for which types of cookies to bake into which numerals? Explain.

(e) The only flaw with the last plan is that your professor might not get to sample all 10 different varieties of cookies. How many choices do you have for which types of cookies to make into which numerals, given that each type of cookie should be present at least once? Explain.

20. For which of the parts of the previous problem (Exercise 1.7.19) does it make sense to interpret the counting question as counting some number of functions? Say what the domain and codomain should be, and whether you are counting all functions, injections, surjections, or something else.

Chapter 7

DMOI 2: Sequences

Chapter 2

Sequences

Investigate!

There is a monastery in Hanoi, as the legend goes, with a great hall containing three tall pillars. Resting on the first pillar are 64 giant disks (or washers), all different sizes, stacked from largest to smallest. The monks are charged with the following task: they must move the entire stack of disks to the third pillar. However, due to the size of the disks, the monks cannot move more than one at a time. Each disk must be placed on one of the pillars before the next disk is moved. And because the disks are so heavy and fragile, the monks may never place a larger disk on top of a smaller disk. When the monks finally complete their task, the world shall come to an end. Your task: figure out how long before we need to start worrying about the end of the world.

1. First, let's find the minimum number of moves required for a smaller number of disks. Collect some data. Make a table.
2. Conjecture a formula for the minimum number of moves required to move n disks. Test your conjecture. How do you know your formula is correct?
3. If the monks were able to move one disk every second without ever stopping, how long before the world ends?

Attempt the above activity before proceeding

This puzzle is called the *Tower of Hanoi*. You are tasked with finding the minimum number of moves to complete the puzzle. This certainly sounds like a counting problem. Perhaps you have an answer? If not, what else could we try?

The answer depends on the number of disks you need to move. In fact, we could answer the puzzle first for 1 disk, then 2, then 3 and so on. If we list out all of the answers for each number of disks, we will get a **sequence** of numbers. The nth term in the sequence is the answer to the question, "what is the smallest number of moves required to complete the Tower of Hanoi puzzle with n disks?" You might wonder why we would create such a sequence instead of just answering the question. By looking at how the sequence of numbers grows, we gain insight into the problem. It is easy to count the number of moves required for a small number of disks. We can then look for a pattern among the first few terms of the sequence. Hopefully this will suggest a method for finding the nth term of the sequence, which is the answer to our question. Of course we will also need to verify that our suspected pattern is correct, and that this correct pattern really does give us the nth term we think it does, but it is impossible to prove that your formula is correct without having a formula to start with.

Sequences are also interesting mathematical objects to study in their own right. Let's see why.

2.1 Definitions

Investigate!

What comes next:

$$1,\ 11,\ 21,\ 1211,\ 111221,\ 312211,\ \ldots$$

Attempt the above activity before proceeding

A **sequence** is simply an ordered list of numbers. For example, here is a sequence: 0, 1, 2, 3, 4, 5, ... This is different from the set $\mathbb{N}$ because, while the sequence is a complete list of every element in the set of natural numbers, in the sequence we very much care what order the numbers come in. For this reason, when we use variables to represent terms in a sequence they will look like this:

$$a_0, a_1, a_2, a_3, \ldots$$

To refer to the *entire* sequence at once, we will write $(a_n)_{n\in\mathbb{N}}$ or $(a_n)_{n\geq 0}$, or sometimes if we are being sloppy, just (a_n) (in which case we assume we start the sequence with a_0).

We might replace the a with another letter, and sometimes we omit a_0, starting with a_1, in which case we would use $(a_n)_{n\geq 1}$ to refer to the sequence as a whole. The numbers in the subscripts are called **indices** (the plural of **index**).

While we often just think of sequences as an ordered list of numbers, they really are a type of function. Specifically, the sequence $(a_n)_{n\geq 0}$ is a function with domain $\mathbb{N}$ where a_n is the image of the natural number n. Later we will manipulate sequences in much the same way you have manipulated functions in algebra or calculus. We can shift a sequence up or down, add two sequences, or ask for the rate of change of a sequence. These are done exactly as you would for functions.

That said, while keeping the rigorous mathematical definition in mind is helpful, we often describe sequences by writing out the first few terms.

Example 2.1.1
Can you find the next term in the following sequences?

1. $7, 7, 7, 7, 7, \ldots$
2. $3, -3, 3, -3, 3, \ldots$
3. $1, 5, 2, 10, 3, 15, \ldots$
4. $1, 2, 4, 8, 16, 32, \ldots$
5. $1, 4, 9, 16, 25, 36, \ldots$
6. $1, 2, 3, 5, 8, 13, 21, \ldots$
7. $1, 3, 6, 10, 15, 21, \ldots$
8. $2, 3, 5, 7, 11, 13, \ldots$
9. $3, 2, 1, 0, -1, \ldots$
10. $1, 1, 2, 6, \ldots$

Solution. No you cannot. You might guess that the next terms are:

1. 7
2. -3
3. 4
4. 64
5. 49
6. 34
7. 28
8. 17

9. -2
10. 24

In fact, those are the next terms of the sequences I had in mind when I made up the example, but there is no way to be sure they are correct.

Still, we will often do this. Given the first few terms of a sequence, we can ask what the pattern in the sequence suggests the next terms are.

Given that no number of initial terms in a sequence is enough to say for certain which sequence we are dealing with, we need to find another way to specify a sequence. We consider two ways to do this:

Closed formula

A **closed formula** for a sequence $(a_n)_{n\in\mathbb{N}}$ is a formula for a_n using a fixed finite number of operations on n. This is what you normally think of as a formula in n, just like if you were defining a function in terms of n (because that is exactly what you are doing).

Recursive definition

A **recursive definition** (sometimes called an **inductive definition**) for a sequence $(a_n)_{n\in\mathbb{N}}$ consists of a **recurrence relation**: an equation relating a term of the sequence to previous terms (terms with smaller index) and an **initial condition**: a list of a few terms of the sequence (one less than the number of terms in the recurrence relation).

It is easier to understand what is going on here with an example:

Example 2.1.2
Here are a few closed formulas for sequences:

- $a_n = n^2$.
- $a_n = \dfrac{n(n+1)}{2}$.
- $a_n = \dfrac{\left(\frac{1+\sqrt{5}}{2}\right)^n - \left(\frac{1+\sqrt{5}}{2}\right)^{-n}}{5}$.

Note in each case, if you are given n, you can calculate a_n directly: just plug in n. For example, to find a_3 in the second sequence, just compute $a_3 = \frac{3(3+1)}{2} = 6$.

Here are a few recursive definitions for sequences:

- $a_n = 2a_{n-1}$ with $a_0 = 1$.
- $a_n = 2a_{n-1}$ with $a_0 = 27$.
- $a_n = a_{n-1} + a_{n-2}$ with $a_0 = 0$ and $a_1 = 1$.

In these cases, if you are given n, you cannot calculate a_n directly, you first need to find a_{n-1} (or a_{n-1} and a_{n-2}). In the second sequence, to find a_3 you would take $2a_2$, but to find $a_2 = 2a_1$ we would need to know $a_1 = 2a_0$. We do know this, so we could trace back through these equations to find $a_1 = 54$, $a_2 = 108$ and finally $a_3 = 216$.

Investigate!

You have a large collection of 1×1 squares and 1×2 dominoes. You want to arrange these to make a 1×15 strip. How many ways can you do this?

1. Start by collecting data. How many length 1×1 strips can you make? How many 1×2 strips? How many 1×3 strips? And so on.
2. How are the 1×3 and 1×4 strips related to the 1×5 strips?
3. How many 1×15 strips can you make?
4. What if I asked you to find the number of 1×1000 strips? Would the method you used to calculate the number fo 1×15 strips be helpful?

Attempt the above activity before proceeding

You might wonder why we would bother with recursive definitions for sequences. After all, it is harder to find a_n with a recursive definition than with a closed formula. This is true, but it is also harder to find a closed formula for a sequence than it is to find a recursive definition. So to find a useful closed formula, we might first find the recursive definition, then use that to find the closed formula.

This is not to say that recursive definitions aren't useful in finding a_n. You can always calculate a_n given a recursive definition, it might just take a while.

Example 2.1.3
Find a_6 in the sequence defined by $a_n = 2a_{n-1} - a_{n-2}$ with $a_0 = 3$ and $a_1 = 4$.

Solution. We know that $a_6 = 2a_5 - a_4$. So to find a_6 we need to find a_5 and a_4. Well

$$a_5 = 2a_4 - a_3 \qquad \text{and} \qquad a_4 = 2a_3 - a_2,$$

so if we can only find a_3 and a_2 we would be set. Of course

$$a_3 = 2a_2 - a_1 \qquad \text{and} \qquad a_2 = 2a_1 - a_0,$$

so we only need to find a_1 and a_0. But we are given these. Thus

$$\begin{aligned}
a_0 &= 3 \\
a_1 &= 4 \\
a_2 &= 2 \cdot 4 - 3 = 5 \\
a_3 &= 2 \cdot 5 - 4 = 6 \\
a_4 &= 2 \cdot 6 - 5 = 7 \\
a_5 &= 2 \cdot 7 - 6 = 8 \\
a_6 &= 2 \cdot 8 - 7 = 9.
\end{aligned}$$

Note that now we can guess a closed formula for the nth term of the sequence: $a_n = n + 3$. To be sure this will always work, we could plug in this formula into the recurrence relation:

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

That is not quite enough though, since there can be multiple closed formulas that satisfy the same recurrence relation; we must also check that our closed formula agrees on the initial terms of the sequence. Since $a_0 = 0 + 3 = 3$ and $a_1 = 1 + 3 = 4$ are the correct initial conditions, we can now conclude we have the correct closed formula.

Finding closed formulas, or even recursive definitions, for sequences is not trivial. There is no one method for doing this. Just like in evaluating integrals or solving differential equations, it is useful to have a bag of tricks you can apply, but sometimes there is no easy answer.

One useful method is to relate a given sequence to another sequence for which we already know the closed formula.

Example 2.1.4

Use the formulas $T_n = \frac{n(n+1)}{2}$ and $a_n = 2^n$ to find closed formulas for the following sequences.

1. (b_n): $1, 2, 4, 7, 11, 16, 22, \ldots$.
2. (c_n): $3, 5, 9, 17, 33, \ldots$.
3. (d_n): $0, 2, 6, 12, 20, 30, 42, \ldots$.
4. (e_n): $3, 6, 10, 15, 21, 28, \ldots$.
5. (f_n): $0, 1, 3, 7, 15, 31, \ldots$.
6. (g_n) $3, 6, 12, 24, 48, \ldots$.
7. (h_n): $6, 10, 18, 34, 66, \ldots$.
8. (j_n): $15, 33, 57, 87, 123, \ldots$.

Solution. Before you say this is impossible, what we are asking for is simply to find a closed formula which agrees with all of the initial terms of the sequences. Of course there is no way to read into the mind of the person who wrote the numbers down, but we can at least do this.

The first few terms of $(T_n)_{n\geq 0}$ are $0, 1, 3, 6, 10, 15, 21, \ldots$ (these are called the **triangular numbers**). The first few terms of $(a_n)_{n\geq 0}$ are $1, 2, 4, 8, 16, \ldots$. Let's try to find formulas for the given sequences:

1. $(1, 2, 4, 7, 11, 16, 22, \ldots)$. Note that if subtract 1 from each term, we get the sequence (T_n). So we have $b_n = T_n + 1$. Therefore a closed formula is $b_n = \frac{n(n+1)}{2} + 1$. A quick check of the first few n confirms we have it right.
2. $(3, 5, 9, 17, 33, \ldots)$. Each term in this sequence is one more than a power of 2, so we might guess the closed formula is $c_n = a_n + 1 = 2^n + 1$. If we try this though, we get $c_0 2^0 + 1 = 2$ and $c_1 = 2^1 + 1 = 3$. We are off because the indices are shifted. What we really want is $c_n = a_{n+1} + 1$ giving $c_n = 2^{n+1} + 1$.
3. $(0, 2, 6, 12, 20, 30, 42, \ldots)$. Notice that all these terms are even. What happens if we factor out a 2? We get (T_n)! More precisely, we find that $d_n/2 = T_n$, so this sequence has closed formula $d_n = n(n + 1)$.
4. $(3, 6, 10, 15, 21, 28, \ldots)$. These are all triangular numbers. However, we are starting with 3 as our initial term instead of as our third term. So if we could plug in 2 instead of 0 into the formula for T_n, we would be set. Therefore the closed formula is $e_n = \frac{(n+2)(n+3)}{2}$ (where $n + 3$ came from $(n + 2) + 1$). Thinking about sequences as functions, we are doing a horizontal shift by 2: $e_n = T_{n+2}$ which would cause the graph to shift 2 units to the left.
5. $(0, 1, 3, 7, 15, 31, \ldots)$. Try adding 1 to each term and we get powers of 2. You might guess this because each term is a little more than twice the previous term (the powers of 2 are *exactly* twice the previous term). Closed formula: $f_n = 2^n - 1$.

6. $(3, 6, 12, 24, 48, \ldots)$. These numbers are also doubling each time, but are also all multiples of 3. Dividing each by 3 gives $1, 2, 4, 8, \ldots$. Aha. We get the closed formula $g_n = 3 \cdot 2^n$.

7. $(6, 10, 18, 34, 66, \ldots)$. To get from one term to the next, we almost double each term. So maybe we can relate this back to 2^n. Yes, each term is 2 more than a power of 2. So we get $h_n = 2^{n+2} + 2$ (the $n + 2$ is because the first term is 2 more than 2^2, not 2^0). Alternatively, we could have related this sequence to the second sequence in this example: starting with $3, 5, 9, 17, \ldots$ we see that this sequence is twice the terms from that sequence. That sequence had closed formula $c_n = 2^{n+1} + 1$. Our sequence here would be twice this, so $h_n = 2(2^n + 1)$, which is the same as we got before.

8. $(15, 33, 57, 87, 123, \ldots)$. Try dividing each term by 3. That gives the sequence $5, 11, 19, 29, 41, \ldots$. Now add 1: $6, 12, 20, 30, 42, \ldots$, which is (d_n) in this example, except starting with 6 instead of 0. So let's start with the formula $d_n = n(n + 1)$. To start with the 6, we shift: $(n + 2)(n + 3)$. But this is one too many, so subtract 1: $(n + 2)(n + 3) - 1$. That gives us our sequence, but divided by 3. So we want $j_n = 3((n + 2)(n + 3) - 1)$.

Exercises

1. Find the closed formula for each of the following sequences by relating them to a well known sequence. Assume the first term given is a_1.

(a) $2, 5, 10, 17, 26, \ldots$

(b) $0, 2, 5, 9, 14, 20, \ldots$

(c) $8, 12, 17, 23, 30, \ldots$

(d) $1, 5, 23, 119, 719, \ldots$

2. For each sequence given below, find a closed formula for a_n, the nth term of the sequence (assume the first terms are a_0) by relating it to another sequence for which you already know the formula. In each case, briefly say how you got your answers.

(a) $4, 5, 7, 11, 19, 35, \ldots$

(b) $0, 3, 8, 15, 24, 35, \ldots$

(c) $6, 12, 20, 30, 42, \ldots$

(d) $0, 2, 7, 15, 26, 40, 57, \ldots$ (Cryptic Hint: these might be called "house numbers")

3. The Fibonacci sequence is $0, 1, 1, 2, 3, 5, 8, 13, \ldots$ (where $F_0 = 0$).

(a) Give the recursive definition for the sequence.

(b) Write out the first few terms of the sequence of partial sums: $0, 0 + 1, 0 + 1 + 1, \ldots$

(c) Give a closed formula for the sequence of partial sums in terms of F_n (for example, you might say $F_0 + F_1 + \cdots + F_n = 3F_{n-1}^2 + n$, although that is definitely not correct).

4. Consider the three sequences below. For each, find a recursive definition. How are these sequences related?

(a) $2, 4, 6, 10, 16, 26, 42, \ldots$.

(b) $5, 6, 11, 17, 28, 45, 73, \ldots$.

(c) $0, 0, 0, 0, 0, 0, 0, \ldots$.

5. Show that $a_n = 3 \cdot 2^n + 7 \cdot 5^n$ is a solution to the recurrence relation $a_n = 7a_{n-1} - 10a_{n-2}$. What would the initial conditions need to be for this to be the closed formula for the sequence?

6. Write out the first few terms of the sequence given by $a_1 = 3$; $a_n = 2a_{n-1} + 4$. Then find a recursive definition for the sequence $10, 24, 52, 108, \ldots$.

7. Write out the first few terms of the sequence given by $a_n = n^2 - 3n + 1$. Then find a closed formula for the sequence (starting with a_1) $0, 2, 6, 12, 20, \ldots$.

8. Find a closed formula for the sequence with recursive definition $a_n = 2a_{n-1} - a_{n-2}$ with $a_1 = 1$ and $a_2 = 2$.

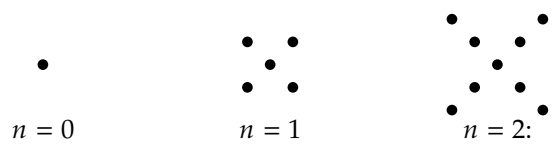
9. Find a recursive definition for the sequence with closed formula $a_n = 3 + 2n$. Bonus points if you can give a recursive definition in which makes use of two previous terms and no constants.

2.2 Arithmetic and Geometric Sequences

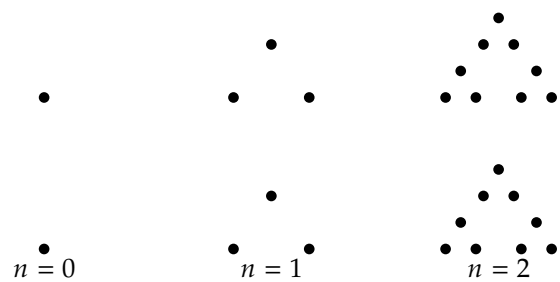
Investigate!

For the patterns of dots below, draw the next pattern in the sequence. Then give a recursive definition and a closed formula for the number of dots in the nth pattern.

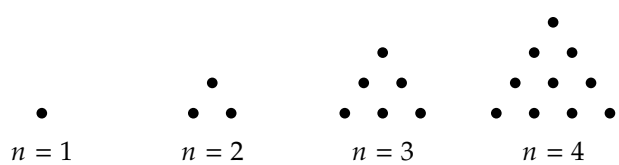
1.



2.



3.



Attempt the above activity before proceeding

We now turn to the question of finding closed formulas for particular types of sequences.

Arithmetic Sequences

If the terms of a sequence differ by a constant, we say the sequence is **arithmetic**. If the initial term (a_0) of the sequence is a and the **common difference** is d, then we have,

Recursive definition: $a_n = a_{n-1} + d$ with $a_0 = a$.

Closed formula: $a_n = a + dn$.

How do we know this? For the recursive definition, we need to specify a_0. Then we need to express a_n in terms of a_{n-1}. If we call the first term a, then $a_0 = a$. For the recurrence relation, by the definition of an arithmetic sequence, the difference between successive terms is some constant, say d. So $a_n - a_{n-1} = d$, or in other words,

$$a_0 = a \qquad a_n = a_{n-1} + d.$$

To find a closed formula, first write out the sequence in general:

$$\begin{aligned}
a_0 &= a \\
a_1 &= a_0 + d = a + d \\
a_2 &= a_1 + d = a + d + d = a + 2d \\
a_3 &= a_2 + d = a + 2d + d = a + 3d \\
&\vdots
\end{aligned}$$

We see that to find the nth term, we need to start with a and then add d a bunch of times. In fact, add it n times. Thus $a_n = a + dn$.

Example 2.2.1
Find recursive definitions and closed formulas for the sequences below. Assume the first term listed is a_0.

1. $2, 5, 8, 11, 14, \ldots$.
2. $50, 43, 36, 29, \ldots$.

Solution. First we should check that these sequences really are arithmetic by taking differences of successive terms. Doing so will reveal the common difference d.

1. $5-2 = 3$, $8-5 = 3$, etc. To get from each term to the next, we add three, so $d = 3$. The recursive definition is therefore $a_n = a_{n-1} + 3$ with $a_0 = 2$. The closed formula is $a_n = 2 + 3n$.
2. Here the common difference is -7, since we add -7 to 50 to get 43, and so on. Thus we have a recursive definition of $a_n = a_{n-1} - 7$ with $a_0 = 50$. The closed formula is $a_n = 50 - 7n$.

What about sequences like $2, 6, 18, 54, \ldots$? This is not arithmetic because the difference between terms is not constant. However, the *ratio* between successive terms is constant. We call such sequences geometric.

The recursive definition for the geometric sequence with initial term a and common ratio r is $a_n = a_n \cdot r; a_0 = a$. To get the next term we multiply the previous term by r. We can find the closed formula like we did for the arithmetic progression. Write

$$\begin{aligned}
a_0 &= a \\
a_1 &= a_0 \cdot r \\
a_2 &= a_1 \cdot r = a_0 \cdot r \cdot r = a_0 \cdot r^2 \\
&\vdots
\end{aligned}$$

We must multiply the first term a by r a number of times, n times to be precise. We get $a_n = a \cdot r^n$.

Geometric Sequences

A sequence is called **geometric** if the ratio between successive terms is constant. Suppose the initial term a_0 is a and the **common ratio** is r. Then we have,
Recursive definition: $a_n = ra_{n-1}$ with $a_0 = a$.
Closed formula: $a_n = a \cdot r^n$.

Example 2.2.2
Find the recursive and closed formula for the sequences below. Again, the first term listed is a_0.

1. $3, 6, 12, 24, 48, \ldots$
2. $27, 9, 3, 1, 1/3, \ldots$

Solution. Again, we should first check that these sequences really are geometric, this time by dividing each term by its previous term. Assuming this ratio is constant, we will have found r.

1. $6/3 = 2$, $12/6 = 2$, $24/12 = 2$, etc. Yes, to get from any term to the next, we multiply by $r = 2$. So the recursive definition is $a_n = 2a_{n-1}$ with $a_0 = 3$. The closed formula is $a_n = 3 \cdot 2^n$.
2. The common ratio is $r = 1/3$. So the sequence has recursive definition $a_n = \frac{1}{3}a_{n-1}$ with $a_0 = 27$ and closed formula $a_n = 27 \cdot \frac{1}{3}^n$.

In the examples and formulas above, we assumed that the *initial* term was a_0. If your sequence starts with a_1, you can easily find the term that would have been a_0 and use that in the formula. For example, if we want a formula for the sequence $2, 5, 8, \ldots$ and insist that $2 = a_1$, then we can find $a_0 = -1$ (since the sequence is arithmetic with common difference 3, we have $a_0 + 3 = a_1$). Then the closed formula will be $a_n = -1 + 3n$.

Remark 2.2.3. If you look at other textbooks or online, you might find that their closed formulas for arithmetic and geometric sequences differ from ours. Specifically, you might find the formulas $a_n = a + (n-1)d$ (arithmetic) and $a_n = a \cdot r^{n-1}$ (geometric). Which is correct? Both! In our case, we take a to be a_0. If instead we had a_1 as our initial term, we would get the (slightly more complicated) formulas you find elsewhere.

Sums of Arithmetic and Geometric Sequences

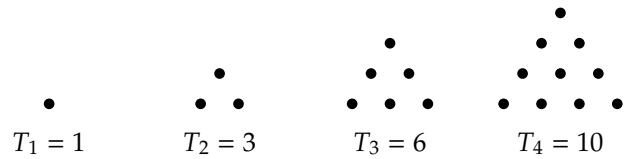
Investigate!

Your neighborhood grocery store has a candy machine full of Skittles.

1. Suppose that the candy machine currently holds exactly 650 Skittles, and every time someone inserts a quarter, exactly 7 Skittles come out of the machine.
   (a) How many Skittles will be left in the machine after 20 quarters have been inserted?
   (b) Will there ever be exactly zero Skittles left in the machine? Explain.
2. What if the candy machine gives 7 Skittles to the first customer who put in a quarter, 10 to the second, 13 to the third, 16 to the fourth, etc. How many Skittles has the machine given out after 20 quarters are put into the machine?
3. Now, what if the machine gives 4 Skittles to the first customer, 7 to the second, 12 to the third, 19 to the fourth, etc. How many Skittles has the machine given out after 20 quarters are put into the machine?

Attempt the above activity before proceeding

Look at the sequence $(T_n)_{n\geq 1}$ which starts $1, 3, 6, 10, 15, \ldots$. These are called the **triangular numbers** since they represent the number of dots in an equilateral triangle (think of how you arrange 10 bowling pins: a row of 4 plus a row of 3 plus a row of 2 and a row of 1).



Is this sequence arithmetic? No, since $3 - 1 = 2$ and $6 - 3 = 3 \neq 2$, so there is no common difference. Is the sequence geometric? No. $3/1 = 3$ but $6/3 = 2$, so there is no common ratio. What to do?

Notice that the differences between terms form an arithmetic sequence: $2, 3, 4, 5, 6, \ldots$. This says that the nth term of the sequence $1, 3, 6, 10, 15, \ldots$ is the *sum* of the first n terms in the sequence $1, 2, 3, 4, 5, \ldots$. We say that the first sequence is the **sequence of partial sums** of the second sequence (partial sums because we are not taking the sum of all infinitely many terms). If we know how to add up the terms of an arithmetic sequence, we could use this to find a closed formula for a sequence whose differences are the terms of that arithmetic sequence.

This should become clearer if we write the triangular numbers like this:

$$\begin{aligned} 1 &= 1 \\ 3 &= 1 + 2 \\ 6 &= 1 + 2 + 3 \\ 10 &= 1 + 2 + 3 + 4 \\ \vdots \quad & \quad \vdots \\ T_n &= 1 + 2 + 3 + \cdots + n. \end{aligned}$$

Consider how we could find the sum of the first 100 positive integers (that is, T_{100}). Instead of adding them in order, we regroup and add $1 + 100 = 101$. The next pair to combine is $2 + 99 = 101$. Then $3 + 98 = 101$. Keep going. This gives 50 pairs which each add up to 101, so $T_{100} = 101 \cdot 50 = 5050$.[1]

In general, using this same sort of regrouping, we find that $T_n = \frac{n(n+1)}{2}$. Incidentally, this is exactly the same as $\binom{n+1}{2}$, which makes sense if you think of the triangular numbers as counting the number of handshakes that take place at a party with $n + 1$ people: the first person shakes n hands, the next shakes an additional $n - 1$ hands and so on.

The point of all of this is that some sequences, while not arithmetic or geometric, can be interpreted as the sequence of partial sums of arithmetic and geometric sequences. Luckily there are methods we can use to compute these sums quickly.

Summing Arithmetic Sequences: Reverse and Add

Here is a technique that allows us to quickly find the sum of an arithmetic sequence.

Example 2.2.4
Find the sum: $2 + 5 + 8 + 11 + 14 + \cdots + 470$.

Solution. The idea is to mimic how we found the formula for triangular numbers. If we add the first and last terms, we get 472. The second term and second-to-last term also add up to 472. To keep track of everything, we might express this as follows. Call the sum S. Then,

$$\begin{array}{rcccccccccc} & S = & 2 & + & 5 & + & 8 & + \cdots + & 467 & + & 470 \\ + & S = & 470 & + & 467 & + & 464 & + \cdots + & 5 & + & 2 \\ \hline & 2S = & 472 & + & 472 & + & 472 & + \cdots + & 472 & + & 472 \end{array}$$

To find $2S$ then we add 472 to itself a number of times. What number? We need to decide how many terms (**summands**) are in the sum. Since the terms form an arithmetic sequence, the nth term

[1]This insight is usually attributed to Carl Friedrich Gauss, one of the greatest mathematicians of all time, who discovered it as a child when his unpleasant elementary teacher thought he would keep the class busy by requiring them to compute the lengthy sum.

in the sum (counting 2 as the 0th term) can be expressed as $2 + 3n$. If $2 + 3n = 470$ then $n = 156$. So n ranges from 0 to 156, giving 157 terms in the sum. This is the number of 472's in the sum for $2S$. Thus

$$2S = 157 \cdot 472 = 74104$$

It is now easy to find S:

$$S = 74104/2 = 37052$$

This will work for any sum of *arithmetic* sequences. Call the sum S. Reverse and add. This produces a single number added to itself many times. Find the number of times. Multiply. Divide by 2. Done.

Example 2.2.5
Find a closed formula for $6 + 10 + 14 + \cdots + (4n - 2)$.

Solution. Again, we have a sum of an arithmetic sequence. We need to know how many terms are in the sequence. Clearly each term in the sequence has the form $4k - 2$ (as evidenced by the last term). For which values of k though? To get 6, $k = 2$. To get $4n - 2$ take $k = n$. So to find the number of terms, we need to know how many integers are in the range $2, 3, \ldots, n$. The answer is $n - 1$. (There are n numbers from 1 to n, so one less if we start with 2.)

Now reverse and add:

$$\begin{array}{rrcccccccc} & S = & 6 & + & 10 & +\cdots+ & 4n-6 & + & 4n-2 \\ + & S = & 4n-2 & + & 4n-6 & +\cdots+ & 10 & + & 6 \\ \hline & 2S = & 4n+4 & + & 4n+4 & +\cdots+ & 4n+4 & + & 4n+4 \end{array}$$

Since there are $n - 2$ terms, we get

$$2S = (n-2)(4n+4) \qquad \text{so} \qquad S = \frac{(n-2)(4n+4)}{2}$$

Besides finding sums, we can use this technique to find closed formulas for sequences we recognize as sequences of partial sums.

Example 2.2.6
Use partial sums to find a closed formula for $(a_n)_{n\geq 0}$ which starts $2, 3, 7, 14, 24, 37, \ldots\ldots$.

Solution. First, if you look at the differences between terms, you get a sequence of differences: $1, 4, 7, 10, 13, \ldots$, which is an arithmetic sequence. Written another way:

$$\begin{aligned} a_0 &= 2 \\ a_1 &= 2 + 1 \\ a_2 &= 2 + 1 + 4 \\ a_3 &= 2 + 1 + 4 + 7 \end{aligned}$$

and so on. We can write the general term of (a_n) in terms of the arithmetic sequence as follows:

$$a_n = 2 + 1 + 4 + 7 + 10 + \cdots + (1 + 3(n-1))$$

(we use $1 + 3(n - 1)$ instead of $1 + 3n$ to get the indices to line up correctly; for a_3 we add up to 7, which is $1 + 3(3 - 1)$).

We can reverse and add, but the initial 2 does not fit our pattern. This just means we need to keep the 2 out of the reverse part:

$$\begin{array}{rcccccc} a_n = & 2 & + & 1 & + & 4 & +\cdots+ \quad 1+3(n-1) \\ +\,a_n = & 2 & + & 1+3(n-1) & + & 1+3(n-2) & +\cdots+ \quad 1 \\ \hline 2a_n = & 4 & + & 2+3(n-1) & + & 2+3(n-1) & +\cdots+ \quad 2+3(n-1) \end{array}$$

Not counting the first term (the 4) there are n summands of $2+3(n-1) = 3n-1$ so the right-hand side becomes $2+(3n-1)n$.

Finally, solving for a_n we get

$$a_n = \frac{4+(3n-1)n}{2}.$$

Just to be sure, we check $a_0 = \frac{4}{2} = 2$, $a_1 = \frac{4+2}{2} = 3$, etc. We have the correct closed formula.

Summing Geometric Sequences: Multiply, Shift and Subtract

To find the sum of a geometric sequence, we cannot just reverse and add. Do you see why? The reason we got the same term added to itself many times is because there was a constant difference. So as we added that difference in one direction, we subtracted the difference going the other way, leaving a constant total. For geometric sums, we have a different technique.

Example 2.2.7
What is $3+6+12+24+\cdots+12288$?

Solution. Multiply each term by 2, the common ratio. You get $2S = 6+12+24+\cdots+24576$. Now subtract: $2S - S = -3+24576 = 24573$. Since $2S - S = S$, we have our answer.

To better see what happened in the above example, try writing it this way:

$$\begin{array}{rcll} S = & 3+ & 6+12+24+\cdots+12288 & \\ -\,2S = & & 6+12+24+\cdots+12288 & +24576 \\ \hline -S = & 3+ & 0+0+0+\cdots+0 & -24576 \end{array}$$

Then divide both sides by -1 and we have the same result for S. The idea is, by multiplying the sum by the common ratio, each term becomes the next term. We shift over the sum to get the subtraction to mostly cancel out, leaving just the first term and new last term.

Example 2.2.8
Find a closed formula for $S(n) = 2+10+50+\cdots+2\cdot 5^n$.

Solution. The common ratio is 5. So we have

$$\begin{array}{rcl} S & = & 2+10+50+\cdots+2\cdot 5^n \\ -\;5S & = & \quad 10+50+\cdots+2\cdot 5^n+2\cdot 5^{n+1} \\ \hline -4S & = & 2-2\cdot 5^{n+1} \end{array}$$

Thus $S = \dfrac{2-2\cdot 5^{n+1}}{-4}$

Even though this might seem like a new technique, you have probably used it before.

Example 2.2.9
Express $0.464646\ldots$ as a fraction.

Solution. Let $N = 0.46464646\ldots$. Consider $0.01N$. We get:

$$\begin{array}{rrl} & N = & 0.4646464\ldots \\ - & 0.01N = & 0.00464646\ldots \\ \hline & 0.99N = & 0.46 \end{array}$$

So $N = \frac{46}{99}$. What have we done? We viewed the repeating decimal $0.464646\ldots$ as a sum of the geometric sequence $0.46, 0.0046, 0.000046, \ldots$ The common ratio is 0.01. The only real difference is that we are now computing an *infinite* geometric sum, we do not have the extra "last" term to consider. Really, this is the result of taking a limit as you would in calculus when you compute *infinite* geometric sums.

$\sum$ and $\prod$ notation

To simplify writing out sums, we will use notation like $\displaystyle\sum_{k=1}^{n} a_k$. This means add up the a_k's where k changes from 1 to n.

Example 2.2.10
Use $\sum$ notation to rewrite the sums:

1. $1 + 2 + 3 + 4 + \cdots + 100$
2. $1 + 2 + 4 + 8 + \cdots + 2^{50}$
3. $6 + 10 + 14 + \cdots + (4n - 2)$.

Solution.

1. $\displaystyle\sum_{k=1}^{100} k$
2. $\displaystyle\sum_{k=0}^{50} 2^k$
3. $\displaystyle\sum_{k=2}^{n} (4k - 2)$

If we want to multiply the a_k instead, we would write $\displaystyle\prod_{k=1}^{n} a_k$. For example, $\displaystyle\prod_{k=1}^{n} k = n!$.

Exercises

1. Consider the sequence $5, 9, 13, 17, 21, \ldots$ with $a_1 = 5$

(a) Give a recursive definition for the sequence.

(b) Give a closed formula for the nth term of the sequence.

(c) Is 2013 a term in the sequence? Explain.

(d) How many terms does the sequence $5, 9, 13, 17, 21, \ldots, 533$ have?

(e) Find the sum: $5 + 9 + 13 + 17 + 21 + \cdots + 533$. Show your work.

(f) Use what you found above to find b_n, the nth term of $1, 6, 15, 28, 45, \ldots$, where $b_0 = 1$

2. Consider the sequence $(a_n)_{n\geq 0}$ which starts $8, 14, 20, 26, \ldots$.

(a) What is the next term in the sequence?

(b) Find a formula for the nth term of this sequence.

(c) Find the sum of the first 100 terms of the sequence: $\sum_{k=0}^{99} a_k$.

3. Consider the sum $4 + 11 + 18 + 25 + \cdots + 249$.

(a) How many terms (summands) are in the sum?

(b) Compute the sum. Remember to show all your work.

4. Consider the sequence $1, 7, 13, 19, \ldots, 6n + 7$.

(a) How many terms are there in the sequence?

(b) What is the second-to-last term?

(c) Find the sum of all the terms in the sequence.

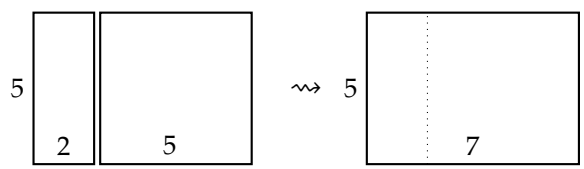
5. Find $5 + 7 + 9 + 11 + \cdots + 521$.

6. Find $5 + 15 + 45 + \cdots + 5 \cdot 3^{20}$.

7. Find $1 - \frac{2}{3} + \frac{4}{9} - \cdots + \frac{2^{30}}{3^{30}}$.

8. Find x and y such that $27, x, y, 1$ is part of an arithmetic sequence. Then find x and y so that the sequence is part of a geometric sequence. (Warning: x and y might not be integers.)

9. Starting with any rectangle, we can create a new, larger rectangle by attaching a square to the longer side. For example, if we start with a 2×5 rectangle, we would glue on a 5×5 square, forming a 5×7 rectangle:



(a) Create a sequence of rectangles using this rule starting with a 1×2 rectangle. Then write out the sequence of *perimeters* for the rectangles (the first term of the sequence would be 6, since the perimeter of a 1×2 rectangle is 6 - the next term would be 10).

(b) Repeat the above part this time starting with a 1×3 rectangle.

(c) Find recursive formulas for each of the sequences of perimeters you found in parts (a) and (b). Don't forget to give the initial conditions as well.

(d) Are the sequences arithmetic? Geometric? If not, are they *close* to being either of these (i.e., are the differences or ratios *almost* constant)? Explain.

10. Consider the sequence $2, 7, 15, 26, 40, 57, \ldots$ (with $a_0 = 2$). By looking at the differences between terms, express the sequence as a sequence of partial sums. Then find a closed formula for the sequence by computing the nth partial sum.

11. If you have enough toothpicks, you can make a large triangular grid. Below, are the triangular grids of size 1 and of size 2. The size 1 grid requires 3 toothpicks, the size 2 grid requires 9 toothpicks.

(a) Let t_n be the number of toothpicks required to make a size n triangular grid. Write out the first 5 terms of the sequence $t_1, t_2, \ldots$.

(b) Find a recursive definition for the sequence. Explain why you are correct.

(c) Is the sequence arithmetic or geometric? If not, is it the sequence of partial sums of an arithmetic or geometric sequence? Explain why your answer is correct.

(d) Use your results from part (c) to find a closed formula for the sequence. Show your work.

12. Use summation ($\sum$) or product ($\prod$) notation to rewrite the following.

(a) $2 + 4 + 6 + 8 + \cdots + 2n$.

(b) $1 + 5 + 9 + 13 + \cdots + 425$.

(c) $1 + \frac{1}{2} + \frac{1}{3} + \frac{1}{4} + \cdots + \frac{1}{50}$.

(d) $2 \cdot 4 \cdot 6 \cdot \cdots \cdot 2n$.

(e) $(\frac{1}{2})(\frac{2}{3})(\frac{3}{4}) \cdots (\frac{100}{101})$.

13. Expand the following sums and products. That is, write them out the long way.

(a) $\displaystyle\sum_{k=1}^{100}(3 + 4k)$.

(b) $\displaystyle\sum_{k=0}^{n} 2^k$.

(c) $\displaystyle\sum_{k=2}^{50} \frac{1}{(k^2 - 1)}$.

(d) $\displaystyle\prod_{k=2}^{100} \frac{k^2}{(k^2 - 1)}$.

(e) $\displaystyle\prod_{k=0}^{n}(2 + 3k)$.

2.3 Polynomial Fitting

Investigate!

A standard 8×8 chessboard contains 64 squares. Actually, this is just the number of unit squares. How many squares of all sizes are there on a chessboard? Start with smaller boards: 1×1, 2×2, 3×3, etc. Find a formula for the total number of squares in an $n \times n$ board.

Attempt the above activity before proceeding

So far we have seen methods for finding the closed formulas for arithmetic and geometric sequences. Since we know how to compute the sum of the first n terms of arithmetic and geometric sequences, we can compute the closed formulas for sequences which have an arithmetic (or geometric) sequence of differences between terms. But what if we consider a sequence which is the sum of the first n terms of a sequence which is itself the sum of an arithmetic sequence?

Before we get too carried away, let's consider an example: How many squares (of all sizes) are there on a chessboard? A chessboard consists of 64 squares, but we also want to consider squares of longer side length. Even though we are only considering an 8×8 board, there is already a lot to count. So instead, let us build a sequence: the first term will be the number of squares on a 1×1 board, the second term will be the number of squares on a 2×2 board, and so on. After a little thought, we arrive at the sequence

$$1, 5, 14, 30, 55, \ldots$$

This sequence is not arithmetic (or geometric for that matter), but perhaps it's sequence of differences is. For differences we get

$$4, 9, 16, 25, \ldots$$

Not a huge surprise: one way to count the number of squares in a 4×4 chessboard is to notice that there are 16 squares with side length 1, 9 with side length 2, 4 with side length 3 and 1 with side length 4. So the original sequence is just the sum of squares. Now this sequence of differences is not arithmetic since it's sequence of differences (the differences of the differences of the original sequence) is not constant. In fact, this sequence of **second differences** is

$$5, 7, 9, \ldots$$

which *is* an arithmetic sequence (with constant difference 2). Notice that our original sequence had **third differences** (that is, differences of differences of differences of the original) constant. We will call such a sequence Δ^3-constant. The sequence $1, 4, 9, 16, \ldots$ has second differences constant, so it will be a Δ^2-constant sequence. In general, we will say a sequence is a Δ^k**-constant** sequence if the kth differences are constant.

Example 2.3.1

Which of the following sequences are Δ^k-constant for some value of k?

1. $2, 3, 7, 14, 24, 37, \ldots$.
2. $1, 8, 27, 64, 125, 216, \ldots$.
3. $1, 2, 4, 8, 16, 64, 128, \ldots$.

Solution.

1. This is the sequence from Example 2.2.6, in which we found a closed formula by recognizing the sequence as the sequence of partial sums of an arithmetic sequence. Indeed, the sequence of first differences is $1, 4, 7, 10, 13, \ldots$, which itself has differences $3, 3, 3, 3, \ldots$. Thus $2, 3, 7, 14, 24, 37, \ldots$ is a Δ^2-constant sequence.
2. These are the perfect cubes. The sequence of first differences is $7, 19, 37, 61, 91, \ldots$; the sequence of second differences is $12, 18, 24, 30, \ldots$; the sequence of third differences is constant: $6, 6, 6, \ldots$. Thus the perfect cubes are a Δ^3-constant sequence.
3. If we take first differences we get $1, 2, 4, 8, 16, \ldots$. Wait, what? That's the sequence we started with. So taking second differences will give us the same sequence again. No matter how many times we repeat this we will always have the same sequence, which in particular means no finite number of differences will be constant. Thus this sequence is not Δ^k-constant for any k.

The Δ^0-constant sequences are themselves constant, so a closed formula for them is easy to compute (it's just the constant). The Δ^1-constant sequences are arithmetic and we have a method for finding closed formulas for them as well. Every Δ^2-constant sequence is the sum of an arithmetic sequence so we can find formulas for these as well. But notice that the format of the closed formula for a Δ^2-constant sequence is always quadratic. For example, the square numbers are Δ^2-constant with closed formula $a_n = n^2$. The triangular numbers (also Δ^2-constant) have closed formula $a_n = \frac{n(n+1)}{2}$, which when multiplied out gives you an n^2 term as well. It appears that every time we increase the complexity of the sequence, that is, increase the number of differences before we get constants, we also increase the degree of the polynomial used for the closed formula. We go from constant to linear to quadratic. The sequence of differences between terms tells us something about the rate of growth of the sequence. If a sequence is growing at a constant rate, then the formula for the sequence will be linear. If the sequence is growing at a rate which itself is growing at a constant rate, then the formula is quadratic. You have seen this elsewhere: if a function has a constant second derivative (rate of change) then the function must be quadratic.

This works in general:

Finite Differences

The closed formula for a sequence will be a degree k polynomial if and only if the sequence is Δ^k-constant (i.e., the kth sequence of differences is constant).

This tells us that the sequence of numbers of squares on a chessboard, $1, 5, 14, 30, 55, \ldots$, which we saw to be Δ^3-constant, will have a cubic (degree 3 polynomial) for its closed formula.

Now once we know what format the closed formula for a sequence will take, it is much easier to actually find the closed formula. In the case that the closed formula is a degree k polynomial, we just need $k+1$ data points to "fit" the polynomial to the data.

Example 2.3.2
Find a formula for the sequence $3, 7, 14, 24, \ldots$. Assume $a_1 = 3$.

Solution. First, check to see if the formula has constant differences at some level. The sequence of first differences is $4, 7, 10, \ldots$ which is arithmetic, so the sequence of second differences is constant. The sequence is Δ^2-constant, so the formula for a_n will be a degree 2 polynomial. That is, we know that for some constants a, b, and c,

$$a_n = an^2 + bn + c.$$

Now to find a, b, and c. First, it would be nice to know what a_0 is, since plugging in $n = 0$ simplifies the above formula greatly. In this case, $a_0 = 2$ (work backwards from the sequence of constant differences). Thus

$$a_0 = 2 = a \cdot 0^2 + b \cdot 0 + c,$$

so $c = 2$. Now plug in $n = 1$ and $n = 2$. We get

$$a_1 = 3 = a + b + 2$$

$$a_2 = 7 = a4 + b2 + 2.$$

At this point we have two (linear) equations and two unknowns, so we can solve the system for a and b (using substitution or elimination or even matrices). We find $a = \frac{3}{2}$ and $b = \frac{-1}{2}$, so $a_n = \frac{3}{2}n^2 - \frac{1}{2}n + 2$.

Example 2.3.3
Find a closed formula for the number of squares on an $n \times n$ chessboard.

Solution. We have seen that the sequence $1, 5, 14, 30, 55, \ldots$ is Δ^3-constant, so we are looking for a degree 3 polynomial. That is,

$$a_n = an^3 + bn^2 + cn + d.$$

We can find d if we know what a_0 is. Working backwards from the third differences, we find $a_0 = 0$ (unsurprisingly, since there are no squares on a 0×0 chessboard). Thus $d = 0$. Now plug in $n = 1$, $n = 2$, and $n = 3$:

$$\begin{aligned} 1 &= a + b + c \\ 5 &= 8a + 4b + 2c \\ 14 &= 27a + 9b + 3c. \end{aligned}$$

If we solve this system of equations we get $a = \frac{1}{3}$, $b = \frac{1}{2}$ and $c = \frac{1}{6}$. Therefore the number of squares on an $n \times n$ chessboard is $a_n = \frac{1}{3}n^3 + \frac{1}{2}n^2 + \frac{1}{6}n$.

Note: Since the squares-on-a-chessboard problem is really asking for the sum of squares, we now have a nice formula for $\sum_{k=1}^{n} k^2$.

Not all sequences will have polynomials as their closed formula. We can use the theory of finite differences to identify these.

Example 2.3.4
Determine whether the following sequences can be described by a polynomial, and if so, of what degree.

1. $1, 2, 4, 8, 16, \ldots$
2. $0, 7, 50, 183, 484, 1055, \ldots$
3. $1, 1, 2, 3, 5, 8, 13, \ldots$

Solution.

1. As we saw in Example 2.3.1, this sequence is not Δ^k-constant for any k. Therefore the closed formula for the sequence is not a polynomial. In fact, we know the closed formula is $a_n = 2^n$, which grows faster than any polynomial (so is not a polynomial).
2. The sequence of first differences is $7, 43, 133, 301, 571, \ldots$. The second differences are: $36, 90, 168, 270, \ldots$. Third difference: $54, 78, 102, \ldots$. Fourth differences: $24, 24, \ldots$. As far as we can tell, this sequence of differences is constant so the sequence is Δ^4-constant and as such the closed formula is a degree 4 polynomial.
3. This is the Fibonacci sequence. The sequence of first differences is $0, 1, 1, 2, 3, 5, 8, \ldots$, the second differences are $1, 0, 1, 1, 2, 3, 5 \ldots$. We notice that after the first few terms, we get the original sequence back. So there will never be constant differences, so the closed formula for the Fibonacci sequence is not a polynomial.

Exercises

1. Use polynomial fitting to find the formula for the nth term of the sequences $(a_n)_{n\geq 0}$ below.

(a) $2, 5, 11, 21, 36, \ldots$
(b) $0, 2, 6, 12, 20, \ldots$
(c) $1, 2, 4, 8, 15, 26 \ldots$
(d) $3, 6, 12, 22, 37, \ldots$. After finding a formula here, compare to part (a).

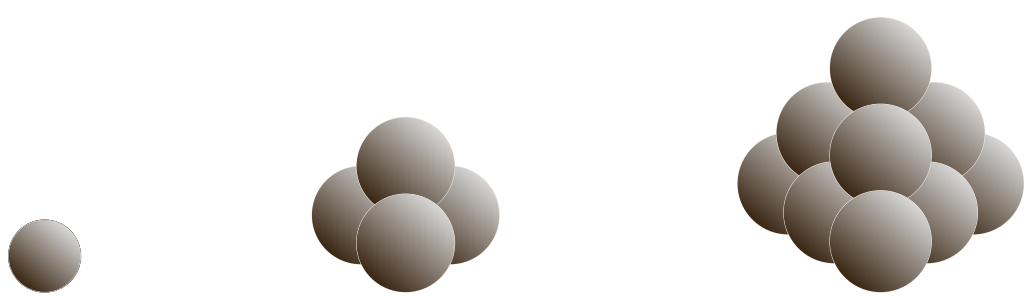
2. Make up a sequences that have

(a) $3, 3, 3, 3, \ldots$ as its second differences.
(b) $1, 2, 3, 4, 5, \ldots$ as its third differences.
(c) $1, 2, 4, 8, 16, \ldots$ as its 100th differences.

3. Consider the sequence $1, 3, 7, 13, 21, \ldots$. Explain how you know the closed formula for the sequence will be quadratic. Then "guess" the correct formula by comparing this sequence to the squares $1, 4, 9, 16, \ldots$ (do not use polynomial fitting).

4. Use a similar technique as in the previous exercise to find a closed formula for the sequence $2, 11, 34, 77, 146, 247, \ldots$.

5. In their down time, ghost pirates enjoy stacking cannonballs in triangular based pyramids (aka, tetrahedrons), like those pictured here:



Note, in the picture on the right, there are some cannonballs (actually just one) you cannot see. The next picture would have 4 cannonballs you cannot see. The stacks are *not* hollow.
The pirates wonder how many cannonballs would be required to build a pyramid 15 layers high (thus breaking the world cannonball stacking record). Can you help?

(a) Let $P(n)$ denote the number of cannonballs needed to create a pyramid n layers high. So $P(1) = 1$, $P(2) = 4$, and so on. Calculate $P(3)$, $P(4)$ and $P(5)$.

(b) Use polynomial fitting to find a closed formula for $P(n)$. Show your work.

(c) Answer the pirate's question: how many cannonballs do they need to make a pyramid 15 layers high?

6. Suppose $a_n = n^2 + 3n + 4$. Find a closed formula for the sequence of differences by computing $a_n - a_{n-1}$.

7. Repeat the above assuming this time $a_n = an^2 + bn + c$. That is, prove that every quadratic sequence has arithmetic differences.

8. Can you use polynomial fitting to find the formula for the nth term of the sequence 4, 7, 11, 18, 29, 47, . . . ? Explain why or why not.

9. Will the nth sequence of differences of $2, 6, 18, 54, 162, \ldots$ ever be constant? Explain.

10. Consider the sequences $2, 5, 12, 29, 70, 169, 408, \ldots$ (with $a_0 = 2$).

(a) Describe the rate of growth of this sequence.

(b) Find a recursive definition for the sequence.

(c) Find a closed formula for the sequence.

(d) If you look at the sequence of differences between terms, and then the sequence of second differences, the sequence of third differences, and so on, will you ever get a constant sequence? Explain how you know.

2.4 Solving Recurrence Relations

Investigate!

Consider the recurrence relation

$$a_n = 5a_{n-1} - 6a_{n-2}.$$

1. What sequence do you get if the initial conditions are $a_0 = 1$, $a_1 = 2$? Give a closed formula for this sequence.
2. What sequence do you get if the initial conditions are $a_0 = 1$, $a_1 = 3$? Give a closed formula.
3. What if $a_0 = 2$ and $a_1 = 5$? Find a closed formula.

Attempt the above activity before proceeding

We have seen that it is often easier to find recursive definitions than closed formulas. Lucky for us, there are a few techniques for converting recursive definitions to closed formulas. Doing so is called **solving a recurrence relation**. Recall that the recurrence relation is a recursive definition without the initial conditions. For example, the recurrence relation for the Fibonacci sequence is $F_n = F_{n-1} + F_{n-2}$. (This, together with the initial conditions $F_0 = 0$ and $F_1 = 1$ give the entire recursive *definition* for the sequence.)

Example 2.4.1
Find a recurrence relation and initial conditions for $1, 5, 17, 53, 161, 485 \ldots$.

Solution. Finding the recurrence relation would be easier if we had some context for the problem (like the Tower of Hanoi, for example). Alas, we have only the sequence. Remember, the recurrence relation tells you how to get from previous terms to future terms. What is going on here? We could

look at the differences between terms: $4, 12, 36, 108, \ldots$. Notice that these are growing by a factor of 3. Is the original sequence as well? $1 \cdot 3 = 3$, $5 \cdot 3 = 15$, $17 \cdot 3 = 51$ and so on. It appears that we always end up with 2 less than the next term. Aha!

So $a_n = 3a_{n-1} + 2$ is our recurrence relation and the initial condition is $a_0 = 1$.

We are going to try to *solve* these recurrence relations. By this we mean something very similar to solving differential equations: we want to find a function of n (a closed formula) which satisfies the recurrence relation, as well as the initial condition.[2] Just like for differential equations, finding a solution might be tricky, but checking that the solution is correct is easy.

Example 2.4.2
Check that $a_n = 2^n + 1$ is a solution to the recurrence relation $a_n = 2a_{n-1} - 1$ with $a_1 = 3$.

Solution. First, it is easy to check the initial condition: a_1 should be $2^1 + 1$ according to our closed formula. Indeed, $2^1 + 1 = 3$, which is what we want. To check that our proposed solution satisfies the recurrence relation, try plugging it in.

$$\begin{aligned} 2a_{n-1} - 1 &= 2(2^{n-1} + 1) - 1 \\ &= 2^n + 2 - 1 \\ &= 2^n + 1 \\ &= a_n. \end{aligned}$$

That's what our recurrence relation says! We have a solution.

Sometimes we can be clever and solve a recurrence relation by inspection. We generate the sequence using the recurrence relation and keep track of what we are doing so that we can see how to jump to finding just the a_n term. Here are two examples of how you might do that.

Telescoping refers to the phenomenon when many terms in a large sum cancel out - so the sum "telescopes." For example:

$$(2-1) + (3-2) + (4-3) + \cdots + (100-99) + (101-100) = -1 + 101$$

because every third term looks like: $2 + -2 = 0$, and then $3 + -3 = 0$ and so on.

We can use this behavior to solve recurrence relations. Here is an example.

Example 2.4.3
Solve the recurrence relation $a_n = a_{n-1} + n$ with initial term $a_0 = 4$.

Solution. To get a feel for the recurrence relation, write out the first few terms of the sequence: $4, 5, 7, 10, 14, 19, \ldots$. Look at the difference between terms. $a_1 - a_0 = 1$ and $a_2 - a_1 = 2$ and so on. The key thing here is that the difference between terms is n. We can write this explicitly: $a_n - a_{n-1} = n$. Of course, we could have arrived at this conclusion directly from the recurrence relation by subtracting a_{n-1} from both sides.

Now use this equation over and over again, changing n each time:

$$\begin{aligned} a_1 - a_0 &= 1 \\ a_2 - a_1 &= 2 \\ a_3 - a_2 &= 3 \\ \vdots \quad & \quad \vdots \\ a_n - a_{n-1} &= n. \end{aligned}$$

[2]Recurrence relations are sometimes called difference equations since they can describe the difference between terms and this highlights the relation to differential equations further.

Add all these equations together. On the right-hand side, we get the sum $1 + 2 + 3 + \cdots + n$. We already know this can be simplified to $\frac{n(n+1)}{2}$. What happens on the left-hand side? We get

$$(a_1 - a_0) + (a_2 - a_1) + (a_3 - a_2) + \cdots (a_{n-1} - a_{n-2}) + (a_n - a_{n-1}).$$

This sum telescopes. We are left with only the $-a_0$ from the first equation and the a_n from the last equation. Putting this all together we have $-a_0 + a_n = \frac{n(n+1)}{2}$ or $a_n = \frac{n(n+1)}{2} + a_0$. But we know that $a_0 = 4$. So the solution to the recurrence relation, subject to the initial condition is

$$a_n = \frac{n(n+1)}{2} + 4.$$

(Now that we know that, we should notice that the sequence is the result of adding 4 to each of the triangular numbers.)

The above example shows a way to solve recurrence relations of the form $a_n = a_{n-1} + f(n)$ where $\sum_{k=1}^{n} f(k)$ has a known closed formula. If you rewrite the recurrence relation as $a_n - a_{n-1} = f(n)$, and then add up all the different equations with n ranging between 1 and n, the left-hand side will always give you $a_n - a_0$. The right-hand side will be $\sum_{k=1}^{n} f(k)$, which is why we need to know the closed formula for that sum.

However, telescoping will not help us with a recursion such as $a_n = 3a_{n-1} + 2$ since the left-hand side will not telescope. You will have $-3a_{n-1}$'s but only one a_{n-1}. However, we can still be clever if we use **iteration**.

We have already seen an example of iteration when we found the closed formula for arithmetic and geometric sequences. The idea is, we *iterate* the process of finding the next term, starting with the known initial condition, up until we have a_n. Then we simplify. In the arithmetic sequence example, we simplified by multiplying d by the number of times we add it to a when we get to a_n, to get from $a_n = a + d + d + d + \cdots + d$ to $a_n = a + dn$.

To see how this works, let's go through the same example we used for telescoping, but this time use iteration.

Example 2.4.4
Use iteration to solve the recurrence relation $a_n = a_{n-1} + n$ with $a_0 = 4$.

Solution. Again, start by writing down the recurrence relation when $n = 1$. This time, don't subtract the a_{n-1} terms to the other side:

$$a_1 = a_0 + 1.$$

Now $a_2 = a_1 + 2$, but we know what a_1 is. By substitution, we get

$$a_2 = (a_0 + 1) + 2.$$

Now go to $a_3 = a_2 + 3$, using our known value of a_2:

$$a_3 = ((a_0 + 1) + 2) + 3.$$

We notice a pattern. Each time, we take the previous term and add the current index. So

$$a_n = ((((a_0 + 1) + 2) + 3) + \cdots + n - 1) + n.$$

Regrouping terms, we notice that a_n is just a_0 plus the sum of the integers from 1 to n. So, since $a_0 = 4$,

$$a_n = 4 + \frac{n(n+1)}{2}.$$

Of course in this case we still needed to know formula for the sum of $1, \ldots, n$. Let's try iteration with a sequence for which telescoping doesn't work.

Example 2.4.5
Solve the recurrence relation $a_n = 3a_{n-1} + 2$ subject to $a_0 = 1$.

Solution. Again, we iterate the recurrence relation, building up to the index n.

$$\begin{aligned}
a_1 &= 3a_0 + 2 \\
a_2 &= 3(a_1) + 2 = 3(3a_0 + 2) + 2 \\
a_3 &= 3[a_2] + 2 = 3[3(3a_0 + 2) + 2] + 2 \\
&\vdots \\
a_n &= 3(a_{n-1}) + 2 = 3(3(3(3 \cdots (3a_0 + 2) + 2) + 2) \cdots + 2) + 2.
\end{aligned}$$

It is difficult to see what is happening here because we have to distribute all those 3's. Let's try again, this time simplifying a bit as we go.

$$\begin{aligned}
a_1 &= 3a_0 + 2 \\
a_2 &= 3(a_1) + 2 = 3(3a_0 + 2) + 2 = 3^2 a_0 + 2 \cdot 3 + 2 \\
a_3 &= 3[a_2] + 2 = 3[3^2 a_0 + 2 \cdot 3 + 2] + 2 = 3^3 a_0 + 2 \cdot 3^2 + 2 \cdot 3 + 2 \\
&\vdots \\
a_n &= 3(a_{n-1}) + 2 = 3(3^{n-1} a_0 + 2 \cdot 3^{n-2} + \cdots + 2) + 2 \\
&\qquad = 3^n a_0 + 2 \cdot 3^{n-1} + 2 \cdot 3^{n-2} + \cdots + 2 \cdot 3 + 2.
\end{aligned}$$

Now we simplify. $a_0 = 1$, so we have $3^n + \langle \text{stuff} \rangle$. Note that all the other terms have a 2 in them. In fact, we have a geometric sum with first term 2 and common ratio 3. We have seen how to simplify $2 + 2 \cdot 3 + 2 \cdot 3^2 + \cdots + 2 \cdot 3^{n-1}$. We get $\frac{2-2\cdot 3^n}{-2}$ which simplifies to $3^n - 1$. Putting this together with the first 3^n term gives our closed formula:

$$a_n = 2 \cdot 3^n - 1.$$

Iteration can be messy, but when the recurrence relation only refers to one previous term (and maybe some function of n) it can work well. However, trying to iterate a recurrence relation such as $a_n = 2a_{n-1} + 3a_{n-2}$ will be way too complicated. We would need to keep track of two sets of previous terms, each of which were expressed by two previous terms, and so on. The length of the formula would grow exponentially (double each time, in fact). Luckily there happens to be a method for solving recurrence relations which works very well on relations like this.

The Characteristic Root Technique

Suppose we want to solve a recurrence relation expressed as a combination of the two previous terms, such as $a_n = a_{n-1} + 6a_{n-2}$. In other words, we want to find a function of n which satisfies $a_n - a_{n-1} - 6a_{n-2} = 0$. Now iteration is too complicated, but think just for a second what would happen if we *did* iterate. In each step, we would, among other things, multiply a previous iteration by 6. So our closed formula would include 6 multiplied some number of times. Thus it is reasonable to guess the solution will contain parts that look geometric. Perhaps the solution will take the form r^n for some constant r.

The nice thing is, we know how to check whether a formula is actually a solution to a recurrence relation: plug it in. What happens if we plug in r^n into the recursion above? We get

$$r^n - r^{n-1} - 6r^{n-2} = 0.$$

Now solve for r:

$$r^{n-2}(r^2 - r - 6) = 0,$$

so by factoring, $r = -2$ or $r = 3$ (or $r = 0$, although this does not help us). This tells us that $a_n = (-2)^n$ is a solution to the recurrence relation, as is $a_n = 3^n$. Which one is correct? They both are, unless we specify initial conditions. Notice we could also have $a_n = (-2)^n + 3^n$. Or $a_n = 7(-2)^n + 4 \cdot 3^n$. In fact, for any a and

b, $a_n = a(-2)^n + b3^n$ is a solution (try plugging this into the recurrence relation). To find the values of a and b, use the initial conditions.

This points us in the direction of a more general technique for solving recurrence relations. Notice we will always be able to factor out the r^{n-2} as we did above. So we really only care about the other part. We call this other part the **characteristic equation** for the recurrence relation. We are interested in finding the roots of the characteristic equation, which are called (surprise) the **characteristic roots**.

Characteristic Roots

Given a recurrence relation $a_n + \alpha a_{n-1} + \beta a_{n-2} = 0$, the **characteristic polynomial** is

$$x^2 + \alpha x + \beta$$

giving the **characteristic equation**:

$$x^2 + \alpha x + \beta = 0.$$

If r_1 and r_2 are two distinct roots of the characteristic polynomial (i.e, solutions to the characteristic equation), then the solution to the recurrence relation is

$$a_n = ar_1^n + br_2^n,$$

where a and b are constants determined by the initial conditions.

Example 2.4.6
Solve the recurrence relation $a_n = 7a_{n-1} - 10a_{n-2}$ with $a_0 = 2$ and $a_1 = 3$.

Solution. Rewrite the recurrence relation $a_n - 7a_{n-1} + 10a_{n-2} = 0$. Now form the characteristic equation:

$$x^2 - 7x + 10 = 0$$

and solve for x:

$$(x-2)(x-5) = 0$$

so $x = 2$ and $x = 5$ are the characteristic roots. We therefore know that the solution to the recurrence relation will have the form

$$a_n = a2^n + b5^n.$$

To find a and b, plug in $n = 0$ and $n = 1$ to get a system of two equations with two unknowns:

$$2 = a2^0 + b5^0 = a + b$$
$$3 = a2^1 + b5^1 = 2a + 5b$$

Solving this system gives $a = \frac{7}{3}$ and $b = -\frac{1}{3}$ so the solution to the recurrence relation is

$$a_n = \frac{7}{3}2^n - \frac{1}{3}3^n.$$

Perhaps the most famous recurrence relation is $F_n = F_{n-1} + F_{n-2}$, which together with the initial conditions $F_0 = 0$ and $F_1 = 1$ defines the Fibonacci sequence. But notice that this is precisely the type of recurrence relation on which we can use the characteristic root technique. When you do, the only thing that changes is that the characteristic equation does not factor, so you need to use the quadratic formula to find the characteristic roots. In fact, doing so gives the third most famous irrational number, φ, the **golden ratio**.

Before leaving the characteristic root technique, we should think about what might happen when you solve the characteristic equation. We have an example above in which the characteristic polynomial has two distinct roots. These roots can be integers, or perhaps irrational numbers (requiring the quadratic formula to find them). In these cases, we know what the solution to the recurrence relation looks like.

However, it is possible for the characteristic polynomial to only have one root. This can happen if the characteristic polynomial factors as $(x-r)^2$. It is still the case that r^n would be a solution to the recurrence

relation, but we won't be able to find solutions for all initial conditions using the general form $a_n = ar_1^n + br_2^n$, since we can't distinguish between r_1^n and r_2^n. We are in luck though:

Characteristic Root Technique for Repeated Roots

Suppose the recurrence relation $a_n = \alpha a_{n-1} + \beta a_{n-2}$ has a characteristic polynomial with only one root r. Then the solution to the recurrence relation is

$$a_n = ar^n + bnr^n$$

where a and b are constants determined by the initial conditions.

Notice the extra n in bnr^n. This allows us to solve for the constants a and b from the initial conditions.

Example 2.4.7
Solve the recurrence relation $a_n = 6a_{n-1} - 9a_{n-2}$ with initial conditions $a_0 = 1$ and $a_1 = 4$.

Solution. The characteristic polynomial is $x^2 - 6x + 9$. We solve the characteristic equation

$$x^2 - 6x + 9 = 0$$

by factoring:

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

so $x = 3$ is the only characteristic root. Therefore we know that the solution to the recurrence relation has the form

$$a_n = a3^n + bn3^n$$

for some constants a and b. Now use the initial conditions:

$$\begin{aligned} a_0 &= 1 = a3^0 + b \cdot 0 \cdot 3^0 = a \\ a_1 &= 4 = a \cdot 3 + b \cdot 1 \cdot 3 = 3a + 3b. \end{aligned}$$

Since $a = 1$, we find that $b = \frac{1}{3}$. Therefore the solution to the recurrence relation is

$$a_n = 3^n + \frac{1}{3}n3^n.$$

Although we will not consider examples more complicated than these, this characteristic root technique can be applied to much more complicated recurrence relations. For example, $a_n = 2a_{n-1} + a_{n-2} - 3a_{n-3}$ has characteristic polynomial $x^3 - 2x^2 - x + 3$. Assuming you see how to factor such a degree 3 (or more) polynomial you can easily find the characteristic roots and as such solve the recurrence relation (the solution would look like $a_n = ar_1^n + br_2^n + cr_3^n$ if there were 3 distinct roots). It is also possible to solve recurrence relations of the form $a_n = \alpha a_{n-1} + \beta a_{n-2} + C$ for some constant C. It is also possible (and acceptable) for the characteristic roots to be complex numbers.

Exercises

1. Find the next two terms in $(a_n)_{n\geq 0}$ beginning $3, 5, 11, 21, 43, 85 \ldots$. Then give a recursive definition for the sequence. Finally, use the characteristic root technique to find a closed formula for the sequence.

2. Solve the recurrence relation $a_n = a_{n-1} + 2^n$ with $a_0 = 5$.

3. Show that 4^n is a solution to the recurrence relation $a_n = 3a_{n-1} + 4a_{n-2}$.

4. Find the solution to the recurrence relation $a_n = 3a_{n-1} + 4a_{n-2}$ with initial terms $a_0 = 2$ and $a_1 = 3$.

5. Find the solution to the recurrence relation $a_n = 3a_{n-1} + 4a_{n-2}$ with initial terms $a_0 = 5$ and $a_1 = 8$.

6. Solve the recurrence relation $a_n = 2a_{n-1} - a_{n-2}$.

(a) What is the solution if the initial terms are $a_0 = 1$ and $a_1 = 2$?

(b) What do the initial terms need to be in order for $a_9 = 30$?

(c) For which x are there initial terms which make $a_9 = x$?

7. Solve the recurrence relation $a_n = 3a_{n-1} + 10a_{n-2}$ with initial terms $a_0 = 4$ and $a_1 = 1$.

8. Suppose that r^n and q^n are both solutions to a recurrence relation of the form $a_n = \alpha a_{n-1} + \beta a_{n-2}$. Prove that $c \cdot r^n + d \cdot q^n$ is also a solution to the recurrence relation, for any constants c, d.

9. Think back to the magical candy machine at your neighborhood grocery store. Suppose that the first time a quarter is put into the machine 1 Skittle comes out. The second time, 4 Skittles, the third time 16 Skittles, the fourth time 64 Skittles, etc.

(a) Find both a recursive and closed formula for how many Skittles the nth customer gets.

(b) Check your solution for the closed formula by solving the recurrence relation using the Characteristic Root technique.

10. You have access to 1×1 tiles which come in 2 different colors and 1×2 tiles which come in 3 different colors. We want to figure out how many different $1 \times n$ path designs we can make out of these tiles.

(a) Find a recursive definition for the sequence a_n of paths of length n.

(b) Solve the recurrence relation using the Characteristic Root technique.

11. Let a_n be the number of $1 \times n$ tile designs you can make using 1×1 squares available in 4 colors and 1×2 dominoes available in 5 colors.

(a) First, find a recurrence relation to describe the problem. Explain why the recurrence relation is correct (in the context of the problem).

(b) Write out the first 6 terms of the sequence $a_1, a_2, \ldots$.

(c) Solve the recurrence relation. That is, find a closed formula for a_n.

12. Consider the recurrence relation $a_n = 4a_{n-1} - 4a_{n-2}$.

(a) Find the general solution to the recurrence relation (beware the repeated root).

(b) Find the solution when $a_0 = 1$ and $a_1 = 2$.

(c) Find the solution when $a_0 = 1$ and $a_1 = 8$.

2.5 Induction

Mathematical induction is a proof technique, not unlike direct proof or proof by contradiction or combinatorial proof.[3] In other words, induction is a style of argument we use to convince ourselves and others that a mathematical statement is always true. Many mathematical statements can be proved by simply explaining what they mean. Others are very difficult to prove—in fact, there are relatively simple mathematical statements which nobody yet knows how to prove. To facilitate the discovery of proofs, it is important to be familiar with some standard styles of arguments. Induction is one such style. Let's start with an example:

Stamps

Investigate!

[3] You might or might not be familiar with these yet. We will consider these in Chapter 3.

Chapter 8

MFCS 10: Recurrences

10 Recurrences

A recurrence describes a sequence of numbers. Early terms are specified explicitly and later terms are expressed as a function of their predecessors. As a trivial example, this recurrence describes the sequence 1, 2, 3, etc.:

$$\begin{aligned} T_1 &= 1 \\ T_n &= T_{n-1} + 1 \qquad (\text{for } n \geq 2). \end{aligned}$$

Here, the first term is defined to be 1 and each subsequent term is one more than its predecessor.

Recurrences turn out to be a powerful tool. In this chapter, we'll emphasize using recurrences to analyze the performance of recursive algorithms. However, recurrences have other applications in computer science as well, such as enumeration of structures and analysis of random processes. And, as we saw in Section 9.4, they also arise in the analysis of problems in the physical sciences.

A recurrence in isolation is not a very useful description of a sequence. One can not easily answer simple questions such as, "What is the hundredth term?" or "What is the asymptotic growth rate?" So one typically wants to *solve* a recurrence; that is, to find a closed-form expression for the nth term.

We'll first introduce two general solving techniques: guess-and-verify and plug-and-chug. These methods are applicable to every recurrence, but their success requires a flash of insight—sometimes an unrealistically brilliant flash. So we'll also introduce two big classes of recurrences, linear and divide-and-conquer, that often come up in computer science. Essentially all recurrences in these two classes are solvable using cookbook techniques; you follow the recipe and get the answer. A drawback is that calculation replaces insight. The "Aha!" moment that is essential in the guess-and-verify and plug-and-chug methods is replaced by a "Huh" at the end of a cookbook procedure.

At the end of the chapter, we'll develop rules of thumb to help you assess many recurrences without any calculation. These rules can help you distinguish promising approaches from bad ideas early in the process of designing an algorithm.

Recurrences are one aspect of a broad theme in computer science: reducing a big problem to progressively smaller problems until easy base cases are reached. This same idea underlies both induction proofs and recursive algorithms. As we'll see, all three ideas snap together nicely. For example, one might describe the running time of a recursive algorithm with a recurrence and use induction to verify the solution.

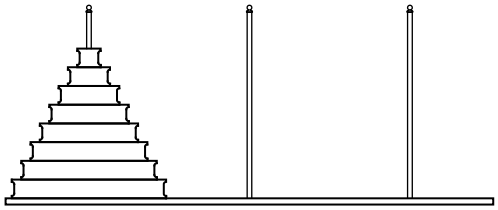


Figure 10.1 The initial configuration of the disks in the Towers of Hanoi problem.

10.1 The Towers of Hanoi

According to legend, there is a temple in Hanoi with three posts and 64 gold disks of different sizes. Each disk has a hole through the center so that it fits on a post. In the misty past, all the disks were on the first post, with the largest on the bottom and the smallest on top, as shown in Figure 10.1.

Monks in the temple have labored through the years since to move all the disks to one of the other two posts according to the following rules:

- The only permitted action is removing the top disk from one post and dropping it onto another post.
- A larger disk can never lie above a smaller disk on any post.

So, for example, picking up the whole stack of disks at once and dropping them on another post is illegal. That's good, because the legend says that when the monks complete the puzzle, the world will end!

To clarify the problem, suppose there were only 3 gold disks instead of 64. Then the puzzle could be solved in 7 steps as shown in Figure 10.2.

The questions we must answer are, "Given sufficient time, can the monks succeed?" If so, "How long until the world ends?" And, most importantly, "Will this happen before the final exam?"

10.1.1 A Recursive Solution

The Towers of Hanoi problem can be solved recursively. As we describe the procedure, we'll also analyze the running time. To that end, let T_n be the minimum number of steps required to solve the n-disk problem. For example, some experimentation shows that $T_1 = 1$ and $T_2 = 3$. The procedure illustrated above shows that T_3 is at most 7, though there might be a solution with fewer steps.

The recursive solution has three stages, which are described below and illustrated in Figure 10.3. For clarity, the largest disk is shaded in the figures.

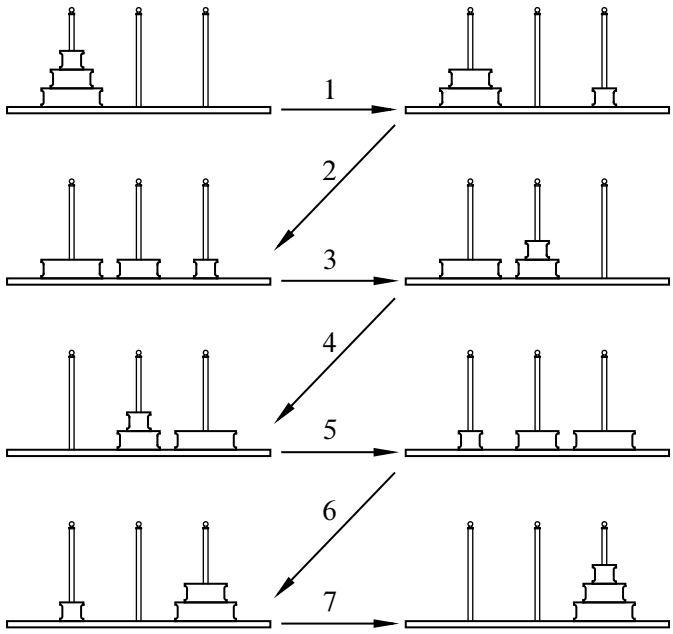


Figure 10.2 The 7-step solution to the Towers of Hanoi problem when there are $n = 3$ disks.

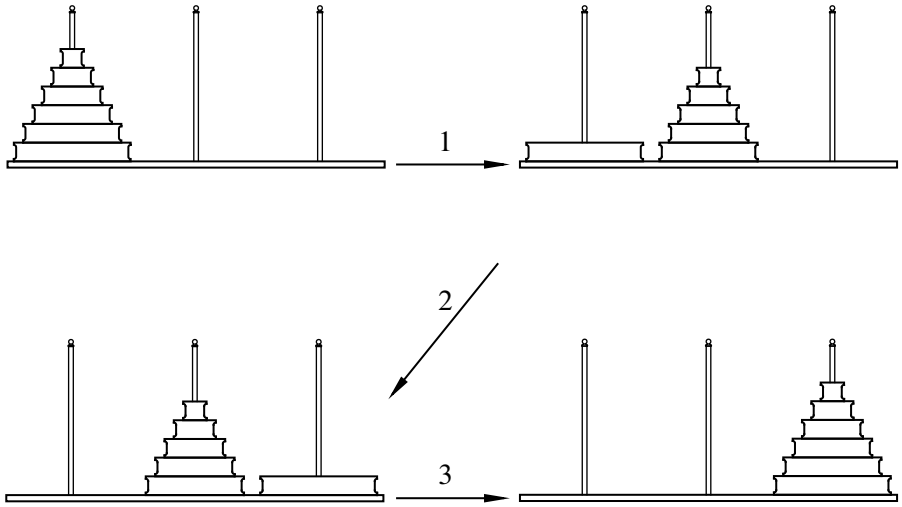


Figure 10.3 A recursive solution to the Towers of Hanoi problem.

Stage 1. Move the top $n-1$ disks from the first post to the second using the solution for $n-1$ disks. This can be done in T_{n-1} steps.

Stage 2. Move the largest disk from the first post to the third post. This takes just 1 step.

Stage 3. Move the $n-1$ disks from the second post to the third post, again using the solution for $n-1$ disks. This can also be done in T_{n-1} steps.

This algorithm shows that T_n, the minimum number of steps required to move n disks to a different post, is at most $T_{n-1}+1+T_{n-1} = 2T_{n-1}+1$. We can use this fact to upper bound the number of operations required to move towers of various heights:

$$T_3 \leq 2 \cdot T_2 + 1 = 7$$
$$T_4 \leq 2 \cdot T_3 + 1 \leq 15$$

Continuing in this way, we could eventually compute an upper bound on T_{64}, the number of steps required to move 64 disks. So this algorithm answers our first question: given sufficient time, the monks can finish their task and end the world. This is a shame. After all that effort, they'd probably want to smack a few high-fives and go out for burgers and ice cream, but nope—world's over.

10.1.2 Finding a Recurrence

We can not yet compute the exact number of steps that the monks need to move the 64 disks, only an upper bound. Perhaps, having pondered the problem since the beginning of time, the monks have devised a better algorithm.

In fact, there is no better algorithm, and here is why. At some step, the monks must move the largest disk from the first post to a different post. For this to happen, the $n-1$ smaller disks must all be stacked out of the way on the only remaining post. Arranging the $n-1$ smaller disks this way requires at least T_{n-1} moves. After the largest disk is moved, at least another T_{n-1} moves are required to pile the $n-1$ smaller disks on top.

This argument shows that the number of steps required is at least $2T_{n-1}+1$. Since we gave an algorithm using exactly that number of steps, we can now write an expression for T_n, the number of moves required to complete the Towers of Hanoi problem with n disks:

$$T_1 = 1$$
$$T_n = 2T_{n-1} + 1 \qquad (\text{for } n \geq 2).$$

This is a typical recurrence. These two lines define a sequence of values, $T_1, T_2, T_3, \ldots$. The first line says that the first number in the sequence, T_1, is equal to 1. The second line defines every other number in the sequence in terms of its predecessor. So we can use this recurrence to compute any number of terms in the sequence:

$$\begin{aligned}
T_1 &= 1 \\
T_2 &= 2 \cdot T_1 + 1 = 3 \\
T_3 &= 2 \cdot T_2 + 1 = 7 \\
T_4 &= 2 \cdot T_3 + 1 = 15 \\
T_5 &= 2 \cdot T_4 + 1 = 31 \\
T_6 &= 2 \cdot T_5 + 1 = 63.
\end{aligned}$$

10.1.3 Solving the Recurrence

We could determine the number of steps to move a 64-disk tower by computing T_7, T_8, and so on up to T_{64}. But that would take a lot of work. It would be nice to have a closed-form expression for T_n, so that we could quickly find the number of steps required for any given number of disks. (For example, we might want to know how much sooner the world would end if the monks melted down one disk to purchase burgers and ice cream *before* the end of the world.)

There are several methods for solving recurrence equations. The simplest is to *guess* the solution and then *verify* that the guess is correct with an induction proof. As a basis for a good guess, let's look for a pattern in the values of T_n computed above: 1, 3, 7, 15, 31, 63. A natural guess is $T_n = 2^n - 1$. But whenever you guess a solution to a recurrence, you should always verify it with a proof, typically by induction. After all, your guess might be wrong. (But why bother to verify in this case? After all, if we're wrong, its not the end of the... no, let's check.)

Claim 10.1.1. $T_n = 2^n - 1$ *satisfies the recurrence:*

$$\begin{aligned}
T_1 &= 1 \\
T_n &= 2T_{n-1} + 1 \qquad \textit{(for } n \geq 2\textit{)}.
\end{aligned}$$

Proof. The proof is by induction on n. The induction hypothesis is that $T_n = 2^n - 1$. This is true for $n = 1$ because $T_1 = 1 = 2^1 - 1$. Now assume that $T_{n-1} = 2^{n-1} - 1$ in order to prove that $T_n = 2^n - 1$, where $n \geq 2$:

$$\begin{aligned}
T_n &= 2T_{n-1} + 1 \\
&= 2(2^{n-1} - 1) + 1 \\
&= 2^n - 1.
\end{aligned}$$

The first equality is the recurrence equation, the second follows from the induction assumption, and the last step is simplification. ■

Such verification proofs are especially tidy because recurrence equations and induction proofs have analogous structures. In particular, the base case relies on the first line of the recurrence, which defines T_1. And the inductive step uses the second line of the recurrence, which defines T_n as a function of preceding terms.

Our guess is verified. So we can now resolve our remaining questions about the 64-disk puzzle. Since $T_{64} = 2^{64} - 1$, the monks must complete more than 18 billion billion steps before the world ends. Better study for the final.

10.1.4 The Upper Bound Trap

When the solution to a recurrence is complicated, one might try to prove that some simpler expression is an upper bound on the solution. For example, the exact solution to the Towers of Hanoi recurrence is $T_n = 2^n - 1$. Let's try to prove the "nicer" upper bound $T_n \leq 2^n$, proceeding exactly as before.

Proof. (Failed attempt.) The proof is by induction on n. The induction hypothesis is that $T_n \leq 2^n$. This is true for $n = 1$ because $T_1 = 1 \leq 2^1$. Now assume that $T_{n-1} \leq 2^{n-1}$ in order to prove that $T_n \leq 2^n$, where $n \geq 2$:

$$\begin{aligned} T_n &= 2T_{n-1} + 1 \\ &\leq 2(2^{n-1}) + 1 \\ &\not\leq 2^n \qquad \leftarrow \text{Uh-oh!} \end{aligned}$$

The first equality is the recurrence relation, the second follows from the induction hypothesis, and the third step is a flaming train wreck. ■

The proof doesn't work! As is so often the case with induction proofs, the argument only goes through with a *stronger* hypothesis. This isn't to say that upper bounding the solution to a recurrence is hopeless, but this is a situation where induction and recurrences do not mix well.

10.1.5 Plug and Chug

Guess-and-verify is a simple and general way to solve recurrence equations. But there is one big drawback: you have to *guess right*. That was not hard for the Towers of Hanoi example. But sometimes the solution to a recurrence has a strange form that is quite difficult to guess. Practice helps, of course, but so can some other methods.

Plug-and-chug is another way to solve recurrences. This is also sometimes called "expansion" or "iteration". As in guess-and-verify, the key step is identifying a pattern. But instead of looking at a sequence of *numbers*, you have to spot a pattern in a sequence of *expressions*, which is sometimes easier. The method consists of three steps, which are described below and illustrated with the Towers of Hanoi example.

Step 1: Plug and Chug Until a Pattern Appears

The first step is to expand the recurrence equation by alternately "plugging" (applying the recurrence) and "chugging" (simplifying the result) until a pattern appears. Be careful: too much simplification can make a pattern harder to spot. The rule to remember—indeed, a rule applicable to the whole of college life—is *chug in moderation*.

$$\begin{aligned}
T_n &= 2T_{n-1} + 1 && \\
&= 2(2T_{n-2} + 1) + 1 && \text{plug} \\
&= 4T_{n-2} + 2 + 1 && \text{chug} \\
&= 4(2T_{n-3} + 1) + 2 + 1 && \text{plug} \\
&= 8T_{n-3} + 4 + 2 + 1 && \text{chug} \\
&= 8(2T_{n-4} + 1) + 4 + 2 + 1 && \text{plug} \\
&= 16T_{n-4} + 8 + 4 + 2 + 1 && \text{chug}
\end{aligned}$$

Above, we started with the recurrence equation. Then we replaced T_{n-1} with $2T_{n-2} + 1$, since the recurrence says the two are equivalent. In the third step, we simplified a little—but not too much! After several similar rounds of plugging and chugging, a pattern is apparent. The following formula seems to hold:

$$\begin{aligned}
T_n &= 2^k T_{n-k} + 2^{k-1} + 2^{k-2} + \ldots + 2^2 + 2^1 + 2^0 \\
&= 2^k T_{n-k} + 2^k - 1
\end{aligned}$$

Once the pattern is clear, simplifying is safe and convenient. In particular, we've collapsed the geometric sum to a closed form on the second line.

Step 2: Verify the Pattern

The next step is to verify the general formula with one more round of plug-and-chug.

$$\begin{aligned} T_n &= 2^k T_{n-k} + 2^k - 1 \\ &= 2^k (2T_{n-(k+1)} + 1) + 2^k - 1 && \text{plug} \\ &= 2^{k+1} T_{n-(k+1)} + 2^{k+1} - 1 && \text{chug} \end{aligned}$$

The final expression on the right is the same as the expression on the first line, except that k is replaced by $k + 1$. Surprisingly, this effectively *proves* that the formula is correct for all k. Here is why: we know the formula holds for $k = 1$, because that's the original recurrence equation. And we've just shown that if the formula holds for some $k \geq 1$, then it also holds for $k + 1$. So the formula holds for all $k \geq 1$ by induction.

Step 3: Write T_n Using Early Terms with Known Values

The last step is to express T_n as a function of early terms whose values are known. Here, choosing $k = n - 1$ expresses T_n in terms of T_1, which is equal to 1. Simplifying gives a closed-form expression for T_n:

$$\begin{aligned} T_n &= 2^{n-1} T_1 + 2^{n-1} - 1 \\ &= 2^{n-1} \cdot 1 + 2^{n-1} - 1 \\ &= 2^n - 1. \end{aligned}$$

We're done! This is the same answer we got from guess-and-verify.

Let's compare guess-and-verify with plug-and-chug. In the guess-and-verify method, we computed several terms at the beginning of the sequence, T_1, T_2, T_3, etc., until a pattern appeared. We generalized to a formula for the nth term, T_n. In contrast, plug-and-chug works backward from the nth term. Specifically, we started with an expression for T_n involving the preceding term, T_{n-1}, and rewrote this using progressively earlier terms, T_{n-2}, T_{n-3}, etc. Eventually, we noticed a pattern, which allowed us to express T_n using the very first term, T_1, whose value we knew. Substituting this value gave a closed-form expression for T_n. So guess-and-verify and plug-and-chug tackle the problem from opposite directions.

Step 3: Write T_n Using Early Terms with Known Values

Finally, we express T_n using early terms whose values are known. Specifically, if we let $k = \log n$, then $T_{n/2^k} = T_1$, which we know is 0:

$$\begin{aligned} T_n &= 2^k T_{n/2^k} + kn - 2^k + 1 \\ &= 2^{\log n} T_{n/2^{\log n}} + n \log n - 2^{\log n} + 1 \\ &= nT_1 + n \log n - n + 1 \\ &= n \log n - n + 1. \end{aligned}$$

We're done! We have a closed-form expression for the maximum number of comparisons used in Merge Sorting a list of n numbers. In retrospect, it is easy to see why guess-and-verify failed: this formula is fairly complicated.

As a check, we can confirm that this formula gives the same values that we computed earlier:

n	T_n	$n \log n - n + 1$
1	0	$1 \log 1 - 1 + 1 = 0$
2	1	$2 \log 2 - 2 + 1 = 1$
4	5	$4 \log 4 - 4 + 1 = 5$
8	17	$8 \log 8 - 8 + 1 = 17$
16	49	$16 \log 16 - 16 + 1 = 49$

As a double-check, we could write out an explicit induction proof. This would be straightforward, because we already worked out the guts of the proof in step 2 of the plug-and-chug procedure.

10.3 Linear Recurrences

So far we've solved recurrences with two techniques: guess-and-verify and plug-and-chug. These methods require spotting a pattern in a sequence of numbers or expressions. In this section and the next, we'll give cookbook solutions for two large classes of recurrences. These methods require no flash of insight; you just follow the recipe and get the answer.

10.3.1 Climbing Stairs

How many different ways are there to climb n stairs, if you can either step up one stair or hop up two? For example, there are five different ways to climb four stairs:

1. step, step, step, step

2. hop, hop
3. hop, step, step
4. step, hop step
5. step, step, hop

Working through this problem will demonstrate the major features of our first cookbook method for solving recurrences. We'll fill in the details of the general solution afterward.

Finding a Recurrence

As special cases, there is 1 way to climb 0 stairs (do nothing) and 1 way to climb 1 stair (step up). In general, an ascent of n stairs consists of either a step followed by an ascent of the remaining $n-1$ stairs or a hop followed by an ascent of $n-2$ stairs. So the total number of ways to climb n stairs is equal to the number of ways to climb $n-1$ plus the number of ways to climb $n-2$. These observations define a recurrence:

$$\begin{aligned} f(0) &= 1 \\ f(1) &= 1 \\ f(n) &= f(n-1) + f(n-2) \qquad \text{for } n \geq 2. \end{aligned}$$

Here, $f(n)$ denotes the number of ways to climb n stairs. Also, we've switched from subscript notation to functional notation, from T_n to f_n. Here the change is cosmetic, but the expressiveness of functions will be useful later.

This is the Fibonacci recurrence, the most famous of all recurrence equations. Fibonacci numbers arise in all sorts of applications and in nature. Fibonacci introduced the numbers in 1202 to study rabbit reproduction. Fibonacci numbers also appear, oddly enough, in the spiral patterns on the faces of sunflowers. And the input numbers that make Euclid's GCD algorithm require the greatest number of steps are consecutive Fibonacci numbers.

Solving the Recurrence

The Fibonacci recurrence belongs to the class of linear recurrences, which are essentially all solvable with a technique that you can learn in an hour. This is somewhat amazing, since the Fibonacci recurrence remained unsolved for almost six centuries!

In general, a *homogeneous linear recurrence* has the form

$$f(n) = a_1 f(n-1) + a_2 f(n-2) + \ldots + a_d f(n-d)$$

where $a_1, a_2, \ldots, a_d$ and d are constants. The *order* of the recurrence is d. Commonly, the value of the function f is also specified at a few points; these are called *boundary conditions*. For example, the Fibonacci recurrence has order $d = 2$ with coefficients $a_1 = a_2 = 1$ and $g(n) = 0$. The boundary conditions are $f(0) = 1$ and $f(1) = 1$. The word "homogeneous" sounds scary, but effectively means "the simpler kind". We'll consider linear recurrences with a more complicated form later.

Let's try to solve the Fibonacci recurrence with the benefit centuries of hindsight. In general, linear recurrences tend to have exponential solutions. So let's guess that

$$f(n) = x^n$$

where x is a parameter introduced to improve our odds of making a correct guess. We'll figure out the best value for x later. To further improve our odds, let's neglect the boundary conditions, $f(0) = 0$ and $f(1) = 1$, for now. Plugging this guess into the recurrence $f(n) = f(n-1) + f(n-2)$ gives

$$x^n = x^{n-1} + x^{n-2}.$$

Dividing both sides by x^{n-2} leaves a quadratic equation:

$$x^2 = x + 1.$$

Solving this equation gives *two* plausible values for the parameter x:

$$x = \frac{1 \pm \sqrt{5}}{2}.$$

This suggests that there are at least two different solutions to the recurrence, neglecting the boundary conditions.

$$f(n) = \left(\frac{1+\sqrt{5}}{2}\right)^n \quad \text{or} \quad f(n) = \left(\frac{1-\sqrt{5}}{2}\right)^n$$

A charming features of homogeneous linear recurrences is that any linear combination of solutions is another solution.

Theorem 10.3.1. *If $f(n)$ and $g(n)$ are both solutions to a homogeneous linear recurrence, then $h(n) = sf(n) + tg(n)$ is also a solution for all $s, t \in \mathbb{R}$.*

Proof.

$$\begin{aligned}
h(n) &= sf(n) + tg(n) \\
&= s\left(a_1 f(n-1) + \ldots + a_d f(n-d)\right) + t\left(a_1 g(n-1) + \ldots + a_d g(n-d)\right) \\
&= a_1(sf(n-1) + tg(n-1)) + \ldots + a_d(sf(n-d) + tg(n-d)) \\
&= a_1 h(n-1) + \ldots + a_d h(n-d)
\end{aligned}$$

The first step uses the definition of the function h, and the second uses the fact that f and g are solutions to the recurrence. In the last two steps, we rearrange terms and use the definition of h again. Since the first expression is equal to the last, h is also a solution to the recurrence. ■

The phenomenon described in this theorem—a linear combination of solutions is another solution—also holds for many differential equations and physical systems. In fact, linear recurrences are so similar to linear differential equations that you can safely snooze through that topic in some future math class.

Returning to the Fibonacci recurrence, this theorem implies that

$$f(n) = s\left(\frac{1+\sqrt{5}}{2}\right)^n + t\left(\frac{1-\sqrt{5}}{2}\right)^n$$

is a solution for all real numbers s and t. The theorem expanded two solutions to a whole spectrum of possibilities! Now, given all these options to choose from, we can find one solution that satisfies the boundary conditions, $f(0) = 1$ and $f(1) = 1$. Each boundary condition puts some constraints on the parameters s and t. In particular, the first boundary condition implies that

$$f(0) = s\left(\frac{1+\sqrt{5}}{2}\right)^0 + t\left(\frac{1-\sqrt{5}}{2}\right)^0 = s + t = 1.$$

Similarly, the second boundary condition implies that

$$f(1) = s\left(\frac{1+\sqrt{5}}{2}\right)^1 + t\left(\frac{1-\sqrt{5}}{2}\right)^1 = 1.$$

Now we have two linear equations in two unknowns. The system is not degenerate, so there is a unique solution:

$$s = \frac{1}{\sqrt{5}} \cdot \frac{1+\sqrt{5}}{2} \qquad t = -\frac{1}{\sqrt{5}} \cdot \frac{1-\sqrt{5}}{2}.$$

These values of s and t identify a solution to the Fibonacci recurrence that also satisfies the boundary conditions:

$$\begin{aligned} f(n) &= \frac{1}{\sqrt{5}} \cdot \frac{1+\sqrt{5}}{2}\left(\frac{1+\sqrt{5}}{2}\right)^n - \frac{1}{\sqrt{5}} \cdot \frac{1-\sqrt{5}}{2}\left(\frac{1-\sqrt{5}}{2}\right)^n \\ &= \frac{1}{\sqrt{5}}\left(\frac{1+\sqrt{5}}{2}\right)^{n+1} - \frac{1}{\sqrt{5}}\left(\frac{1-\sqrt{5}}{2}\right)^{n+1}. \end{aligned}$$

It is easy to see why no one stumbled across this solution for almost six centuries! All Fibonacci numbers are integers, but this expression is full of square roots of five! Amazingly, the square roots always cancel out. This expression really does give the Fibonacci numbers if we plug in $n = 0, 1, 2$, etc.

This closed-form for Fibonacci numbers has some interesting corollaries. The first term tends to infinity because the base of the exponential, $(1 + \sqrt{5})/2 = 1.618\ldots$ is greater than one. This value is often denoted ϕ and called the "golden ratio". The second term tends to zero, because $(1 - \sqrt{5})/2 = -0.618033988\ldots$ has absolute value less than 1. This implies that the nth Fibonacci number is:

$$f(n) = \frac{\phi^{n+1}}{\sqrt{5}} + o(1).$$

Remarkably, this expression involving irrational numbers is actually very close to an integer for all large n—namely, a Fibonacci number! For example:

$$\frac{\phi^{20}}{\sqrt{5}} = 6765.000029\ldots \approx f(19).$$

This also implies that the ratio of consecutive Fibonacci numbers rapidly approaches the golden ratio. For example:

$$\frac{f(20)}{f(19)} = \frac{10946}{6765} = 1.618033998\ldots.$$

10.3.2 Solving Homogeneous Linear Recurrences

The method we used to solve the Fibonacci recurrence can be extended to solve any homogeneous linear recurrence; that is, a recurrence of the form

$$f(n) = a_1 f(n-1) + a_2 f(n-2) + \ldots + a_d f(n-d)$$

where $a_1, a_2, \ldots, a_d$ and d are constants. Substituting the guess $f(n) = x^n$, as with the Fibonacci recurrence, gives

$$x^n = a_1 x^{n-1} + a_2 x^{n-2} + \ldots + a_d x^{n-d}.$$

Dividing by x^{n-d} gives

$$x^d = a_1 x^{d-1} + a_2 x^{d-2} + \ldots + a_{d-1} x + a_d.$$

This is called the *characteristic equation* of the recurrence. The characteristic equation can be read off quickly since the coefficients of the equation are the same as the coefficients of the recurrence.

The solutions to a linear recurrence are defined by the roots of the characteristic equation. Neglecting boundary conditions for the moment:

- If r is a nonrepeated root of the characteristic equation, then r^n is a solution to the recurrence.
- If r is a repeated root with multiplicity k then r^n, nr^n, n^2r^n, $\ldots$, $n^{k-1}r^n$ are all solutions to the recurrence.

Theorem 10.3.1 implies that every linear combination of these solutions is also a solution.

For example, suppose that the characteristic equation of a recurrence has roots s, t, and u twice. These four roots imply four distinct solutions:

$$f(n) = s^n \qquad f(n) = t^n \qquad f(n) = u^n \qquad f(n) = nu^n.$$

Furthermore, every linear combination

$$f(n) = a \cdot s^n + b \cdot t^n + c \cdot u^n + d \cdot nu^n \tag{10.1}$$

is also a solution.

All that remains is to select a solution consistent with the boundary conditions by choosing the constants appropriately. Each boundary condition implies a linear equation involving these constants. So we can determine the constants by solving a system of linear equations. For example, suppose our boundary conditions were $f(0) = 0$, $f(1) = 1$, $f(2) = 4$, and $f(3) = 9$. Then we would obtain four equations in four unknowns:

$$\begin{aligned}
f(0) = 0 \quad &\Rightarrow \quad a \cdot s^0 + b \cdot t^0 + c \cdot u^0 + d \cdot 0u^0 = 0 \\
f(1) = 1 \quad &\Rightarrow \quad a \cdot s^1 + b \cdot t^1 + c \cdot u^1 + d \cdot 1u^1 = 1 \\
f(2) = 4 \quad &\Rightarrow \quad a \cdot s^2 + b \cdot t^2 + c \cdot u^2 + d \cdot 2u^2 = 4 \\
f(3) = 9 \quad &\Rightarrow \quad a \cdot s^3 + b \cdot t^3 + c \cdot u^3 + d \cdot 3u^3 = 9
\end{aligned}$$

This looks nasty, but remember that s, t, and u are just constants. Solving this system gives values for a, b, c, and d that define a solution to the recurrence consistent with the boundary conditions.

10.3.3 Solving General Linear Recurrences

We can now solve all linear homogeneous recurrences, which have the form

$$f(n) = a_1 f(n-1) + a_2 f(n-2) + \ldots + a_d f(n-d).$$

Many recurrences that arise in practice do not quite fit this mold. For example, the Towers of Hanoi problem led to this recurrence:

$$\begin{aligned}
&f(1) = 1 \\
&f(n) = 2f(n-1) + 1 \qquad (\text{for } n \geq 2).
\end{aligned}$$

The problem is the extra $+1$; that is not allowed in a homogeneous linear recurrence. In general, adding an extra function $g(n)$ to the right side of a linear recurrence gives an *inhomogeneous linear recurrence*:

$$f(n) = a_1 f(n-1) + a_2 f(n-2) + \ldots + a_d f(n-d) + g(n).$$

Solving inhomogeneous linear recurrences is neither very different nor very difficult. We can divide the whole job into five steps:

1. Replace $g(n)$ by 0, leaving a homogeneous recurrence. As before, find roots of the characteristic equation.
2. Write down the solution to the homogeneous recurrence, but do not yet use the boundary conditions to determine coefficients. This is called the *homogeneous solution*.
3. Now restore $g(n)$ and find a single solution to the recurrence, ignoring boundary conditions. This is called a *particular solution*. We'll explain how to find a particular solution shortly.
4. Add the homogeneous and particular solutions together to obtain the *general solution*.
5. Now use the boundary conditions to determine constants by the usual method of generating and solving a system of linear equations.

As an example, let's consider a variation of the Towers of Hanoi problem. Suppose that moving a disk takes time proportional to its size. Specifically, moving the smallest disk takes 1 second, the next-smallest takes 2 seconds, and moving the nth disk then requires n seconds instead of 1. So, in this variation, the time to complete the job is given by a recurrence with a $+n$ term instead of a $+1$:

$$\begin{aligned} f(1) &= 1 \\ f(n) &= 2f(n-1) + n \qquad \text{for } n \geq 2. \end{aligned}$$

Clearly, this will take longer, but how much longer? Let's solve the recurrence with the method described above.

In Steps 1 and 2, dropping the $+n$ leaves the homogeneous recurrence $f(n) = 2f(n-1)$. The characteristic equation is $x = 2$. So the homogeneous solution is $f(n) = c2^n$.

In Step 3, we must find a solution to the full recurrence $f(n) = 2f(n-1) + n$, without regard to the boundary condition. Let's guess that there is a solution of the

form $f(n) = an + b$ for some constants a and b. Substituting this guess into the recurrence gives

$$\begin{aligned} an + b &= 2(a(n-1) + b) + n \\ 0 &= (a+1)n + (b - 2a). \end{aligned}$$

The second equation is a simplification of the first. The second equation holds for all n if both $a + 1 = 0$ (which implies $a = -1$) and $b - 2a = 0$ (which implies that $b = -2$). So $f(n) = an + b = -n - 2$ is a particular solution.

In the Step 4, we add the homogeneous and particular solutions to obtain the general solution

$$f(n) = c2^n - n - 2.$$

Finally, in step 5, we use the boundary condition, $f(1) = 1$, determine the value of the constant c:

$$\begin{aligned} f(1) = 1 \quad &\Rightarrow \quad c2^1 - 1 - 2 = 1 \\ &\Rightarrow \quad c = 2. \end{aligned}$$

Therefore, the function $f(n) = 2 \cdot 2^n - n - 2$ solves this variant of the Towers of Hanoi recurrence. For comparison, the solution to the original Towers of Hanoi problem was $2^n - 1$. So if moving disks takes time proportional to their size, then the monks will need about twice as much time to solve the whole puzzle.

10.3.4 How to Guess a Particular Solution

Finding a particular solution can be the hardest part of solving inhomogeneous recurrences. This involves guessing, and you might guess wrong.[1] However, some rules of thumb make this job fairly easy most of the time.

- Generally, look for a particular solution with the same form as the inhomogeneous term $g(n)$.
- If $g(n)$ is a constant, then guess a particular solution $f(n) = c$. If this doesn't work, try polynomials of progressively higher degree: $f(n) = bn + c$, then $f(n) = an^2 + bn + c$, etc.
- More generally, if $g(n)$ is a polynomial, try a polynomial of the same degree, then a polynomial of degree one higher, then two higher, etc. For example, if $g(n) = 6n + 5$, then try $f(n) = bn + c$ and then $f(n) = an^2 + bn + c$.

[1] In Chapter 12, we will show how to solve linear recurrences with generating functions—it's a little more complicated, but it does not require guessing.

- If $g(n)$ is an exponential, such as 3^n, then first guess that $f(n) = c3^n$. Failing that, try $f(n) = bn3^n + c3^n$ and then $an^2 3^n + bn3^n + c3^n$, etc.

The entire process is summarized on the following page.

10.4 Divide-and-Conquer Recurrences

We now have a recipe for solving general linear recurrences. But the Merge Sort recurrence, which we encountered earlier, is not linear:

$$\begin{aligned} T(1) &= 0 \\ T(n) &= 2T(n/2) + n - 1 \qquad (\text{for } n \geq 2). \end{aligned}$$

In particular, $T(n)$ is not a linear combination of a fixed number of immediately preceding terms; rather, $T(n)$ is a function of $T(n/2)$, a term halfway back in the sequence.

Merge Sort is an example of a divide-and-conquer algorithm: it divides the input, "conquers" the pieces, and combines the results. Analysis of such algorithms commonly leads to *divide-and-conquer* recurrences, which have this form:

$$T(n) = \sum_{i=1}^{k} a_i T(b_i n) + g(n)$$

Here $a_1, \ldots a_k$ are positive constants, $b_1, \ldots, b_k$ are constants between 0 and 1, and $g(n)$ is a nonnegative function. For example, setting $a_1 = 2$, $b_1 = 1/2$, and $g(n) = n - 1$ gives the Merge Sort recurrence.

10.4.1 The Akra-Bazzi Formula

The solution to virtually all divide and conquer solutions is given by the amazing *Akra-Bazzi formula*. Quite simply, the asymptotic solution to the general divide-and-conquer recurrence

$$T(n) = \sum_{i=1}^{k} a_i T(b_i n) + g(n)$$

is

$$T(n) = \Theta\left(n^p \left(1 + \int_1^n \frac{g(u)}{u^{p+1}}\, du\right)\right) \tag{10.2}$$

Chapter 9

MFCS 11: The Pigeonhole Principle

Proof. We give a combinatorial proof. Let S be all n-card hands that can be dealt from a deck containing n red cards (numbered $1, \ldots, n$) and $2n$ black cards (numbered $1, \ldots, 2n$). First, note that every $3n$-element set has

$$|S| = \binom{3n}{n}$$

n-element subsets.

From another perspective, the number of hands with exactly r red cards is

$$\binom{n}{r}\binom{2n}{n-r}$$

since there are $\binom{n}{r}$ ways to choose the r red cards and $\binom{2n}{n-r}$ ways to choose the $n-r$ black cards. Since the number of red cards can be anywhere from 0 to n, the total number of n-card hands is:

$$|S| = \sum_{r=0}^{n} \binom{n}{r}\binom{2n}{n-r}.$$

Equating these two expressions for $|S|$ proves the theorem. ■

Combinatorial proofs are almost magical. Theorem 11.9.1 looks pretty scary, but we proved it without any algebraic manipulations at all. The key to constructing a combinatorial proof is choosing the set S properly, which can be tricky. Generally, the simpler side of the equation should provide some guidance. For example, the right side of Theorem 11.9.1 is $\binom{3n}{n}$, which suggests that it will be helpful to choose S to be all n-element subsets of some $3n$-element set.

11.10 The Pigeonhole Principle

Here is an old puzzle:

> A drawer in a dark room contains red socks, green socks, and blue socks. How many socks must you withdraw to be sure that you have a matching pair?

For example, picking out three socks is not enough; you might end up with one red, one green, and one blue. The solution relies on the *Pigeonhole Principle*, which is a friendly name for the contrapositive of the injective case of the Mapping Rule.

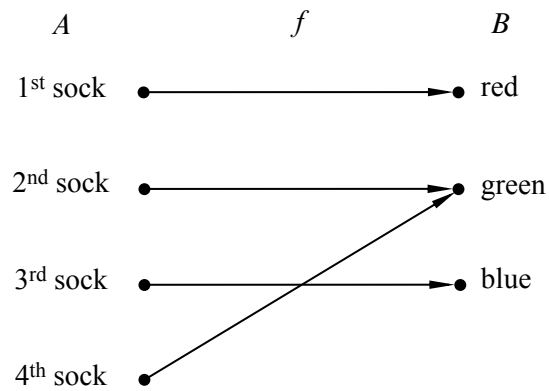


Figure 11.3 One possible mapping of four socks to three colors.

Rule 11.10.1 (Pigeonhole Principle)**.** *If $|X| > |Y|$, then for every total function*[5] *$f : X \to Y$, there exist two different elements of X that are mapped to the same element of Y.*

What this abstract mathematical statement has to do with selecting footwear under poor lighting conditions is maybe not obvious. However, let A be the set of socks you pick out, let B be the set of colors available, and let f map each sock to its color. The Pigeonhole Principle says that if $|A| > |B| = 3$, then at least two elements of A (that is, at least two socks) must be mapped to the same element of B (that is, the same color). Therefore, four socks are enough to ensure a matched pair. For example, one possible mapping of four socks to three colors is shown in Figure 11.3.

Not surprisingly, the pigeonhole principle is often described in terms of pigeons:

> *If there are more pigeons than holes they occupy, then at least two pigeons must be in the same hole.*

In this case, the pigeons form set A, the pigeonholes are set B, and f describes which hole each pigeon flies into.

Mathematicians have come up with many ingenious applications for the pigeonhole principle. If there were a cookbook procedure for generating such arguments, we'd give it to you. Unfortunately, there isn't one. One helpful tip, though: when you try to solve a problem with the pigeonhole principle, the key is to clearly identify three things:

[5]This Mapping Rule applies even if f is a total injective *relation*. Recall that a function is total if $\forall x \in X\ \exists y \in Y.\ f(x) = y$.

1. The set A (the pigeons).
2. The set B (the pigeonholes).
3. The function f (the rule for assigning pigeons to pigeonholes).

11.10.1 Hairs on Heads

There are a number of generalizations of the pigeonhole principle. For example:

Rule 11.10.2 (Generalized Pigeonhole Principle)**.** *If $|X| > k \cdot |Y|$, then every total function $f : X \to Y$ maps at least $k + 1$ different elements of X to the same element of Y.*

For example, if you pick two people at random, surely they are extremely unlikely to have *exactly* the same number of hairs on their heads. However, in the remarkable city of Boston, Massachusetts there are actually *three* people who have exactly the same number of hairs! Of course, there are many bald people in Boston, and they all have zero hairs. But we're talking about non-bald people; say a person is non-bald if they have at least ten thousand hairs on their head.

Boston has about 500,000 non-bald people, and the number of hairs on a person's head is at most 200,000. Let A be the set of non-bald people in Boston, let $B = \{10,000, 10,001, \ldots, 200,000\}$, and let f map a person to the number of hairs on his or her head. Since $|A| > 2|B|$, the Generalized Pigeonhole Principle implies that at least three people have exactly the same number of hairs. We don't know who they are, but we know they exist!

11.10.2 Subsets with the Same Sum

For your reading pleasure, we have displayed ninety 25-digit numbers in Figure 11.4. Are there two different subsets of these 25-digit numbers that have the same sum? For example, maybe the sum of the last ten numbers in the first column is equal to the sum of the first eleven numbers in the second column?

Finding two subsets with the same sum may seem like a silly puzzle, but solving these sorts of problems turns out to be useful in diverse applications such as finding good ways to fit packages into shipping containers and decoding secret messages.

It turns out that it is hard to find different subsets with the same sum, which is why this problem arises in cryptography. But it is easy to prove that two such subsets *exist*. That's where the Pigeonhole Principle comes in.

Let A be the collection of all subsets of the 90 numbers in the list. Now the sum of any subset of numbers is at most $90 \cdot 10^{25}$, since there are only 90 numbers and every 25-digit number is less than 10^{25}. So let B be the set of integers $\{0, 1, \ldots, 90 \cdot 10^{25}\}$, and let f map each subset of numbers (in A) to its sum (in B).

0020480135385502964448038	3171004832173501394113017
5763257331083479647409398	8247331000042995311646021
0489445991866915676240992	3208234421597368647019265
5800949123548989122628663	8496243997123475922766310
1082662032430379651370981	3437254656355157864869113
6042900801199280218026001	8518399140676002660747477
1178480894769706178994993	3574883393058653923711365
6116171789137737896701405	8543691283470191452333763
1253127351683239693851327	3644909946040480189969149
6144868973001582369723512	8675309258374137092461352
1301505129234077811069011	3790044132737084094417246
6247314593851169234746152	8694321112363996867296665
1311567111143866433882194	3870332127437971355322815
6814428944266874963488274	8772321203608477245851154
1470029452721203587686214	4080505804577801451363100
6870852945543886849147881	8791422161722582546341091
1578271047286257499433886	4167283461025702348124920
6914955508120950093732397	9062628024592126283973285
1638243921852176243192354	4235996831123777788211249
6949632451365987152423541	9137845566925526349897794
1763580219131985963102365	4670939445749439042111220
7128211143613619828415650	9153762966803189291934419
1826227795601842231029694	4815379351865384279613427
7173920083651862307925394	9270880194077636406984249
1843971862675102037201420	4837052948212922604442190
7215654874211755676220587	9324301480722103490379204
2396951193722134526177237	5106389423855018550671530
7256932847164391040233050	9436090832146695147140581
2781394568268599801096354	5142368192004769218069910
7332822657075235431620317	9475308159734538249013238
2796605196713610405408019	5181234096130144084041856
7426441829541573444964139	9492376623917486974923202
2931016394761975263190347	5198267398125617994391348
7632198126531809327186321	9511972558779880288252979
2933458058294405155197296	5317592940316231219758372
7712154432211912882310511	9602413424619187112552264
3075514410490975920315348	5384358126771794128356947
7858918664240262356610010	9631217114906129219461111
8149436716871371161932035	3157693105325111284321993
3111474985252793452860017	5439211712248901995423441
7898156786763212963178679	9908189853102753335981319
3145621587936120118438701	5610379826092838192760458
8147591017037573337848616	9913237476341764299813987
3148901255628881103198549	5632317555465228677676044
5692168374637019617423712	8176063831682536571306791

Figure 11.4 Ninety 25-digit numbers. Can you find two different subsets of these numbers that have the same sum?

We proved that an n-element set has 2^n different subsets in Section 11.2. Therefore:

$$|A| = 2^{90} \geq 1.237 \times 10^{27}$$

On the other hand:

$$|B| = 90 \cdot 10^{25} + 1 \leq 0.901 \times 10^{27}.$$

Both quantities are enormous, but $|A|$ is a bit greater than $|B|$. This means that f maps at least two elements of A to the same element of B. In other words, by the Pigeonhole Principle, two different subsets must have the same sum!

Notice that this proof gives no indication *which* two sets of numbers have the same sum. This frustrating variety of argument is called a *nonconstructive proof*. To see if was possible to actually *find* two different subsets of the ninety 25-digit numbers with the same sum, we offered a $100 prize to the first student who did it. We didn't expect to have to pay off this bet, but we underestimated the ingenuity and initiative of the students. One computer science major wrote a program that cleverly searched only among a reasonably small set of "plausible" sets, sorted them by their sums, and actually found a couple with the same sum. He won the prize. A few days later, a math major figured out how to reformulate the sum problem as a "lattice basis reduction" problem; then he found a software package implementing an efficient basis reduction procedure, and using it, he very quickly found lots of pairs of subsets with the same sum. He didn't win the prize, but he got a standing ovation from the class—staff included.

11.11 A Magic Trick

There is a Magician and an Assistant. The Assistant goes into the audience with a deck of 52 cards while the Magician looks away.

Five audience members each select one card from the deck. The Assistant then gathers up the five cards and holds up four of them so the Magician can see them. The Magician concentrates for a short time and then correctly names the secret, fifth card!

Since we don't really believe the Magician can read minds, we know the Assistant has somehow communicated the secret card to the Magician. Since real Magicians and Assistants are not to be trusted, we can expect that the Assistant would illegitimately signal the Magician with coded phrases or body language, but they don't have to cheat in this way. In fact, the Magician and Assistant could be

Chapter 10

MFCS 5: Graph Theory

5 Graph Theory

Informally, a graph is a bunch of dots and lines where the lines connect some pairs of dots. An example is shown in Figure 5.1. The dots are called *nodes* (or *vertices*) and the lines are called *edges*.

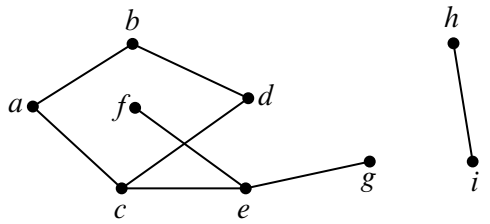


Figure 5.1 An example of a graph with 9 nodes and 8 edges.

Graphs are ubiquitous in computer science because they provide a handy way to represent a relationship between pairs of objects. The objects represent items of interest such as programs, people, cities, or web pages, and we place an edge between a pair of nodes if they are related in a certain way. For example, an edge between a pair of people might indicate that they like (or, in alternate scenarios, that they don't like) each other. An edge between a pair of courses might indicate that one needs to be taken before the other.

In this chapter, we will focus our attention on simple graphs where the relationship denoted by an edge is symmetric. Afterward, in Chapter 6, we consider the situation where the edge denotes a one-way relationship, for example, where one web page points to the other.[1]

5.1 Definitions

5.1.1 Simple Graphs

Definition 5.1.1. A *simple graph* G consists of a nonempty set V, called the *vertices* (aka *nodes*[2]) of G, and a set E of two-element subsets of V. The members of E are called the *edges* of G, and we write $G = (V, E)$.

[1] Two Stanford students analyzed such a graph to become multibillionaires. So, pay attention to graph theory, and who knows what might happen!

[2] We will use the terms vertex and node interchangeably.

The vertices correspond to the dots in Figure 5.1, and the edges correspond to the lines. The graph in Figure 5.1 is expressed mathematically as $G = (V, E)$, where:

$$V = \{a, b, c, d, e, f, g, h, i\}$$
$$E = \{\{a, b\}, \{a, c\}, \{b, d\}, \{c, d\}, \{c, e\}, \{e, f\}, \{e, g\}, \{h, i\}\}.$$

Note that $\{a, b\}$ and $\{b, a\}$ are different descriptions of the same edge, since sets are unordered. In this case, the graph $G = (V, E)$ has 9 nodes and 8 edges.

Definition 5.1.2. Two vertices in a simple graph are said to be *adjacent* if they are joined by an edge, and an edge is said to be *incident* to the vertices it joins. The number of edges incident to a vertex v is called the *degree* of the vertex and is denoted by $\deg(v)$; equivalently, the degree of a vertex is equals the number of vertices adjacent to it.

For example, in the simple graph shown in Figure 5.1, vertex a is adjacent to b and b is adjacent to d, and the edge $\{a, c\}$ is incident to vertices a and c. Vertex h has degree 1, d has degree 2, and $\deg(e) = 3$. It is possible for a vertex to have degree 0, in which case it is not adjacent to any other vertices. A simple graph does not need to have any edges at all —in which case, the degree of every vertex is zero and $|E| = 0$[3] —but it does need to have at least one vertex, that is, $|V| \geq 1$.

Note that simple graphs do *not* have any *self-loops* (that is, an edge of the form $\{a, a\}$) since an edge is defined to be a set of *two* vertices. In addition, there is at most one edge between any pair of vertices in a simple graph. In other words, a simple graph does not contain *multiedges* or *multiple edges*. That is because E is a set. Lastly, and most importantly, simple graphs do not contain *directed edges* (that is, edges of the form (a, b) instead of $\{a, b\}$).

There's no harm in relaxing these conditions, and some authors do, but we don't need self-loops, multiple edges between the same two vertices, or graphs with no vertices, and it's simpler not to have them around. We will consider graphs with directed edges (called *directed graphs* or *digraphs*) at length in Chapter 6. Since we'll only be considering simple graphs in this chapter, we'll just call them "graphs" from now on.

5.1.2 Some Common Graphs

Some graphs come up so frequently that they have names. The *complete graph* on n vertices, denoted K_n, has an edge between every two vertices, for a total of $n(n-1)/2$ edges. For example, K_5 is shown in Figure 5.2.

The *empty graph* has no edges at all. For example, the empty graph with 5 nodes is shown in Figure 5.3.

[3]The *cardinality*, $|E|$, of the set E is the number of elements in E.

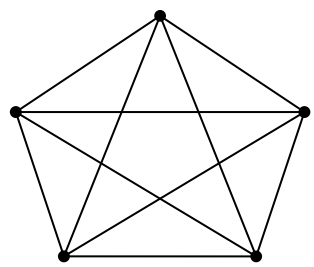


Figure 5.2 The complete graph on 5 nodes, K_5.

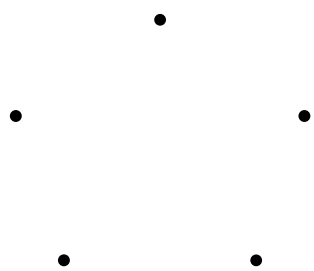


Figure 5.3 The empty graph with 5 nodes.

The n-node graph containing $n - 1$ edges in sequence is known as the *line graph* L_n. More formally, $L_n = (V, E)$ where

$$V = \{v_1, v_2, \ldots, v_n\}$$

and

$$E = \{\{v_1, v_2\}, \{v_2, v_3\}, \ldots, \{v_{n-1}, v_n\}\}$$

For example, L_5 is displayed in Figure 5.4.

If we add the edge $\{v_n, v_1\}$ to the line graph L_n, we get the graph C_n consisting of a simple cycle. For example, C_5 is illustrated in Figure 5.5.

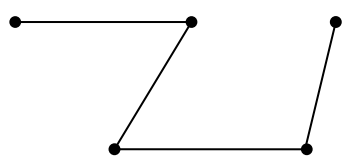


Figure 5.4 The 5-node line graph L_5.

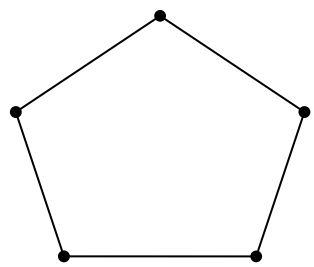


Figure 5.5 The 5-node cycle graph C_5.

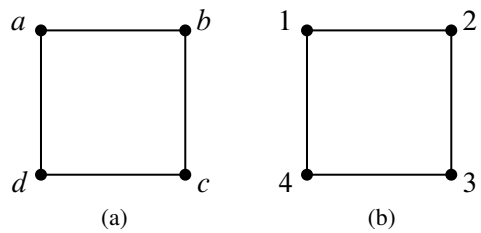


Figure 5.6 Two graphs that are isomorphic to C_4.

5.1.3 Isomorphism

Two graphs that look the same might actually be different in a formal sense. For example, the two graphs in Figure 5.6 are both simple cycles with 4 vertices, but one graph has vertex set $\{a, b, c, d\}$ while the other has vertex set $\{1, 2, 3, 4\}$. Strictly speaking, these graphs are different mathematical objects, but this is a frustrating distinction since the graphs *look the same*!

Fortunately, we can neatly capture the idea of "looks the same" through the notion of graph isomorphism.

Definition 5.1.3. If $G_1 = (V_1, E_1)$ and $G_2 = (V_2, E_2)$ are two graphs, then we say that G_1 is *isomorphic* to G_2 iff there exists a *bijection*[4] $f : V_1 \to V_2$ such that for every pair of vertices $u, v \in V_1$:

$$\{u, v\} \in E_1 \quad \text{iff} \quad \{f(u), f(v)\} \in E_2.$$

The function f is called an *isomorphism* between G_1 and G_2.

In other words, two graphs are isomorphic if they are the same up to a relabeling of their vertices. For example, here is an isomorphism between vertices in the two

[4]A bijection $f : V_1 \to V_2$ is a function that associates every node in V_1 with a unique node in V_2 and vice-versa. We will study bijections more deeply in Part III.

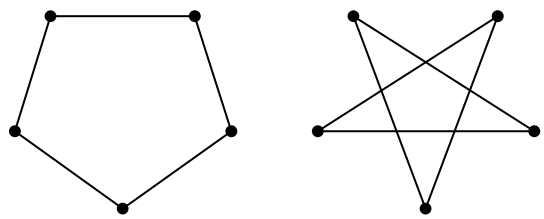


Figure 5.7 Two ways of drawing C_5.

graphs shown in Figure 5.6:

a corresponds to 1	b corresponds to 2
d corresponds to 4	c corresponds to 3.

You can check that there is an edge between two vertices in the graph on the left if and only if there is an edge between the two corresponding vertices in the graph on the right.

Two isomorphic graphs may be drawn very differently. For example, we have shown two different ways of drawing C_5 in Figure 5.7.

Isomorphism preserves the connection properties of a graph, abstracting out what the vertices are called, what they are made out of, or where they appear in a drawing of the graph. More precisely, a property of a graph is said to be *preserved under isomorphism* if whenever G has that property, every graph isomorphic to G also has that property. For example, isomorphic graphs must have the same number of vertices. What's more, if f is a graph isomorphism that maps a vertex, v, of one graph to the vertex, $f(v)$, of an isomorphic graph, then by definition of isomorphism, every vertex adjacent to v in the first graph will be mapped by f to a vertex adjacent to $f(v)$ in the isomorphic graph. This means that v and $f(v)$ will have the same degree. So if one graph has a vertex of degree 4 and another does not, then they can't be isomorphic. In fact, they can't be isomorphic if the number of degree 4 vertices in each of the graphs is not the same.

Looking for preserved properties can make it easy to determine that two graphs are not isomorphic, or to actually find an isomorphism between them if there is one. In practice, it's frequently easy to decide whether two graphs are isomorphic. However, no one has yet found a *general* procedure for determining whether two graphs are isomorphic that is *guaranteed* to run in polynomial time[5] in $|V|$.

Having such a procedure would be useful. For example, it would make it easy to search for a particular molecule in a database given the molecular bonds. On

[5] *I.e.*, in an amount of time that is upper-bounded by $|V|^c$ where c is a fixed number independent of $|V|$.

the other hand, knowing there is no such efficient procedure would also be valuable: secure protocols for encryption and remote authentication can be built on the hypothesis that graph isomorphism is computationally exhausting.

5.1.4 Subgraphs

Definition 5.1.4. A graph $G_1 = (V_1, E_1)$ is said to be a *subgraph* of a graph $G_2 = (V_2, E_2)$ if $V_1 \subseteq V_2$ and $E_1 \subseteq E_2$.

For example, the empty graph on n nodes is a subgraph of L_n, L_n is a subgraph of C_n, and C_n is a subgraph of K_n. Also, the graph $G = (V, E)$ where

$$V = \{g, h, i\} \quad \text{and} \quad E = \{\{h, i\}\}$$

is a subgraph of the graph in Figure 5.1. On the other hand, any graph containing an edge $\{g, h\}$ would not be a subgraph of the graph in Figure 5.1 because the graph in Figure 5.1 does not contain this edge.

Note that since a subgraph is itself a graph, the endpoints of any edge in a subgraph must also be in the subgraph. In other words if $G' = (V', E')$ is a subgraph of some graph G, and $\{v_i, v_j\} \in E'$, then it must be the case that $v_i \in V'$ and $v_j \in V'$.

5.1.5 Weighted Graphs

Sometimes, we will use edges to denote a connection between a pair of nodes where the connection has a *capacity* or *weight*. For example, we might be interested in the capacity of an Internet fiber between a pair of computers, the resistance of a wire between a pair of terminals, the tension of a spring connecting a pair of devices in a dynamical system, the tension of a bond between a pair of atoms in a molecule, or the distance of a highway between a pair of cities.

In such cases, it is useful to represent the system with an *edge-weighted* graph (aka a *weighted graph*). A weighted graph is the same as a simple graph except that we associate a real number (that is, the weight) with each edge in the graph. Mathematically speaking, a weighted graph consists of a graph $G = (V, E)$ and a weight function $w : E \to \mathbb{R}$. For example, Figure 5.8 shows a weighted graph where the weight of edge $\{a, b\}$ is 5.

5.1.6 Adjacency Matrices

There are many ways to represent a graph. We have already seen two ways: you can draw it, as in Figure 5.8 for example, or you can represent it with sets —as in $G = (V, E)$. Another common representation is with an adjacency matrix.

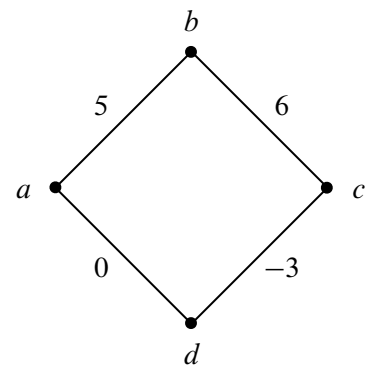


Figure 5.8 A 4-node weighted graph where the edge $\{a, b\}$ has weight 5.

$$\begin{pmatrix} 0 & 1 & 0 & 1 \\ 1 & 0 & 1 & 0 \\ 0 & 1 & 0 & 1 \\ 1 & 0 & 1 & 0 \end{pmatrix} \qquad \begin{pmatrix} 0 & 5 & 0 & 0 \\ 5 & 0 & 6 & 0 \\ 0 & 6 & 0 & -3 \\ 0 & 0 & -3 & 0 \end{pmatrix}$$

(a) (b)

Figure 5.9 Examples of adjacency matrices. (a) shows the adjacency matrix for the graph in Figure 5.6(a) and (b) shows the adjacency matrix for the weighted graph in Figure 5.8. In each case, we set $v_1 = a$, $v_2 = b$, $v_3 = c$, and $v_4 = d$ to construct the matrix.

Definition 5.1.5. Given an n-node graph $G = (V, E)$ where $V = \{v_1, v_2, \ldots, v_n\}$, the *adjacency matrix* for G is the $n \times n$ matrix $A_G = \{a_{ij}\}$ where

$$a_{ij} = \begin{cases} 1 & \text{if } \{v_i, v_j\} \in E \\ 0 & \text{otherwise.} \end{cases}$$

If G is a weighted graph with edge weights given by $w : E \to \mathbb{R}$, then the adjacency matrix for G is $A_G = \{a_{ij}\}$ where

$$a_{ij} = \begin{cases} w(\{v_i, v_j\}) & \text{if } \{v_i, v_j\} \in E \\ 0 & \text{otherwise.} \end{cases}$$

For example, Figure 5.9 displays the adjacency matrices for the graphs shown in Figures 5.6(a) and 5.8 where $v_1 = a$, $v_2 = b$, $v_3 = c$, and $v_4 = d$.

One way to avoid an accidental build-up error is to use a "shrink down, grow back" process in the inductive step; that is, start with a size $n + 1$ graph, remove a vertex (or edge), apply the inductive hypothesis $P(n)$ to the smaller graph, and then add back the vertex (or edge) and argue that $P(n + 1)$ holds. Let's see what would have happened if we'd tried to prove the claim above by this method:

Revised inductive step: We must show that $P(n)$ implies $P(n + 1)$ for all $n \geq 1$. Consider an $(n + 1)$-vertex graph G in which every vertex has degree at least 1. Remove an arbitrary vertex v, leaving an n-vertex graph G' in which every vertex has degree... uh oh!

The reduced graph G' might contain a vertex of degree 0, making the inductive hypothesis $P(n)$ inapplicable! We are stuck—and properly so, since the claim is false!

Always use shrink-down, grow-back arguments and you'll never fall into this trap.

5.6 Around and Around We Go

5.6.1 Cycles and Closed Walks

Definition 5.6.1. A *closed walk*[15] in a graph G is a sequence of vertices

$$v_0, v_1, \ldots, v_k$$

and edges

$$\{v_0, v_1\}, \{v_1, v_2\}, \ldots, \{v_{k-1}, v_k\}$$

where v_0 is the same node as v_k and $\{v_i, v_{i+1}\}$ is an edge of G for all i where $0 \leq i < k$. The *length* of the closed walk is k. A closed walk is said to be a *cycle* if $k \geq 3$ and $v_0, v_1, \ldots, v_{k-1}$ are all different.

For example, b, c, d, e, c, b is a closed walk of length 5 in the graph shown in Figure 5.18. It is not a cycle since it contains node c twice. On the other hand, c, d, e, c is a cycle of length 3 in this graph since every node appears just once.

There are many ways to represent the same closed walk or cycle. For example, b, c, d, e, c, b is the same as c, d, e, c, b, c (just starting at node c instead of node b) and the same as b, c, e, d, c, b (just reversing the direction).

[15] Some texts use the word *cycle* for our definition of closed walk and *simple cycle* for our definition of cycle.

Cycles are similar to paths, except that the last node is the first node and the notion of first and last does not matter. Indeed, there are many possible vertex orders that can be used to describe cycles and closed walks, whereas walks and paths have a prescribed beginning, end, and ordering.

5.6.2 Odd Cycles and 2-Colorability

We have already seen that determining the chromatic number of a graph is a challenging problem. There is a special case where this problem is very easy; namely, the case where every cycle in the graph has even length. In this case, the graph is 2-colorable! Of course, this is optimal if the graph has any edges at all. More generally, we will prove

Theorem 5.6.2. *The following properties of a graph are equivalent (that is, if the graph has any one of the properties, then it has all of the properties):*

1. *The graph is bipartite.*
2. *The graph is 2-colorable.*
3. *The graph does not contain any cycles with odd length.*
4. *The graph does not contain any closed walks with odd length.*

Proof. We will show that property 1 IMPLIES property 2, property 2 IMPLIES property 3, property 3 IMPLIES property 4, and property 4 IMPLIES property 1. This will show that all four properties are equivalent by repeated application of Rule 2.1.2 in Section 2.1.2.

1 IMPLIES 2 Assume that $G = (V, E)$ is a bipartite graph. Then V can be partitioned into two sets L and R so that no edge connects a pair of nodes in L nor a pair of nodes in R. Hence, we can use one color for all the nodes in L and a second color for all the nodes in R. Hence $\chi(G) = 2$.

2 IMPLIES 3 Let $G = (V, E)$ be a 2-colorable graph and

$$C ::= v_0, v_1, \ldots, v_k$$

be any cycle in G. Consider any 2-coloring for the nodes of G. Since $\{v_i, v_{i+1}\} \in E$, v_i and v_{i+1} must be differently colored for $0 \leq i < k$. Hence v_0, v_2, v_4, $\ldots$, have one color and v_1, v_3, v_5, $\ldots$, have the other color. Since C is a cycle, v_k is the same node as v_0, which means they must have the same color, and so k must be an even number. This means that C has even length.

3 IMPLIES 4 The proof is by contradiction. Assume for the purposes of contradiction that G is a graph that does not contain any cycles with odd length (that is, G satisfies Property 3) but that G *does* contain a closed walk with odd length (that is, G does not satisfy Property 4).

Let

$$w ::= v_0, v_1, v_2, \ldots, v_k$$

be the *shortest* closed walk with odd length in G. Since G has no odd-length cycles, w cannot be a cycle. Hence $v_i = v_j$ for some $0 \leq i < j < k$. This means that w is the union of two closed walks:

$$v_0, v_1, \ldots, v_i, v_{j+1}, v_{j+2}, \ldots, v_k$$

and

$$v_i, v_{i+1}, \ldots, v_j.$$

Since w has odd length, one of these two closed walks must also have odd length and be shorter than w. This contradicts the minimality of w. Hence 3 IMPLIES 4.

4 IMPLIES 1 Once again, the proof is by contradiction. Assume for the purposes of contradictin that G is a graph without any closed walks with odd length (that is, G satisfies Property 4) but that G is *not* bipartite (that is, G does not satisfy Property 1).

Since G is not bipartite, it must contain a connected component $G' = (V', E')$ that is not bipartite. Let v be some node in V'. For every node $u \in V'$, define

$$\begin{aligned} \text{dist}(u) ::= &\text{ the length of the shortest path from } u \text{ to } v \text{ in } G'. \\ &\text{If } u = v, \text{ the distance is zero.} \end{aligned}$$

Partition V' into sets L and R so that

$$\begin{aligned} L &= \{\, u \mid \text{dist}(u) \text{ is even} \,\}, \\ R &= \{\, u \mid \text{dist}(u) \text{ is odd} \,\}. \end{aligned}$$

Since G' is not bipartite, there must be a pair of adjacent nodes u_1 and u_2 that are both in L or both in R. Let e denote the edge incident to u_1 and u_2.

Let P_i denote a shortest path in G' from u_i to v for $i = 1, 2$. Because u_1 and u_2 are both in L or both in R, it must be the case that P_1 and P_2 both have even length or they both have odd length. In either case, the union of P_1, P_2, and e forms a closed walk with odd length, which is a contradiction. Hence 4 IMPLIES 1. ■

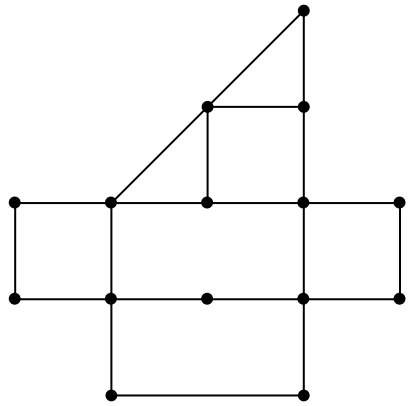


Figure 5.23 A possible floor plan for a museum. Can you find a walk that traverses every edge exactly once?

Theorem 5.6.2 turns out to be useful since bipartite graphs come up fairly often in practice. We'll see examples when we talk about planar graphs in Section 5.8 and when we talk about packet routing in communication networks in Chapter 6.

5.6.3 Euler Tours

Can you walk every hallway in the Museum of Fine Arts *exactly once*? If we represent hallways and intersections with edges and vertices, then this reduces to a question about graphs. For example, could you visit every hallway exactly once in a museum with the floor plan in Figure 5.23?

The entire field of graph theory began when Euler asked whether the seven bridges of Königsberg could all be traversed exactly once—essentially the same question we asked about the Museum of Fine Arts. In his honor, an *Euler walk* is a defined to be a walk that traverses every edge in a graph exactly once. Similarly, an *Euler tour* is an Euler walk that starts and finishes at the same vertex. Graphs with Euler tours and Euler walks both have simple characterizations.

Theorem 5.6.3. *A connected graph has an Euler tour if and only if every vertex has even degree.*

Proof. We first show that if a graph has an Euler tour, then every vertex has even degree. Assume that a graph $G = (V, E)$ has an Euler tour $v_0, v_1, \ldots, v_k$ where $v_k = v_0$. Since every edge is traversed once in the tour, $k = |E|$ and the degree of a node u in G is the number of times that node appears in the sequence $v_0, v_1, \ldots, v_{k-1}$ times two. We multiply by two since if $u = v_i$ for some i where $0 < i < k$, then both $\{v_{i-1}, v_i\}$ and $\{v_i, v_{i+1}\}$ are edges incident to u in G. If $u = v_0 = v_k$,

then both $\{v_{k-1}, v_k\}$ and $\{v_0, v_1\}$ are edges incident to u in G. Hence, the degree of every node is even.

We next show that if the degree of every node is even in a graph $G = (V, E)$, then there is an Euler tour. Let

$$W ::= v_0, v_1, \ldots, v_k$$

be the longest walk in G that traverses *no edge more than once*[16]. W must traverse every edge incident to v_k; otherwise the walk could be extended and W would not be the longest walk that traverses all edges at most once. Moreover, it must be that $v_k = v_0$ and that W is a closed walk, since otherwise v_k would have odd degree in W (and hence in G), which is not possible by assumption.

We conclude the argument with a proof by contradiction. Suppose that W is not an Euler tour. Because G is a connected graph, we can find an edge not in W but incident to some vertex in W. Call this edge $\{u, v_i\}$. But then we can construct a walk W' that is longer than W but that still uses no edge more than once:

$$W' ::= u, v_i, v_{i+1}, \ldots, v_k, v_1, v_2, \ldots, v_i.$$

This contradicts the definition of W, so W must be an Euler tour after all. ■

It is not difficult to extend Theorem 5.6.3 to prove that a connected graph G has an Euler walk if and only if precisely 0 or 2 nodes in G have odd degree. Hence, we can conclude that the graph shown in Figure 5.23 has an Euler walk but not an Euler tour since the graph has precisely two nodes with odd degree.

Although the proof of Theorem 5.6.3 does not explicitly define a method for finding an Euler tour when one exists, it is not hard to modify the proof to produce such a method. The idea is to grow a tour by continually splicing in closed walks until all the edges are consumed.

5.6.4 Hamiltonian Cycles

Hamiltonian cycles are the unruly cousins of Euler tours.

Definition 5.6.4. A *Hamiltonian cycle* in a graph G is a cycle that visits every *node* in G exactly once. Similarly, a *Hamiltonian* path is a path in G that visits every node exactly once.

[16]Did you notice that we are using a variation of the Well Ordering Principle here when we implicitly assume that a longest walk exists? This is ok since the length of a walk where no edge is used more than once is at most $|E|$.

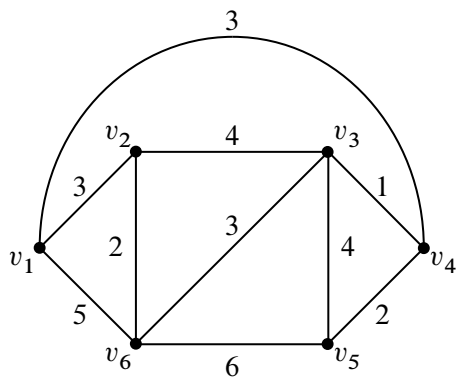


Figure 5.24 A weighted graph. Can you find a cycle with weight 15 that visits every node exactly once?

Although Hamiltonian cycles sound similar to Euler tours—one visits every node once while the other visits every edge once—finding a Hamiltonian cycle can be a lot harder than finding an Euler tour. The same is true for Hamiltonian paths. This is because no one has discovered a simple characterization of all graphs with a Hamiltonian cycle. In fact, determining whether a graph has a Hamiltonian cycle is the same category of problem as the SAT problem of Section 1.5 and the coloring problem in Section 5.3; you get a million dollars for finding an efficient way to determine when a graph has a Hamiltonian cycle—or proving that no procedure works efficiently on all graphs.

5.6.5 The Traveling Salesperson Problem

As if the problem of finding a Hamiltonian cycle is not hard enough, when the graph is weighted, we often want to find a Hamiltonian cycle that has least possible weight. This is a very famous optimization problem known as the Traveling Salesperson Problem.

Definition 5.6.5. Given a weighted graph G, the *weight* of a cycle in G is defined as the sum of the weights of the edges in the cycle.

For example, consider the graph shown in Figure 5.24 and suppose that you would like to visit every node once and finish at the node where you started. Can you find way to do this by traversing a cycle with weight 15?

Needless to say, if you can figure out a fast procedure that finds the optimal cycle for the traveling salesperson, let us know so that we can win a million dollars.

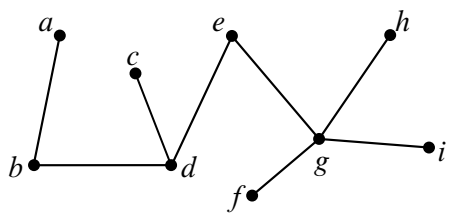


Figure 5.25 A 9-node tree.

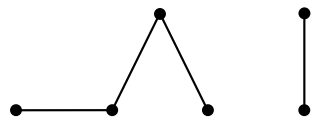


Figure 5.26 A 6-node forest consisting of 2 component trees. Note that this 6-node graph is not itself a tree since it is not connected.

5.7 Trees

As we have just seen, finding good cycles in a graph can be trickier than you might first think. But what if a graph has no cycles at all? Sounds pretty dull. But graphs without cycles (called *acyclic graphs*) are probably the most important graphs of all when it comes to computer science.

5.7.1 Definitions

Definition 5.7.1. A connected acyclic graph is called a *tree*.

For example, Figure 5.25 shows an example of a 9-node tree.

The graph shown in Figure 5.26 is not a tree since it is not connected, but it is a forest. That's because, of course, it consists of a collection of trees.

Definition 5.7.2. If every connected component of a graph G is a tree, then G is a *forest*.

One of the first things you will notice about trees is that they tend to have a lot of nodes with degree one. Such nodes are called *leaves*.

Definition 5.7.3. A *leaf* is a node with degree 1 in a tree (or forest).

For example, the tree in Figure 5.25 has 5 leaves and the forest in Figure 5.26 has 4 leaves.

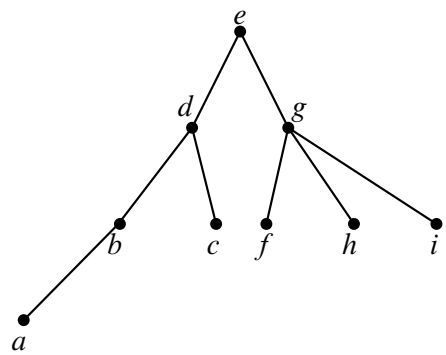


Figure 5.27 The tree from Figure 5.25 redrawn in a leveled fashion, with node E as the root.

Trees are a fundamental data structure in computer science. For example, information is often stored in tree-like data structures and the execution of many recursive programs can be modeled as the traversal of a tree. In such cases, it is often useful to draw the tree in a leveled fashion where the node in the top level is identified as the *root*, and where every edge joins a *parent* to a *child*. For example, we have redrawn the tree from Figure 5.25 in this fashion in Figure 5.27. In this example, node d is a child of node e and a parent of nodes b and c.

In the special case of *ordered binary trees*, every node is the parent of at most 2 children and the children are labeled as being a left-child or a right-child.

5.7.2 Properties

Trees have many unique properties. We have listed some of them in the following theorem.

Theorem 5.7.4. *Every tree has the following properties:*

1. *Any connected subgraph is a tree.*
2. *There is a unique simple path between every pair of vertices.*
3. *Adding an edge between nonadjacent nodes in a tree creates a graph with a cycle.*
4. *Removing any edge disconnects the graph.*
5. *If the tree has at least two vertices, then it has at least two leaves.*
6. *The number of vertices in a tree is one larger than the number of edges.*

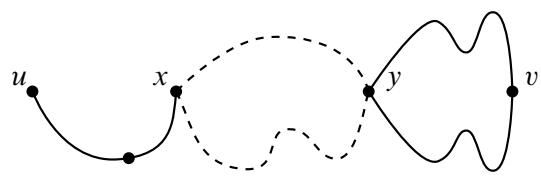


Figure 5.28 If there are two paths between u and v, the graph must contain a cycle.

Proof.

1. A cycle in a subgraph is also a cycle in the whole graph, so any subgraph of an acyclic graph must also be acyclic. If the subgraph is also connected, then by definition, it is a tree.

2. Since a tree is connected, there is at least one path between every pair of vertices. Suppose for the purposes of contradiction, that there are two different paths between some pair of vertices u and v. Beginning at u, let x be the first vertex where the paths diverge, and let y be the next vertex they share. (For example, see Figure 5.28.) Then there are two paths from x to y with no common edges, which defines a cycle. This is a contradiction, since trees are acyclic. Therefore, there is exactly one path between every pair of vertices.

3. An additional edge $\{u, v\}$ together with the unique path between u and v forms a cycle.

4. Suppose that we remove edge $\{u, v\}$. Since the tree contained a unique path between u and v, that path must have been $\{u, v\}$. Therefore, when that edge is removed, no path remains, and so the graph is not connected.

5. Let $v_1, \ldots, v_m$ be the sequence of vertices on a longest path in the tree. Then $m \geq 2$, since a tree with two vertices must contain at least one edge. There cannot be an edge $\{v_1, v_i\}$ for $2 < i \leq m$; otherwise, vertices $v_1, \ldots, v_i$ would from a cycle. Furthermore, there cannot be an edge $\{u, v_1\}$ where u is not on the path; otherwise, we could make the path longer. Therefore, the only edge incident to v_1 is $\{v_1, v_2\}$, which means that v_1 is a leaf. By a symmetric argument, v_m is a second leaf.

6. We use induction on the proposition $P(n) ::=$ there are $n - 1$ edges in any n-vertex tree.

   Base Case ($n = 1$): $P(1)$ is true since a tree with 1 node has 0 edges and $1 - 1 = 0$.

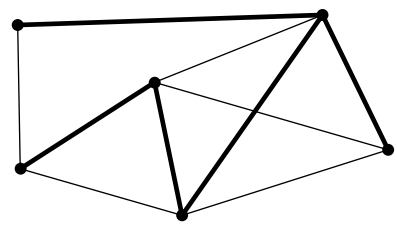


Figure 5.29 A graph where the edges of a spanning tree have been thickened.

Inductive step: Now suppose that $P(n)$ is true and consider an $(n+1)$-vertex tree, T. Let v be a leaf of the tree. You can verify that deleting a vertex of degree 1 (and its incident edge) from any connected graph leaves a connected subgraph. So by part 1 of Theorem 5.7.4, deleting v and its incident edge gives a smaller tree, and this smaller tree has $n-1$ edges by induction. If we re-attach the vertex v and its incident edge, then we find that T has $n = (n+1) - 1$ edges. Hence, $P(n+1)$ is true, and the induction proof is complete. ■

Various subsets of properties in Theorem 5.7.4 provide alternative characterizations of trees, though we won't prove this. For example, a *connected* graph with a number of vertices one larger than the number of edges is necessarily a tree. Also, a graph with unique paths between every pair of vertices is necessarily a tree.

5.7.3 Spanning Trees

Trees are everywhere. In fact, every connected graph contains a subgraph that is a tree with the same vertices as the graph. This is a called a *spanning tree* for the graph. For example, Figure 5.29 is a connected graph with a spanning tree highlighted.

Theorem 5.7.5. *Every connected graph contains a spanning tree.*

Proof. By contradiction. Assume there is some connected graph G that has no spanning tree and let T be a connected subgraph of G, with the same vertices as G, and with the smallest number of edges possible for such a subgraph. By the assumption, T is not a spanning tree and so it contains some cycle:

$$\{v_0, v_1\}, \{v_1, v_2\}, \ldots, \{v_k, v_0\}$$

Suppose that we remove the last edge, $\{v_k, v_0\}$. If a pair of vertices x and y was joined by a path not containing $\{v_k, v_0\}$, then they remain joined by that path. On the other hand, if x and y were joined by a path containing $\{v_k, v_0\}$, then they

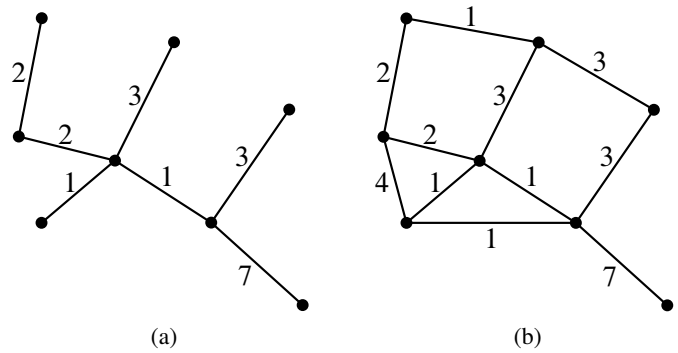


Figure 5.30 A spanning tree (a) with weight 19 for a graph (b).

remain joined by a walk containing the remainder of the cycle. By Lemma 5.4.2, they must also then be joined by a path. So all the vertices of G are still connected after we remove an edge from T. This is a contradiction, since T was defined to be a minimum size connected subgraph with all the vertices of G. So the theorem must be true. ■

5.7.4 Minimum Weight Spanning Trees

Spanning trees are interesting because they connect all the nodes of a graph using the smallest possible number of edges. For example the spanning tree for the 6-node graph shown in Figure 5.29 has 5 edges.

Spanning trees are very useful in practice, but in the real world, not all spanning trees are equally desirable. That's because, in practice, there are often costs associated with the edges of the graph.

For example, suppose the nodes of a graph represent buildings or towns and edges represent connections between buildings or towns. The cost to actually make a connection may vary a lot from one pair of buildings or towns to another. The cost might depend on distance or topography. For example, the cost to connect LA to NY might be much higher than that to connect NY to Boston. Or the cost of a pipe through Manhattan might be more than the cost of a pipe through a cornfield.

In any case, we typically represent the cost to connect pairs of nodes with a weighted edge, where the weight of the edge is its cost. The weight of a spanning tree is then just the sum of the weights of the edges in the tree. For example, the weight of the spanning tree shown in Figure 5.30 is 19.

The goal, of course, is to find the spanning tree with minimum weight, called the min-weight spanning tree (MST for short).

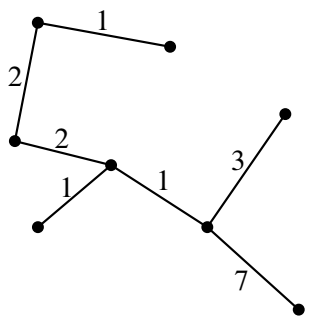


Figure 5.31 An MST with weight 17 for the graph in Figure 5.30(b).

Definition 5.7.6. The *min-weight spanning tree* (MST) of an edge-weighted graph G is the spanning tree of G with the smallest possible sum of edge weights.

Is the spanning tree shown in Figure 5.30(a) an MST of the weighted graph shown in Figure 5.30(b)? Actually, it is not, since the tree shown in Figure 5.31 is also a spanning tree of the graph shown in Figure 5.30(b), and this spanning tree has weight 17.

What about the tree shown in Figure 5.31? Is it an MST? It seems to be, but how do we prove it? In general, how do we find an MST? We could, of course, enumerate all trees, but this could take forever for very large graphs.

Here are two possible algorithms:

Algorithm 1. *Grow a tree one edge at a time by adding the minimum weight edge possible to the tree, making sure that you have a tree at each step.*

Algorithm 2. *Grow a subgraph one edge at a time by adding the minimum-weight edge possible to the subgraph, making sure that you have an acyclic subgraph at each step.*

For example, in the weighted graph we have been considering, we might run Algorithm 1 as follows. We would start by choosing one of the weight 1 edges, since this is the smallest weight in the graph. Suppose we chose the weight 1 edge on the bottom of the triangle of weight 1 edges in our graph. This edge is incident to two weight 1 edges, a weight 4 edge, a weight 7 edge, and a weight 3 edge. We would then choose the incident edge of minimum weight. In this case, one of the two weight 1 edges. At this point, we cannot choose the third weight 1 edge since this would form a cycle, but we can continue by choosing a weight 2 edge. We might end up with the spanning tree shown in Figure 5.32, which has weight 17, the smallest we've seen so far.

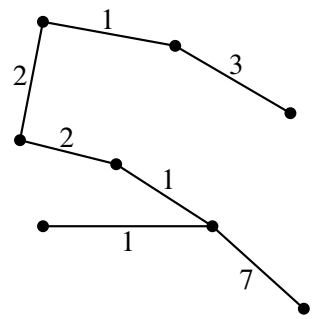


Figure 5.32 A spanning tree found by Algorithm 1.

Now suppose we instead ran Algorithm 2 on our graph. We might again choose the weight 1 edge on the bottom of the triangle of weight 1 edges in our graph. Now, instead of choosing one of the weight 1 edges it touches, we might choose the weight 1 edge on the top of the graph. Note that this edge still has minimum weight, and does not cause us to form a cycle, so Algorithm 2 can choose it. We would then choose one of the remaining weight 1 edges. Note that neither causes us to form a cycle. Continuing the algorithm, we may end up with the same spanning tree in Figure 5.32, though this need not always be the case.

It turns out that both algorithms work, but they might end up with different MSTs. The MST is not necessarily unique—indeed, if all edges of an n-node graph have the same weight ($= 1$), then all spanning trees have weight $n - 1$.

These are examples of greedy approaches to optimization. Sometimes it works and sometimes it doesn't. The good news is that it works to find the MST. In fact, both variations work. It's a little easier to prove it for Algorithm 2, so we'll do that one here.

Theorem 5.7.7. *For any connected, weighted graph G, Algorithm 2 produces an MST for G.*

Proof. The proof is a bit tricky. We need to show the algorithm terminates, that is, that if we have selected fewer than $n - 1$ edges, then we can always find an edge to add that does not create a cycle. We also need to show the algorithm creates a tree of minimum weight.

The key to doing all of this is to show that the algorithm never gets stuck or goes in a bad direction by adding an edge that will keep us from ultimately producing an MST. The natural way to prove this is to show that the set of edges selected at any point is contained in some MST—that is, we can always get to where we need to be. We'll state this as a lemma.

Lemma 5.7.8. *For any $m \geq 0$, let S consist of the first m edges selected by Algorithm 2. Then there exists some MST $T = (V, E)$ for G such that $S \subseteq E$, that is, the set of edges that we are growing is always contained in some MST.*

We'll prove this momentarily, but first let's see why it helps to prove the theorem. Assume the lemma is true. Then how do we know Algorithm 2 can always find an edge to add without creating a cycle? Well, as long as there are fewer than $n - 1$ edges picked, there exists some edge in $E - S$ and so there is an edge that we can add to S without forming a cycle. Next, how do we know that we get an MST at the end? Well, once $m = n - 1$, we know that S is an MST.

Ok, so the theorem is an easy corollary of the lemma. To prove the lemma, we'll use induction on the number of edges chosen by the algorithm so far. This is very typical in proving that an algorithm preserves some kind of invariant condition—induct on the number of steps taken, that is, the number of edges added.

Our inductive hypothesis $P(m)$ is the following: for any G and any set S of m edges initially selected by Algorithm 2, there exists an MST $T = (V, E)$ of G such that $S \subseteq E$.

For the base case, we need to show $P(0)$. In this case, $S = \emptyset$, so $S \subseteq E$ trivially holds for any MST $T = (V, E)$.

For the inductive step, we assume $P(m)$ holds and show that it implies $P(m+1)$. Let e denote the $(m+1)$st edge selected by Algorithm 2, and let S denote the first m edges selected by Algorithm 2. Let $T^* = (V^*, E^*)$ be the MST such that $S \subseteq E^*$, which exists by the inductive hypothesis. There are now two cases:

Case 1: $e \in E^*$, in which case $S \cup \{e\} \subseteq E^*$, and thus $P(m + 1)$ holds.

Case 2: $e \notin E^*$, as illustrated in Figure 5.33. Now we need to find a different MST that contains S and e.

What happens when we add e to T^*? Since T^* is a tree, we get a cycle. (Here we used part 3 of Theorem 5.7.4.) Moreover, the cycle cannot only contains edges in S, since e was chosen so that together with the edges in S, it does not form a cycle. This implies that $\{e\} \cup T^*$ contains a cycle that contains an edge e' of $E^* - S$. For example, such an e' is shown in Figure 5.33.

Note that the weight of e is at most that of e'. This is because Algorithm 2 picks the minimum weight edge that does not make a cycle with S. Since $e' \in T^*$, e' cannot make a cycle with S and if the weight of e were greater than the weight of e', Algorithm 2 would not have selected e ahead of e'.

Okay, we're almost done. Now we'll make an MST that contains $S \cup \{e\}$. Let $T^{**} = (V, E^{**})$ where $E^{**} = (E^* - \{e'\}) \cup \{e\}$, that is, we swap e and e' in T^*.

Claim 5.7.9. *T^{**} is an MST.*

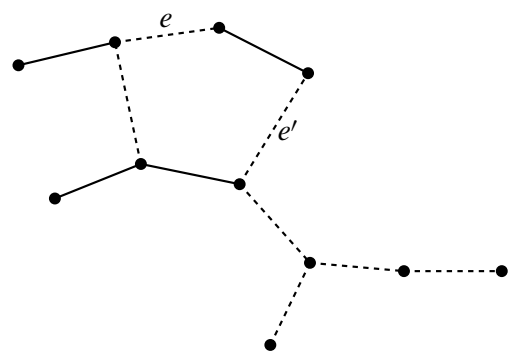


Figure 5.33 The graph formed by adding e to T^*. Edges of S are denoted with solid lines and edges of $E^* - S$ are denoted with dashed lines.

Proof of claim. We first show that T^{**} is a spanning tree. T^{**} is acyclic because it was produced by removing an edge from the only cycle in $T^* \cup \{e\}$. T^{**} is connected since the edge we deleted from $T^* \cup \{e\}$ was on a cycle. Since T^{**} contains all the nodes of G, it must be a spanning tree for G.

Now let's look at the weight of T^{**}. Well, since the weight of e was at most that of e', the weight of T^{**} is at most that of T^*, and thus T^{**} is an MST for G. ■

Since $S \cup \{e\} \subseteq E^{**}$, $P(m+1)$ holds. Thus, Algorithm 2 must eventually produce an MST. This will happens when it adds $n-1$ edges to the subgraph it builds. ■

So now we know for sure that the MST for our example graph has weight 17 since it was produced by Algorithm 2. And we have a fast algorithm for finding a minimum-weight spanning tree for any graph.

5.8 Planar Graphs

5.8.1 Drawing Graphs in the Plane

Suppose there are three dog houses and three human houses, as shown in Figure 5.34. Can you find a route from each dog house to each human house such that no route crosses any other route?

A *quadrapus* is a little-known animal similar to an octopus, but with four arms. Suppose there are five quadrapi resting on the sea floor, as shown in Figure 5.35.

Chapter 11

MFCS 6: Directed Graphs

6 Directed Graphs

6.1 Definitions

So far, we have been working with graphs with undirected edges. A *directed edge* is an edge where the endpoints are distinguished—one is the *head* and one is the *tail*. In particular, a directed edge is specified as an ordered pair of vertices u, v and is denoted by (u, v) or $u \to v$. In this case, u is the *tail* of the edge and v is the *head*. For example, see Figure 6.1.

A graph with directed edges is called a *directed graph* or *digraph*.

Definition 6.1.1. A directed graph $G = (V, E)$ consists of a nonempty set of nodes V and a set of directed edges E. Each edge e of E is specified by an ordered pair of vertices $u, v \in V$. A directed graph is *simple* if it has no *loops* (that is, edges of the form $u \to u$) and no multiple edges.

Since we will focus on the case of simple directed graphs in this chapter, we will generally omit the word *simple* when referring to them. Note that such a graph can contain an edge $u \to v$ as well as the edge $v \to u$ since these are different edges (for example, they have a different tail).

Directed graphs arise in applications where the relationship represented by an edge is 1-way or asymmetric. Examples include: a 1-way street, one person likes another but the feeling is not necessarily reciprocated, a communication channel such as a cable modem that has more capacity for downloading than uploading, one entity is larger than another, and one job needs to be completed before another job can begin. We'll see several such examples in this chapter and also in Chapter 7.

Most all of the definitions for undirected graphs from Chapter 5 carry over in a natural way for directed graphs. For example, two directed graphs $G_1 = (V_1, E_1)$ and $G_2 = (V_2, E_2)$ are *isomorphic* if there exists a bijection $f : V_1 \to V_2$ such that for every pair of vertices $u, v \in V_1$,

$$u \to v \in E_1 \qquad \text{IFF} \qquad f(u) \to f(v) \in E_2.$$

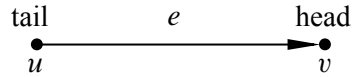


Figure 6.1 A directed edge $e = (u, v)$. u is the tail of e and v is the head of e.

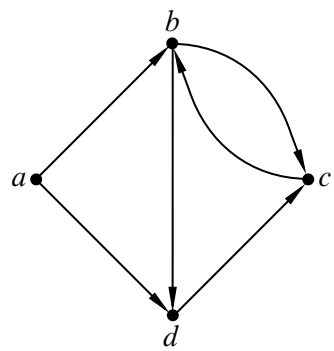


Figure 6.2 A 4-node directed graph with 6 edges.

Directed graphs have adjacency matrices just like undirected graphs. In the case of a directed graph $G = (V, E)$, the adjacency matrix $A_G = \{a_{ij}\}$ is defined so that

$$a_{ij} = \begin{cases} 1 & \text{if } i \to j \in E \\ 0 & \text{otherwise.} \end{cases}$$

The only difference is that the adjacency matrix for a directed graph is not necessarily symmetric (that is, it may be that $A_G^T \neq A_G$).

6.1.1 Degrees

With directed graphs, the notion of degree splits into *indegree* and *outdegree*. For example, indegree$(c) = 2$ and outdegree$(c) = 1$ for the graph in Figure 6.2. If a node has outdegree 0, it is called a *sink*; if it has indegree 0, it is called a *source*. The graph in Figure 6.2 has one source (node a) and no sinks.

6.1.2 Directed Walks, Paths, and Cycles

The definitions for (directed) walks, paths, and cycles in a directed graph are similar to those for undirected graphs except that the direction of the edges need to be consistent with the order in which the walk is traversed.

Definition 6.1.2. A *directed walk* (or more simply, a *walk*) in a directed graph G is a sequence of vertices $v_0, v_1, \ldots, v_k$ and edges

$$v_0 \to v_1, v_1 \to v_2, \ldots, v_{k-1} \to v_k$$

such that $v_{i-1} \to v_i$ is an edge of G for all i where $0 \leq i < k$. A *directed path* (or *path*) in a directed graph is a walk where the nodes in the walk are all different. A *directed closed walk* (or *closed walk*) in a directed graph is a walk

where $v_0 = v_k$. A *directed cycle* (or *cycle*) in a directed graph is a closed walk where all the vertices v_i are different for $0 \leq i < k$.

As with undirected graphs, we will typically refer to a walk in a directed graph by a sequence of vertices. For example, for the graph in Figure 6.2,

- a, b, c, b, d is a walk,
- a, b, d is a path,
- d, c, b, c, b, d is a closed walk, and
- b, d, c, b is a cycle.

Note that b, c, b is also a cycle for the graph in Figure 6.2. This is a cycle of length 2. Such cycles are not possible with undirected graphs.

Also note that

$$c, b, a, d$$

is *not* a walk in the graph shown in Figure 6.2, since $b \to a$ is not an edge in this graph. (You are *not* allowed to traverse edges in the wrong direction as part of a walk.)

A path or cycle in a directed graph is said to be *Hamiltonian* if it visits every node in the graph. For example, a, b, d, c is the only Hamiltonian path for the graph in Figure 6.2. The graph in Figure 6.2 does not have a Hamiltonian cycle.

A walk in a directed graph is said to be *Eulerian* if it contains every edge. The graph shown in Figure 6.2 does not have an Eulerian walk. Can you see why not? (Hint: Look at node a.)

6.1.3 Strong Connectivity

The notion of being connected is a little more complicated for a directed graph than it is for an undirected graph. For example, should we consider the graph in Figure 6.2 to be connected? There is a path from node a to every other node so on that basis, we might answer "Yes." But there is no path from nodes b, c, or d to node a, and so on that basis, we might answer "No." For this reason, graph theorists have come up with the notion of *strong* connectivity for directed graphs.

Definition 6.1.3. A directed graph $G = (V, E)$ is said to be *strongly connected* if for every pair of nodes $u, v \in V$, there is a directed path from u to v (and vice-versa) in G.

For example, the graph in Figure 6.2 is not strongly connected since there is no directed path from node b to node a. But if node a is removed, the resulting graph would be strongly connected.

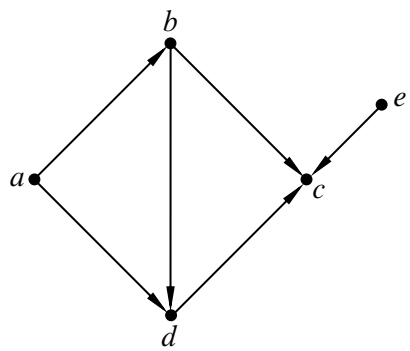


Figure 6.3 A 4-node directed acyclic graph (DAG).

A directed graph is said to be *weakly connected* (or, more simply, *connected*) if the corresponding undirected graph (where directed edges $u \to v$ and/or $v \to u$ are replaced with a single undirected edge $\{u, v\}$ is connected. For example, the graph in Figure 6.2 is weakly connected.

6.1.4 DAGs

If an undirected graph does not have any cycles, then it is a tree or a forest. But what does a directed graph look like if it has no cycles? For example, consider the graph in Figure 6.3. This graph is weakly connected and has no directed cycles but it certainly does not look like a tree.

Definition 6.1.4. A directed graph is called a *directed acyclic graph* (or, *DAG*) if it does not contain any directed cycles.

A first glance, DAGs don't appear to be particularly interesting. But first impressions are not always accurate. In fact, DAGs arise in many scheduling and optimization problems and they have several interesting properties. We will study them extensively in Chapter 7.

6.2 Tournament Graphs

Suppose that n players compete in a round-robin tournament and that for every pair of players u and v, either u beats v or v beats u. Interpreting the results of a round-robin tournament can be problematic—there might be all sorts of cycles where x beats y and y beats z, yet z beats x. Who is the best player? Graph theory does not solve this problem but it can provide some interesting perspectives.

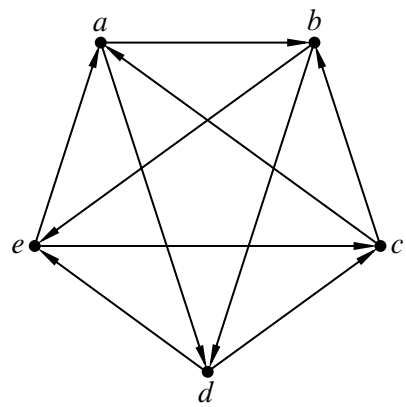


Figure 6.4 A 5-node tournament graph.

The results of a round-robin tournament can be represented with a *tournament graph*. This is a directed graph in which the vertices represent players and the edges indicate the outcomes of games. In particular, an edge from u to v indicates that player u defeated player v. In a round-robin tournament, every pair of players has a match. Thus, in a tournament graph there is either an edge from u to v or an edge from v to u (but not both) for *every* pair of distinct vertices u and v. An example of a tournament graph is shown in Figure 6.4.

6.2.1 Finding a Hamiltonian Path in a Tournament Graph

We're going to prove that in every round-robin tournament, there exists a ranking of the players such that each player lost to the player one position higher. For example, in the tournament corresponding to Figure 6.4, the ranking

$$a > b > d > e > c$$

satisfies this criterion, because b lost to a, d lost to b, e lost to d, and c lost to e. In graph terms, proving the existence of such a ranking amounts to proving that every tournament graph has a Hamiltonian path.

Theorem 6.2.1. *Every tournament graph contains a directed Hamiltonian path.*

Proof. We use strong induction. Let $P(n)$ be the proposition that every tournament graph with n vertices contains a directed Hamiltonian path.

Base case: $P(1)$ is trivially true; every graph with a single vertex has a Hamiltonian path consisting of only that vertex.

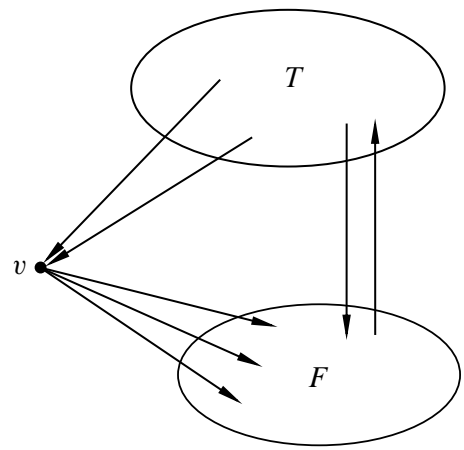


Figure 6.5 The sets T and F in a tournament graph.

Inductive step: For $n \geq 1$, we assume that $P(1), \ldots, P(n)$ are all true and prove $P(n+1)$. Consider a tournament graph $G = (V, E)$ with $n+1$ players. Select one vertex v arbitrarily. Every other vertex in the tournament either has an edge *to* vertex v or an edge *from* vertex v. Thus, we can partition the remaining vertices into two corresponding sets, T and F, each containing at most n vertices, where $T = \{ u \mid u \to v \in E \}$ and $F = \{ u \mid v \to u \in E \}$. For example, see Figure 6.5.

The vertices in T together with the edges that join them form a smaller tournament. Thus, by strong induction, there is a Hamiltonian path within T. Similarly, there is a Hamiltonian path within the tournament on the vertices in F. Joining the path in T to the vertex v followed by the path in F gives a Hamiltonian path through the whole tournament. As special cases, if T or F is empty, then so is the corresponding portion of the path. ■

The ranking defined by a Hamiltonian path is not entirely satisfactory. For example, in the tournament associated with Figure 6.4, notice that the lowest-ranked player, c, actually defeated the highest-ranked player, a.

In practice, players are typically ranked according to how many victories they achieve. This makes sense for several reasons. One not-so-obvious reason is that if the player with the most victories does not beat some other player v, he is guaranteed to have at least beaten a third player who beat v. We'll prove this fact shortly. But first, let's talk about chickens.

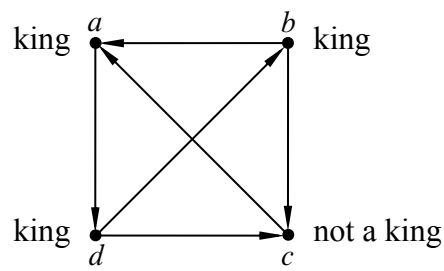


Figure 6.6 A 4-chicken tournament in which chickens a, b, and d are kings.

.

6.2.2 The King Chicken Theorem

Suppose that there are n chickens in a farmyard. Chickens are rather aggressive birds that tend to establish dominance in relationships by pecking. (Hence the term "pecking order.") In particular, for each pair of distinct chickens, either the first pecks the second or the second pecks the first, but not both. We say that chicken u *virtually pecks* chicken v if either:

- Chicken u directly pecks chicken v, or
- Chicken u pecks some other chicken w who in turn pecks chicken v.

A chicken that virtually pecks every other chicken is called a *king chicken*.

We can model this situation with a tournament digraph. The vertices are chickens, and an edge $u \to v$ indicates that chicken u pecks chicken v. In the tournament shown in Figure 6.6, three of the four chickens are kings. Chicken c is not a king in this example since it does not peck chicken b and it does not peck any chicken that pecks chicken b. Chicken a *is* a king since it pecks chicken d, who in turn pecks chickens b and c.

Theorem 6.2.2 (King Chicken Theorem)**.** *The chicken with the largest outdegree in an n-chicken tournament is a king.*

Proof. By contradiction. Let u be a node in a tournament graph $G = (V, E)$ with maximum outdegree and suppose that u is not a king. Let $Y = \{ v \mid u \to v \in E \}$ be the set of chickens that chicken u pecks. Then $\text{outdegree}(u) = |Y|$.

Since u is not a king, there is a chicken $x \notin Y$ (that is, x is not pecked by chicken u) and that is not pecked by any chicken in Y. Since for any pair of chickens, one pecks the other, this means that x pecks u as well as every chicken in Y. This means that

$$\text{outdegree}(x) = |Y| + 1 > \text{outdegree}(u).$$

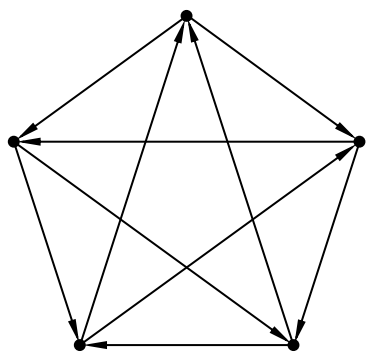


Figure 6.7 A 5-chicken tournament in which every chicken is a king.

But u was assumed to be the node with the largest degree in the tournament, so we have a contradiction. Hence, u must be a king. ■

Theorem 6.2.2 means that if the player with the most victories is defeated by another player x, then at least he/she defeats some third player that defeats x. In this sense, the player with the most victories has some sort of bragging rights over every other player. Unfortunately, as Figure 6.6 illustrates, there can be many other players with such bragging rights, even some with fewer victories. Indeed, for some tournaments, it is possible that every player is a "king." For example, consider the tournament illustrated in Figure 6.7.

6.3 Communication Networks

While reasoning about chickens pecking each other may be amusing (to mathematicians, at least), the use of directed graphs to model communication networks is very serious business. In the context of communication problems, vertices represent computers, processors, or switches, and edges represent wires, fiber, or other transmission lines through which data flows. For some communication networks, like the Internet, the corresponding graph is enormous and largely chaotic. Highly structured networks, such as an array or butterfly, by contrast, find application in telephone switching systems and the communication hardware inside parallel computers.

6.3.1 Packet Routing

Whatever architecture is chosen, the goal of a communication network is to get data from *inputs* to *outputs*. In this text, we will focus on a model in which the data to be communicated is in the form of a *packet*. In practice, a packet would consist of a fixed amount of data, and a message (such as a web page or a movie) would consist of many packets.

For simplicity, we will restrict our attention to the scenario where there is just one packet at every input and where there is just one packet destined for each output. We will denote the number of inputs and output by N and we will often assume that N is a power of two.

We will specify the desired destinations of the packets by a permutation[1] of 0, 1, ..., $N-1$. So a permutation, π, defines a *routing problem*: get a packet that starts at input i to output $\pi(i)$ for $0 \leq i < N$. A *routing* P that *solves* a routing problem π is a set of paths from each input to its specified output. That is, P is a set of paths, P_i, for $i = 0, \ldots, N-1$, where P_i goes from input i to output $\pi(i)$.

Of course, the goal is to get all the packets to their destinations as quickly as possible using as little hardware as possible. The time needed to get the packages to their destinations depends on several factors, such as how many switches they need to go through and how many packets will need to cross the same wire. We will assume that only one packet can cross a wire at a time. The complexity of the hardware depends on factors such as the number of switches needed and the size of the switches.

Let's see how all this works with an example—routing packets on a complete binary tree.

6.3.2 The Complete Binary Tree

One of the simplest structured communications networks is a *complete binary tree*. A complete binary tree with 4 inputs and 4 outputs is shown in Figure 6.8.

In this diagram and many that follow, the squares represent *terminals* (that is, the inputs and outputs), and the circles represent *switches*, which direct packets through the network. A switch receives packets on incoming edges and relays them forward along the outgoing edges. Thus, you can imagine a data packet hopping through the network from an input terminal, through a sequence of switches joined by directed edges, to an output terminal.

Recall that there is a unique simple path between every pair of vertices in a tree. So the natural way to route a packet of data from an input terminal to an output terminal in the complete binary tree is along the corresponding directed path. For

[1] A permutation of a sequence is a reordering of the sequence.

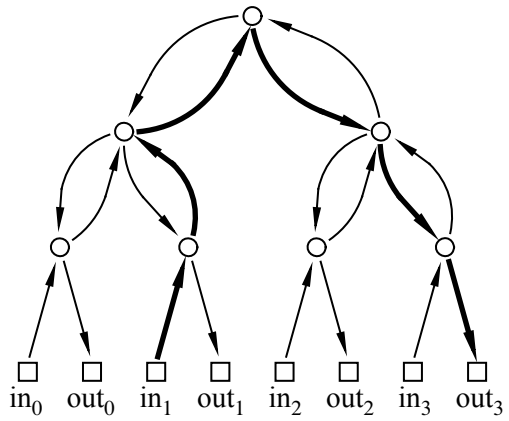


Figure 6.8 A 4-input, 4-output complete binary tree. The squares represent terminals (input and output registers) and the circles represent switches. Directed edges represent communication channels in the network through which data packets can move. The unique path from input 1 to output 3 is shown in bold.

example, the route of a packet traveling from input 1 to output 3 is shown in bold in Figure 6.8.

6.3.3 Network Diameter

The delay between the time that a packet arrives at an input and the time that it reaches its designated output is referred to as *latency* and it is a critical issue in communication networks. If congestion is not a factor, then this delay is generally proportional to the length of the path a packet follows. Assuming it takes one time unit to travel across a wire, and that there are no additional delays at switches, the delay of a packet will be the number of wires it crosses going from input to output.[2]

Generally a packet is routed from input to output using the shortest path possible. The length of this shortest path is the *distance* between the input and output. With a shortest path routing, the worst possible delay is the distance between the input and output that are farthest apart. This is called the *diameter* of the network. In other words, the diameter of a network[3] is the maximum length of any shortest

[2]Latency can also be measured as the number of switches that a packet must pass through when traveling between the most distant input and output, since switches usually have the biggest impact on network speed. For example, in the complete binary tree example, the packet traveling from input 1 to output 3 crosses 5 switches, which is 1 less than the number of edges traversed.

[3]The usual definition of *diameter* for a general graph (simple or directed) is the largest distance between *any* two vertices, but in the context of a communication network, we're only interested in the distance between inputs and outputs, not between arbitrary pairs of vertices.

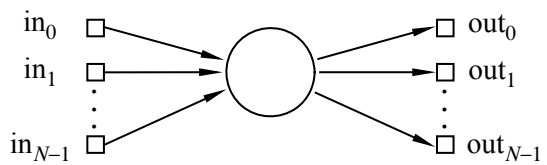


Figure 6.9 A monster $N \times N$ switch.

path between an input and an output. For example, in the complete binary tree shown in Figure 6.8, the distance from input 1 to output 3 is six. No input and output are farther apart than this, so the diameter of this tree is also six.

More generally, the diameter of a complete binary tree with N inputs and outputs is $2 \log N + 2$. (All logarithms in this lecture—and in most of computer science—are base 2.) This is quite good, because the logarithm function grows very slowly. We could connect $2^{20} = 1{,}048{,}576$ inputs and outputs using a complete binary tree and the worst input-output delay for any packet would be this diameter, namely, $2 \log(2^{20}) + 2 = 42$.

6.3.4 Switch Size

One way to reduce the diameter of a network (and hence the latency needed to route packets) is to use larger switches. For example, in the complete binary tree, most of the switches have three incoming edges and three outgoing edges, which makes them 3×3 switches. If we had 4×4 switches, then we could construct a complete *ternary* tree with an even smaller diameter. In principle, we could even connect up all the inputs and outputs via a single monster $N \times N$ switch, as shown in Figure 6.9. In this case, the "network" would consist of a single switch and the latency would be 2.

This isn't very productive, however, since we've just concealed the original network design problem inside this abstract monster switch. Eventually, we'll have to design the internals of the monster switch using simpler components, and then we're right back where we started. So the challenge in designing a communication network is figuring out how to get the functionality of an $N \times N$ switch using fixed size, elementary devices, like 3×3 switches.

6.3.5 Switch Count

Another goal in designing a communication network is to use as few switches as possible. The number of switches in a complete binary tree is $1 + 2 + 4 + 8 + \cdots + N = 2N - 1$, since there is 1 switch at the top (the "root switch"), 2 below it, 4 below those, and so forth. This is nearly the best possible with 3×3 switches,

since at least one switch will be needed for each pair of inputs and outputs.

6.3.6 Congestion

The complete binary tree has a fatal drawback: the root switch is a bottleneck. At best, this switch must handle an enormous amount of traffic: every packet traveling from the left side of the network to the right or vice-versa. Passing all these packets through a single switch could take a long time. At worst, if this switch fails, the network is broken into two equal-sized pieces.

The traffic through the root depends on the routing problem. For example, if the routing problem is given by the identity permutation, $\pi(i) ::= i$, then there is an easy routing P that solves the problem: let P_i be the path from input i up through one switch and back down to output i. On the other hand, if the problem was given by $\pi(i) ::= (N-1)-i$, then in *any* solution P for π, each path P_i beginning at input i must eventually loop all the way up through the root switch and then travel back down to output $(N-1)-i$.

We can distinguish between a "good" set of paths and a "bad" set based on congestion. The *congestion* of a routing, P, is equal to the largest number of paths in P that pass through a single switch. Generally, lower congestion is better since packets can be delayed at an overloaded switch.

By extending the notion of congestion to networks, we can also distinguish between "good" and "bad" networks with respect to bottleneck problems. For each routing problem, π, for the network, we assume a routing is chosen that optimizes congestion, that is, that has the minimum congestion among all routings that solve π. Then the largest congestion that will ever be suffered by a switch will be the maximum congestion among these optimal routings. This "maxi-min" congestion is called the *congestion of the network*.

You may find it helpful to think about max congestion in terms of a value game. You design your spiffy, new communication network; this defines the game. Your opponent makes the first move in the game: she inspects your network and specifies a permutation routing problem that will strain your network. You move second: given her specification, you choose the precise paths that the packets should take through your network; you're trying to avoid overloading any one switch. Then her next move is to pick a switch with as large as possible a number of packets passing through it; this number is her score in the competition. The max congestion of your network is the largest score she can ensure; in other words, it is precisely the max-value of this game.

For example, if your enemy were trying to defeat the complete binary tree, she would choose a permutation like $\pi(i) = (N-1)-i$. Then for *every* packet i, you would be forced to select a path $P_{i,\pi(i)}$ passing through the root switch. Then, your

network	diameter	switch size	# switches	congestion
complete binary tree	$2\log N + 2$	3×3	$2N - 1$	N

Table 6.1 A summary of the attributes of the complete binary tree.

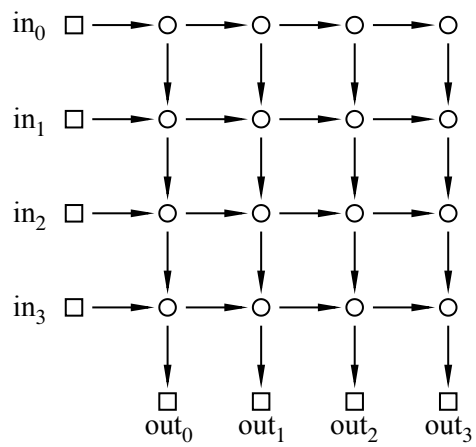


Figure 6.10 A 4×4 2-dimensional array.

enemy would choose the root switch and achieve a score of N. In other words, the max congestion of the complete binary tree is N—which is horrible!

We have summarized the results of our analysis of the complete binary tree in Table 6.1. Overall, the complete binary tree does well in every category except the last—congestion, and that is a killer in practice. Next, we will look at a network that solves the congestion problem, but at a very high cost.

6.3.7 The 2-d Array

An illustration of the $N \times N$ 2-d *array* (also known as the *grid* or *crossbar*) is shown in Figure 6.10 for the case when $N = 4$.

The diameter of the 4×4 2-d array is 8, which is the number of edges between input 0 and output 3. More generally, the diameter of a 2-d array with N inputs and outputs is $2N$, which is much worse than the diameter of the complete binary tree ($2\log N + 2$). On the other hand, replacing a complete binary tree with a 2-d array almost eliminates congestion.

Theorem 6.3.1. *The congestion of an N-input 2-d array is 2.*

Proof. First, we show that the congestion is at most 2. Let π be any permutation. Define a solution, P, for π to be the set of paths, P_i, where P_i goes to the right

network	diameter	switch size	# switches	congestion
complete binary tree	$2\log N + 2$	3×3	$2N - 1$	N
2-D array	$2N$	2×2	N^2	2

Table 6.2 Comparing the N-input 2-d array to the N-input complete binary tree.

from input i to column $\pi(i)$ and then goes down to output $\pi(i)$. In this solution, the switch in row i and column j encounters at most two packets: the packet originating at input i and the packet destined for output j.

Next, we show that the congestion is at least 2. This follows because in any routing problem, π, where $\pi(0) = 0$ and $\pi(N-1) = N-1$, two packets must pass through the lower left switch. ■

The characteristics of the 2-d array are recorded in Table 6.2. The crucial entry in this table is the number of switches, which is N^2. This is a major defect of the 2-d array; a network with $N = 1000$ inputs would require a *million* 2×2 switches! Still, for applications where N is small, the simplicity and low congestion of the array make it an attractive choice.

6.3.8 The Butterfly

The Holy Grail of switching networks would combine the best properties of the complete binary tree (low diameter, few switches) and the array (low congestion). The *butterfly* is a widely-used compromise between the two. A butterfly network with $N = 8$ inputs is shown in Figure 6.11.

The structure of the butterfly is certainly more complicated than that of the complete binary or 2-d array. Let's see how it is constructed.

All the terminals and switches in the network are in N rows. In particular, input i is at the left end of row i, and output i is at the right end of row i. Now let's label the rows in *binary* so that the label on row i is the binary number $b_1 b_2 \ldots b_{\log N}$ that represents the integer i.

Between the inputs and outputs, there are $\log(N) + 1$ levels of switches, numbered from 0 to $\log N$. Each level consists of a column of N switches, one per row. Thus, each switch in the network is uniquely identified by a sequence $(b_1, b_2, \ldots, b_{\log N}, l)$, where $b_1 b_2 \ldots b_{\log N}$ is the switch's row in binary and l is the switch's level.

All that remains is to describe how the switches are connected up. The basic

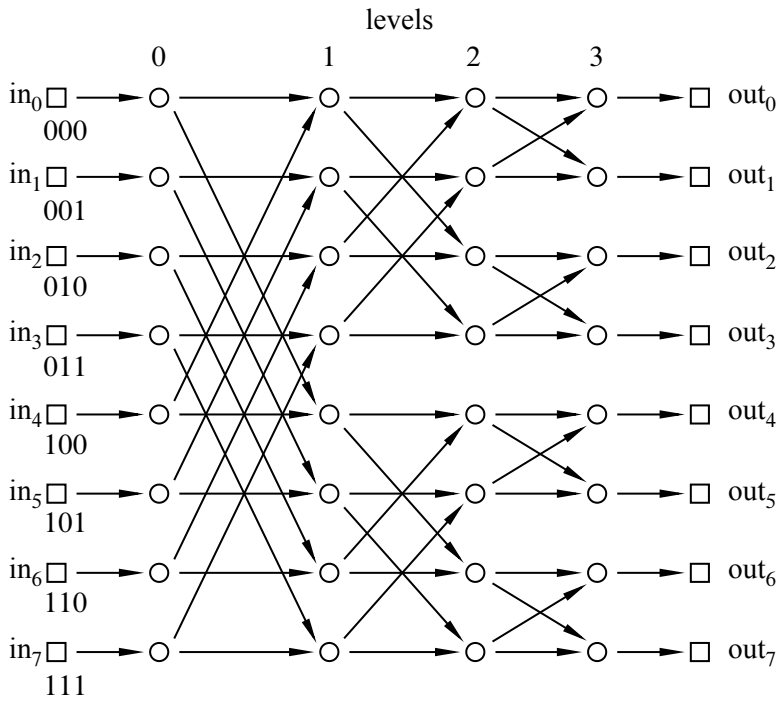


Figure 6.11 An 8-input/output butterfly.

connection pattern is expressed below in a compact notation:

$$(b_1, b_2, \ldots b_{l+1}, \ldots b_{\log N}, l) \begin{array}{l} \nearrow \quad (b_1, b_2, \ldots b_{l+1}, \ldots b_{\log N}, l+1) \\ \\ \searrow \quad (b_1, b_2, \ldots \overline{b_{l+1}}, \ldots b_{\log N}, l+1) \end{array}$$

This says that there are directed edges from switch $(b_1, b_2, \ldots, b_{\log N}, l)$ to two switches in the next level. One edges leads to the switch in the *same* row, and the other edge leads to the switch in the row obtained by *inverting* the $(l+1)$st bit b_{l+1}. For example, referring back to the illustration of the size $N = 8$ butterfly, there is an edge from switch $(0, 0, 0, 0)$ to switch $(0, 0, 0, 1)$, which is in the same row, and to switch $(1, 0, 0, 1)$, which is in the row obtained by inverting bit $l + 1 = 1$.

The butterfly network has a recursive structure; specifically, a butterfly of size $2N$ consists of two butterflies of size N and one additional level of switches. Each switch in the additional level has directed edges to a corresponding switch in each of the smaller butterflies. For example, see Figure 6.12.

Despite the relatively complicated structure of the butterfly, there is a simple way to route packets through its switches. In particular, suppose that we want to send a packet from input $x_1 x_2 \ldots x_{\log N}$ to output $y_1 y_2 \ldots y_{\log N}$. (Here we are specifying the input and output numbers in binary.) Roughly, the plan is to "correct" the first bit on the first level, correct the second bit on the second level, and so forth. Thus, the sequence of switches visited by the packet is:

$$\begin{aligned} (x_1, x_2, x_3, \ldots, x_{\log N}, 0) &\to (y_1, x_2, x_3, \ldots, x_{\log N}, 1) \\ &\to (y_1, y_2, x_3, \ldots, x_{\log N}, 2) \\ &\to (y_1, y_2, y_3, \ldots, x_{\log N}, 3) \\ &\to \quad \ldots \\ &\to (y_1, y_2, y_3, \ldots, y_{\log N}, \log N) \end{aligned}$$

In fact, this is the *only* path from the input to the output!

The congestion of the butterfly network is about $\sqrt{N}$. More precisely, the congestion is $\sqrt{N}$ if N is an even power of 2 and $\sqrt{N/2}$ if N is an odd power of 2. The task of proving this fact has been left to the problem section.[4]

A comparison of the butterfly with the complete binary tree and the 2-d array is provided in Table 6.3. As you can see, the butterfly has lower congestion than the complete binary tree. And it uses fewer switches and has lower diameter than the

[4]The routing problems that result in $\sqrt{N}$ congestion do arise in practice, but for most routing problems, the congestion is much lower (around $\log N$), which is one reason why the butterfly is useful in practice.

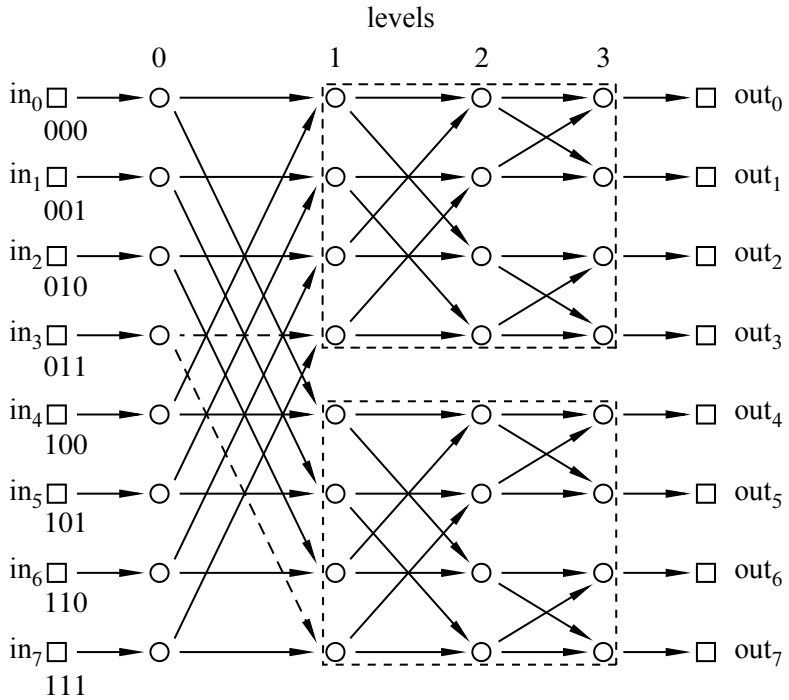


Figure 6.12 An N-input butterfly contains two $N/2$-input butterflies (shown in the dashed boxes). Each switch on the first level is adjacent to a corresponding switch in each of the sub-butterflies. For example, we have used dashed lines to show these edges for the node $(0, 1, 1, 0)$.

network	diameter	switch size	# switches	congestion
complete binary tree	$2\log N + 2$	3×3	$2N - 1$	N
2-D array	$2N$	2×2	N^2	2
butterfly	$\log N + 2$	2×2	$N(\log(N) + 1)$	$\sqrt{N}$ or $\sqrt{N/2}$

Table 6.3 A comparison of the N-input butterfly with the N-input complete binary tree and the N-input 2-d array.

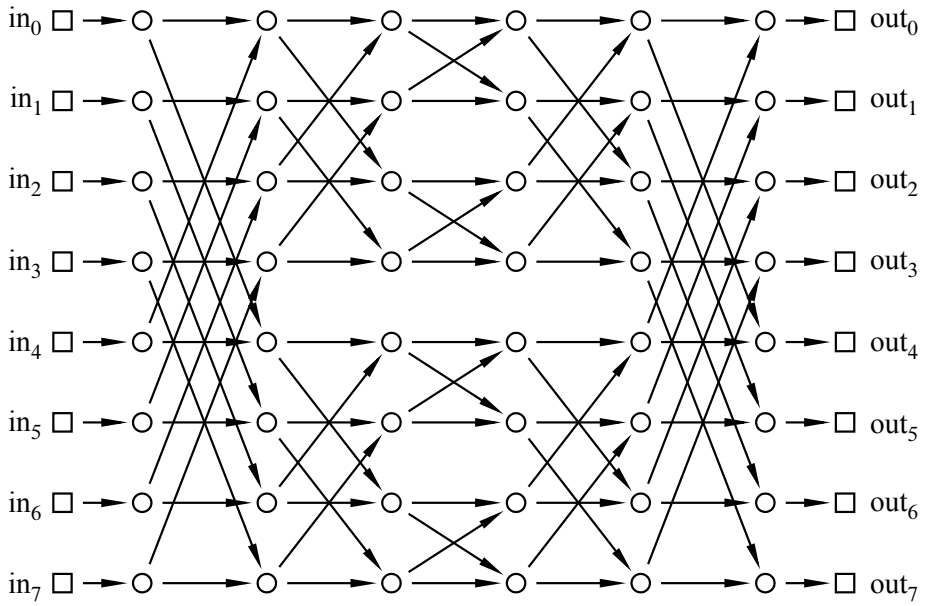


Figure 6.13 The 8-input Beneš network.

array. However, the butterfly does not capture the best qualities of each network, but rather is a compromise somewhere between the two. So our quest for the Holy Grail of routing networks goes on.

6.3.9 Beneš Network

In the 1960's, a researcher at Bell Labs named Václav Beneš had a remarkable idea. He obtained a marvelous communication network with congestion 1 by placing *two* butterflies back-to-back. For example, the 8-input Beneš network is shown in Figure 6.13.

Putting two butterflies back-to-back roughly doubles the number of switches and the diameter of a single butterfly, but it completely eliminates congestion problems! The proof of this fact relies on a clever induction argument that we'll come to in a

network	diameter	switch size	# switches	congestion
complete binary tree	$2\log N + 2$	3×3	$2N - 1$	N
2-D array	$2N$	2×2	N^2	2
butterfly	$\log N + 2$	2×2	$N(\log(N) + 1)$	$\sqrt{N}$ or $\sqrt{N/2}$
Beneš	$2\log N + 1$	2×2	$2N \log N$	1

Table 6.4 A comparison of the N-input Beneš network with the N-input complete binary tree, 2-d array, and butterfly.

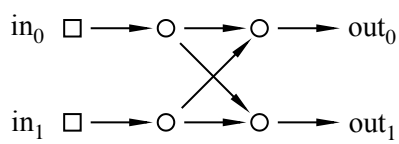


Figure 6.14 The 2-input Beneš network.

moment. Let's first see how the Beneš network stacks up against the other networks we have been studying. As you can see in Table 6.4, the Beneš network has small size and diameter, and completely eliminates congestion. The Holy Grail of routing networks is in hand!

Theorem 6.3.2. *The congestion of the N-input Beneš network is 1 for any N that is a power of 2.*

Proof. We use induction. Let $P(a)$ be the proposition that the congestion of the 2^a-input Beneš network is 1.

Base case ($a = 1$): We must show that the congestion of the 2^1-input Beneš network is 1. The network is shown in Figure 6.14.

There are only two possible permutation routing problems for a 2-input network. If $\pi(0) = 0$ and $\pi(1) = 1$, then we can route both packets along the straight edges. On the other hand, if $\pi(0) = 1$ and $\pi(1) = 0$, then we can route both packets along the diagonal edges. In both cases, a single packet passes through each switch.

Inductive step: We must show that $P(a)$ implies $P(a+1)$ where $a \geq 1$. Thus, we assume that the congestion of a 2^a-input Beneš network is 1 in order to prove that the congestion of a 2^{a+1}-input Beneš network is also 1.

Digression

Time out! Let's work through an example, develop some intuition, and then complete the proof. Notice that inside a Beneš network of size $2N$ lurk two Beneš subnetworks of size N. This follows from our earlier observation that a butterfly

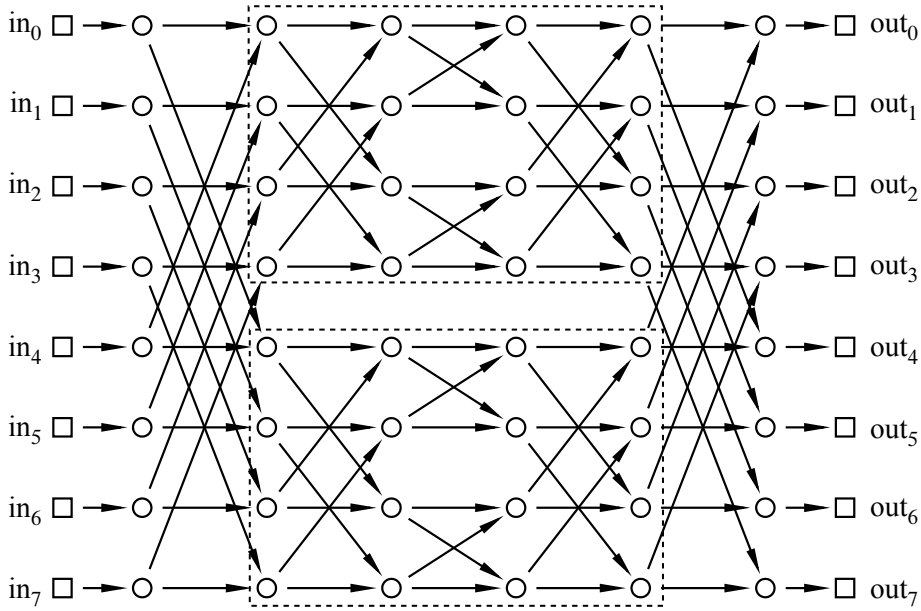


Figure 6.15 A $2N$-input Beneš network contains two N-input Beneš networks—shown here for $N = 4$.

of size $2N$ contains two butterflies of size N. In the Beneš network shown in Figure 6.15 with $N = 8$ inputs and outputs, the two 4-input/output subnetworks are shown in dashed boxes.

By the inductive assumption, the subnetworks can each route an arbitrary permutation with congestion 1. So if we can guide packets safely through just the first and last levels, then we can rely on induction for the rest! Let's see how this works in an example. Consider the following permutation routing problem:

$$\begin{aligned} \pi(0) &= 1 & \qquad \pi(4) &= 3 \\ \pi(1) &= 5 & \qquad \pi(5) &= 6 \\ \pi(2) &= 4 & \qquad \pi(6) &= 0 \\ \pi(3) &= 7 & \qquad \pi(7) &= 2 \end{aligned}$$

We can route each packet to its destination through either the upper subnetwork or the lower subnetwork. However, the choice for one packet may constrain the choice for another. For example, we can not route the packets at inputs 0 and 4 both through the same network since that would cause two packets to collide at a single switch, resulting in congestion. So one packet must go through the upper network and the other through the lower network. Similarly, the packets at inputs 1 and 5,

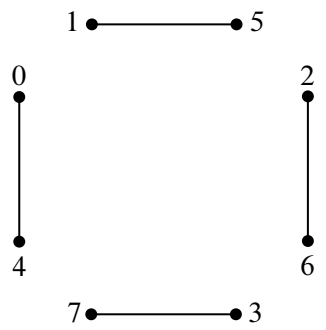


Figure 6.16 The beginnings of a constraint graph for our packet routing problem. Adjacent packets cannot be routed using the same sub-Beneš network.

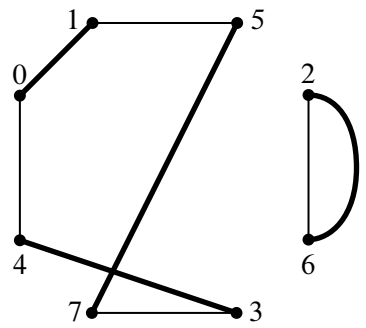


Figure 6.17 The updated constraint graph.

2 and 6, and 3 and 7 must be routed through different networks. Let's record these constraints in a graph. The vertices are the 8 packets (labeled according to their input position). If two packets must pass through different networks, then there is an edge between them. The resulting constraint graph is illustrated in Figure 6.16. Notice that at most one edge is incident to each vertex.

The output side of the network imposes some further constraints. For example, the packet destined for output 0 (which is packet 6) and the packet destined for output 4 (which is packet 2) can not both pass through the same network since that would require both packets to arrive from the same switch. Similarly, the packets destined for outputs 1 and 5, 2 and 6, and 3 and 7 must also pass through different switches. We can record these additional constraints in our constraint graph with gray edges, as is illustrated in Figure 6.17.

Notice that at most one new edge is incident to each vertex. The two lines drawn between vertices 2 and 6 reflect the two different reasons why these packets must be routed through different networks. However, we intend this to be a simple graph;

the two lines still signify a single edge.

Now here's the key insight: *a 2-coloring of the graph corresponds to a solution to the routing problem*. In particular, suppose that we could color each vertex either red or blue so that adjacent vertices are colored differently. Then all constraints are satisfied if we send the red packets through the upper network and the blue packets through the lower network.

The only remaining question is whether the constraint graph is 2-colorable. Fortunately, this is easy to verify:

Lemma 6.3.3. *If the edges of an undirected graph G can be grouped into two sets such that every vertex is incident to at most 1 edge from each set, then the graph is 2-colorable.*

Proof. Since the two sets of edges may overlap, let's call an edge that is in both sets a *doubled edge*. Note that no other edge can be incident to either of the endpoints of a doubled edge, since that endpoint would then be incident to two edges from the same set. This means that doubled edges form connected components with 2 nodes. Such connected components are easily colored with 2 colors and so we can henceforth ignore them and focus on the remaining nodes and edges, which form a simple graph.

By Theorem 5.6.2, we know that if a simple graph has no odd cycles, then it is 2-colorable. So all we need to do is show that every cycle in G has even length. This is easy since any cycle in G must traverse successive edges that alternate from one set to the other. In particular, a closed walk must traverse a path of alternating edges that begins and ends with edges from different sets. This means that the cycle has to be of even length. ■

For example, a 2-coloring of the constraint graph in Figure 6.17 is shown in Figure 6.18. The solution to this graph-coloring problem provides a start on the packet routing problem. We can complete the routing in the two smaller Beneš networks by induction. With this insight in hand, the digression is over and we can now complete the proof of Theorem 6.3.2.

Proof of Theorem 6.3.2 (*continued*). Let π be an arbitrary permutation of $0, 1, \ldots, N-1$. Let G be the graph whose vertices are packet numbers $0, 1, \ldots, N-1$ and whose edges come from the union of these two sets:

$$E_1 ::= \{\{u, v\} \mid |u - v| = N/2\}, \text{ and}$$
$$E_2 ::= \{\{u, w\} \mid |\pi(u) - \pi(w)| = N/2\}.$$

Now any vertex, u, is incident to at most two edges: a unique edge $\{u, v\} \in E_1$ and a unique edge $\{u, w\} \in E_2$. So according to Lemma 6.3.3, there is a 2-coloring for

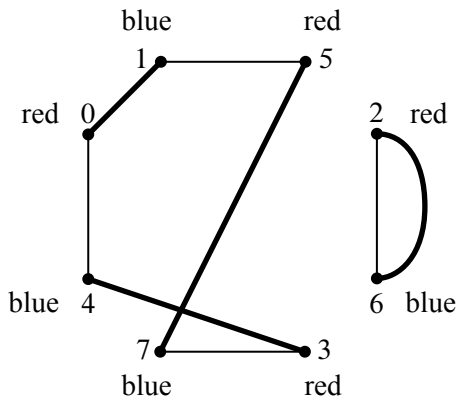


Figure 6.18 A 2-coloring of the constraint graph in Figure 6.17.

the vertices of G. Now route packets of one color through the upper subnetwork and packets of the other color through the lower subnetwork. Since for each edge in E_1, one vertex goes to the upper subnetwork and the other to the lower subnetwork, there will not be any conflicts in the first level. Since for each edge in E_2, one vertex comes from the upper subnetwork and the other from the lower subnetwork, there will not be any conflicts in the last level. We can complete the routing within each subnetwork by the induction hypothesis $P(n)$. ■

Chapter 12

MFCS 14: Events and Probability Spaces

14 Events and Probability Spaces

14.1 Let's Make a Deal

In the September 9, 1990 issue of *Parade* magazine, columnist Marilyn vos Savant responded to this letter:

> *Suppose you're on a game show, and you're given the choice of three doors. Behind one door is a car, behind the others, goats. You pick a door, say number 1, and the host, who knows what's behind the doors, opens another door, say number 3, which has a goat. He says to you, "Do you want to pick door number 2?" Is it to your advantage to switch your choice of doors?*
>
> Craig. F. Whitaker
> Columbia, MD

The letter describes a situation like one faced by contestants in the 1970's game show *Let's Make a Deal*, hosted by Monty Hall and Carol Merrill. Marilyn replied that the contestant should indeed switch. She explained that if the car was behind either of the two unpicked doors—which is twice as likely as the the car being behind the picked door—the contestant wins by switching. But she soon received a torrent of letters, many from mathematicians, telling her that she was wrong. The problem became known as the *Monty Hall Problem* and it generated thousands of hours of heated debate.

This incident highlights a fact about probability: the subject uncovers lots of examples where ordinary intuition leads to completely wrong conclusions. So until you've studied probabilities enough to have refined your intuition, a way to avoid errors is to fall back on a rigorous, systematic approach such as the Four Step Method that we will describe shortly. First, let's make sure we really understand the setup for this problem. This is always a good thing to do when you are dealing with probability.

14.1.1 Clarifying the Problem

Craig's original letter to Marilyn vos Savant is a bit vague, so we must make some assumptions in order to have any hope of modeling the game formally. For example, we will assume that:

1. The car is equally likely to be hidden behind each of the three doors.
2. The player is equally likely to pick each of the three doors, regardless of the car's location.
3. After the player picks a door, the host *must* open a different door with a goat behind it and offer the player the choice of staying with the original door or switching.
4. If the host has a choice of which door to open, then he is equally likely to select each of them.

In making these assumptions, we're reading a lot into Craig Whitaker's letter. Other interpretations are at least as defensible, and some actually lead to different answers. But let's accept these assumptions for now and address the question, "What is the probability that a player who switches wins the car?"

14.2 The Four Step Method

Every probability problem involves some sort of randomized experiment, process, or game. And each such problem involves two distinct challenges:

1. How do we model the situation mathematically?
2. How do we solve the resulting mathematical problem?

In this section, we introduce a four step approach to questions of the form, "What is the probability that…?" In this approach, we build a probabilistic model step-by-step, formalizing the original question in terms of that model. Remarkably, the structured thinking that this approach imposes provides simple solutions to many famously-confusing problems. For example, as you'll see, the four step method cuts through the confusion surrounding the Monty Hall problem like a Ginsu knife.

14.2.1 Step 1: Find the Sample Space

Our first objective is to identify all the possible outcomes of the experiment. A typical experiment involves several randomly-determined quantities. For example, the Monty Hall game involves three such quantities:

1. The door concealing the car.
2. The door initially chosen by the player.

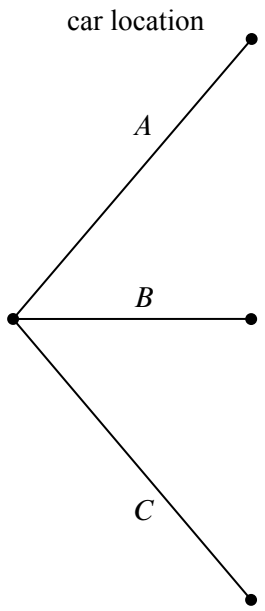


Figure 14.1 The first level in a tree diagram for the Monty Hall Problem. The branches correspond to the door behind which the car is located.

3. The door that the host opens to reveal a goat.

Every possible combination of these randomly-determined quantities is called an *outcome*. The set of all possible outcomes is called the *sample space* for the experiment.

A *tree diagram* is a graphical tool that can help us work through the four step approach when the number of outcomes is not too large or the problem is nicely structured. In particular, we can use a tree diagram to help understand the sample space of an experiment. The first randomly-determined quantity in our experiment is the door concealing the prize. We represent this as a tree with three branches, as shown in Figure 14.1. In this diagram, the doors are called A, B, and C instead of 1, 2, and 3, because we'll be adding a lot of other numbers to the picture later.

For each possible location of the prize, the player could initially choose any of the three doors. We represent this in a second layer added to the tree. Then a third layer represents the possibilities of the final step when the host opens a door to reveal a goat, as shown in Figure 14.2.

Notice that the third layer reflects the fact that the host has either one choice or two, depending on the position of the car and the door initially selected by the player. For example, if the prize is behind door A and the player picks door B, then

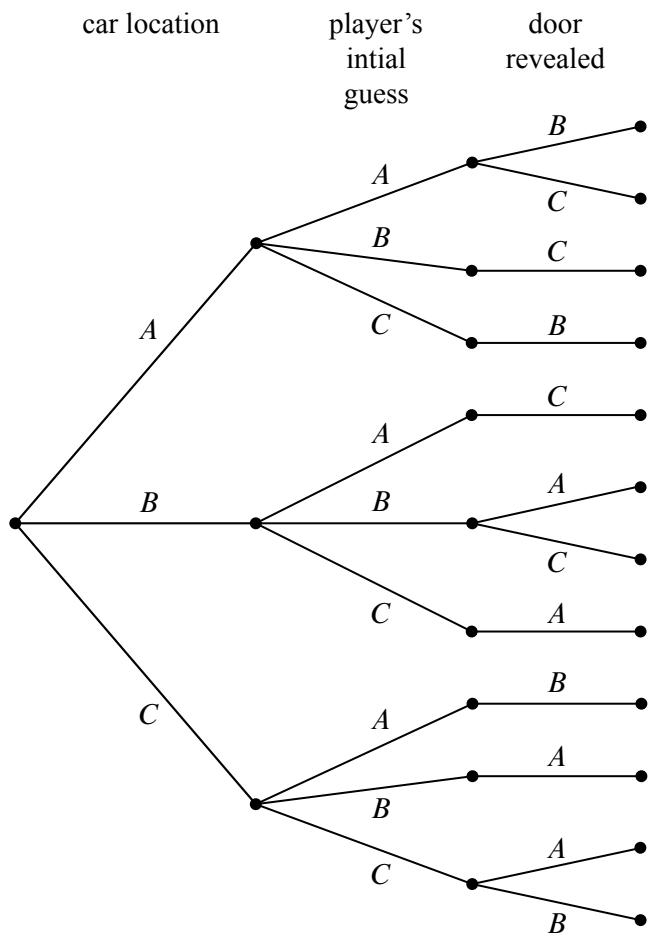


Figure 14.2 The full tree diagram for the Monty Hall Problem. The second level indicates the door initially chosen by the player. The third level indicates the door revealed by Monty Hall.

the host must open door C. However, if the prize is behind door A and the player picks door A, then the host could open either door B or door C.

Now let's relate this picture to the terms we introduced earlier: the leaves of the tree represent *outcomes* of the experiment, and the set of all leaves represents the *sample space*. Thus, for this experiment, the sample space consists of 12 outcomes. For reference, we've labeled each outcome in Figure 14.3 with a triple of doors indicating:

$$(\text{door concealing prize},\ \text{door initially chosen},\ \text{door opened to reveal a goat}).$$

In these terms, the sample space is the set

$$\mathcal{S} = \left\{ \begin{array}{l} (A, A, B),\ (A, A, C),\ (A, B, C),\ (A, C, B),\ (B, A, C),\ (B, B, A), \\ (B, B, C),\ (B, C, A),\ (C, A, B),\ (C, B, A),\ (C, C, A),\ (C, C, B) \end{array} \right\}$$

The tree diagram has a broader interpretation as well: we can regard the whole experiment as following a path from the root to a leaf, where the branch taken at each stage is "randomly" determined. Keep this interpretation in mind; we'll use it again later.

14.2.2 Step 2: Define Events of Interest

Our objective is to answer questions of the form "What is the probability that . . . ?", where, for example, the missing phrase might be "the player wins by switching", "the player initially picked the door concealing the prize", or "the prize is behind door C". Each of these phrases characterizes a set of outcomes. For example, the outcomes specified by "the prize is behind door C" is:

$$\{(C, A, B), (C, B, A), (C, C, A), (C, C, B)\}.$$

A set of outcomes is called an *event* and it is a subset of the sample space. So the event that the player initially picked the door concealing the prize is the set:

$$\{(A, A, B), (A, A, C), (B, B, A), (B, B, C), (C, C, A), (C, C, B)\}.$$

And what we're really after, the event that the player wins by switching, is the set of outcomes:

$$\{(A, B, C), (A, C, B), (B, A, C), (B, C, A), (C, A, B), (C, B, A)\}.$$

These outcomes are denoted with a check mark in Figure 14.4.

Notice that exactly half of the outcomes are checked, meaning that the player wins by switching in half of all outcomes. You might be tempted to conclude that a player who switches wins with probability $1/2$. *This is wrong.* The reason is that these outcomes are not all equally likely, as we'll see shortly.

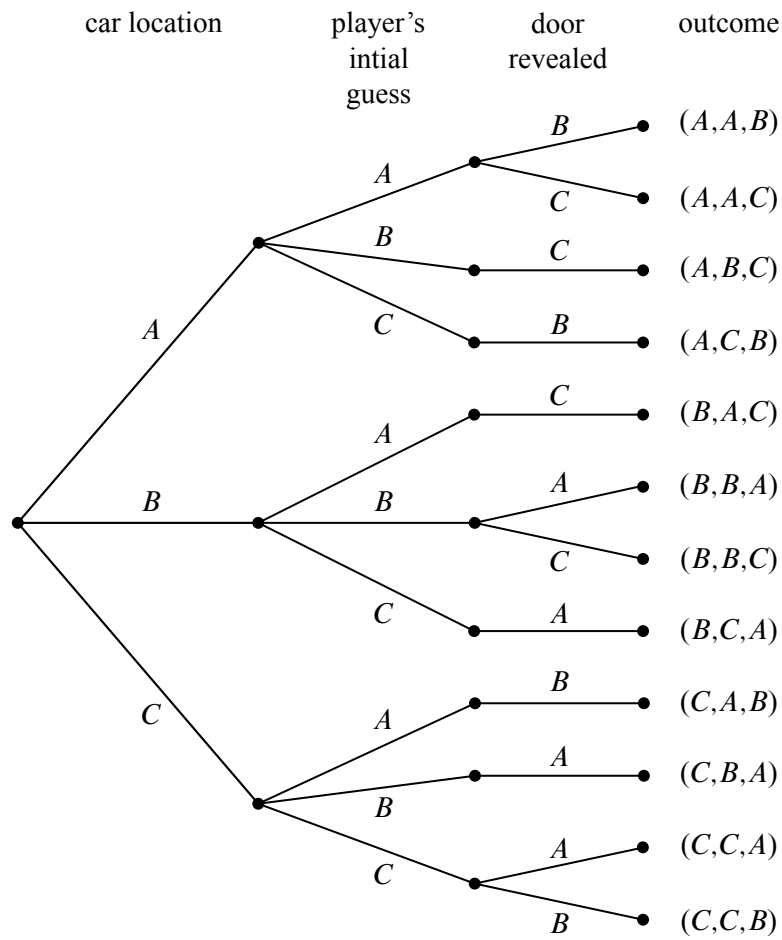


Figure 14.3 The tree diagram for the Monty Hal Problem with the outcomes labeled for each path from root to leaf. For example, outcome (A, A, B) corresponds to the car being behind door A, the player initially choosing door A, and Monty Hall revealing the goat behind door B.

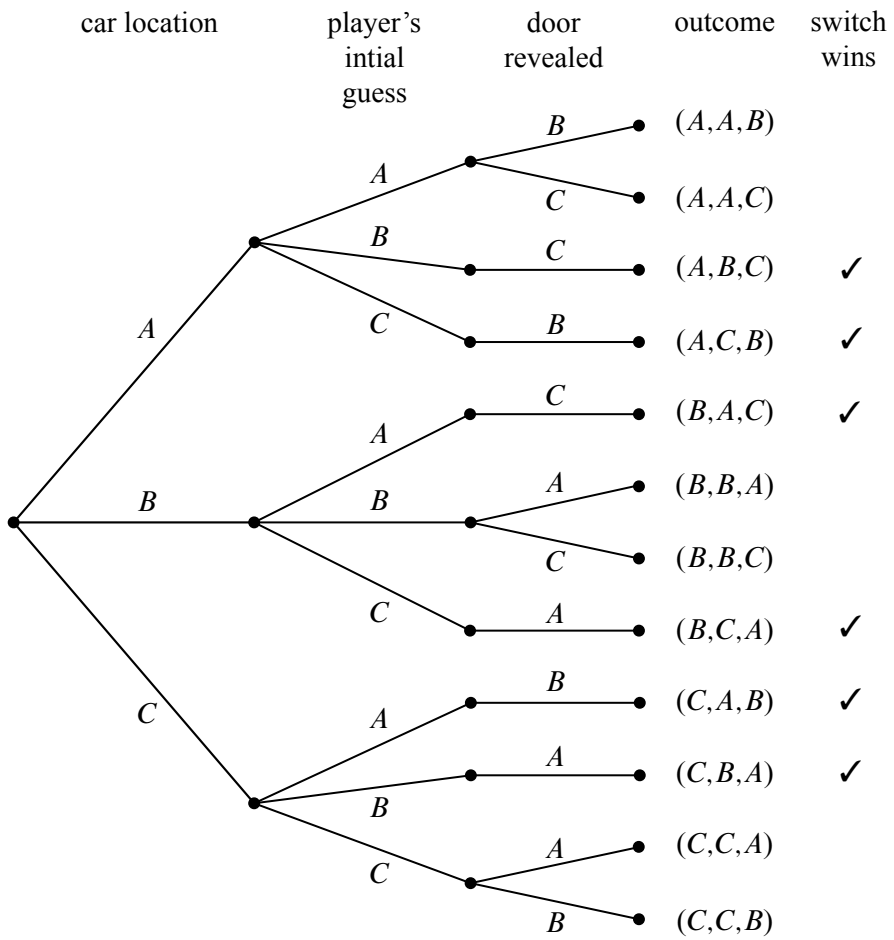


Figure 14.4 The tree diagram for the Monty Hall Problem where the outcomes in the event where the player wins by switching are denoted with a check mark.

14.2.3 Step 3: Determine Outcome Probabilities

So far we've enumerated all the possible outcomes of the experiment. Now we must start assessing the likelihood of those outcomes. In particular, the goal of this step is to assign each outcome a probability, indicating the fraction of the time this outcome is expected to occur. The sum of all outcome probabilities must be one, reflecting the fact that there always is an outcome.

Ultimately, outcome probabilities are determined by the phenomenon we're modeling and thus are not quantities that we can derive mathematically. However, mathematics can help us compute the probability of every outcome *based on fewer and more elementary modeling decisions.* In particular, we'll break the task of determining outcome probabilities into two stages.

Step 3a: Assign Edge Probabilities

First, we record a probability on each *edge* of the tree diagram. These edge-probabilities are determined by the assumptions we made at the outset: that the prize is equally likely to be behind each door, that the player is equally likely to pick each door, and that the host is equally likely to reveal each goat, if he has a choice. Notice that when the host has no choice regarding which door to open, the single branch is assigned probability 1. For example, see Figure 14.5.

Step 3b: Compute Outcome Probabilities

Our next job is to convert edge probabilities into outcome probabilities. This is a purely mechanical process: *the probability of an outcome is equal to the product of the edge-probabilities on the path from the root to that outcome*. For example, the probability of the topmost outcome in Figure 14.5, (A, A, B), is

$$\frac{1}{3} \cdot \frac{1}{3} \cdot \frac{1}{2} = \frac{1}{18}.$$

There's an easy, intuitive justification for this rule. As the steps in an experiment progress randomly along a path from the root of the tree to a leaf, the probabilities on the edges indicate how likely the path is to proceed along each branch. For example, a path starting at the root in our example is equally likely to go down each of the three top-level branches.

How likely is such a path to arrive at the topmost outcome, (A, A, B)? Well, there is a 1-in-3 chance that a path would follow the A-branch at the top level, a 1-in-3 chance it would continue along the A-branch at the second level, and 1-in-2 chance it would follow the B-branch at the third level. Thus, it seems that 1 path in 18 should arrive at the (A, A, B) leaf, which is precisely the probability we assign it.

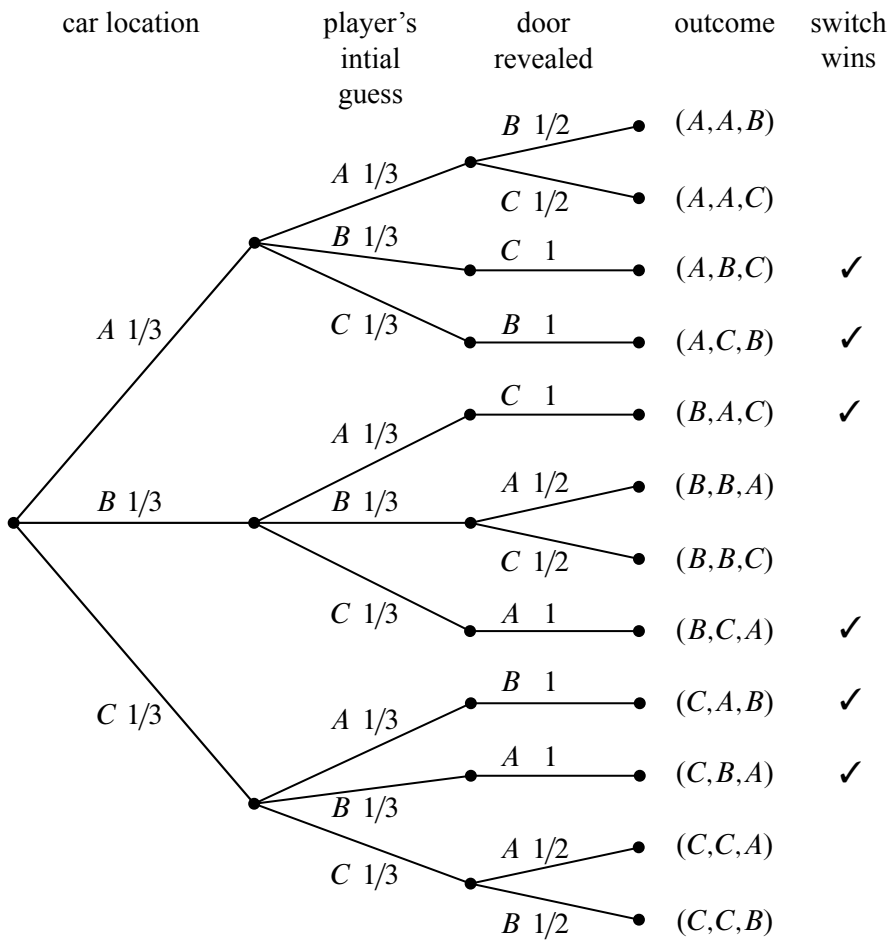


Figure 14.5 The tree diagram for the Monty Hall Problem where edge weights denote the probability of that branch being taken given that we are at the parent of that branch. For example, if the car is behind door A, then there is a 1/3 chance that the player's initial selection is door B.

We have illustrated all of the outcome probabilities in Figure 14.6.

Specifying the probability of each outcome amounts to defining a function that maps each outcome to a probability. This function is usually called **Pr**. In these terms, we've just determined that:

$$\begin{aligned}
\Pr[(A, A, B)] &= \frac{1}{18}, \\
\Pr[(A, A, C)] &= \frac{1}{18}, \\
\Pr[(A, B, C)] &= \frac{1}{9}, \\
&\text{etc.}
\end{aligned}$$

14.2.4 Step 4: Compute Event Probabilities

We now have a probability for each *outcome*, but we want to determine the probability of an *event*. The probability of an event E is denoted by $\Pr[E]$ and it is the sum of the probabilities of the outcomes in E. For example, the probability of the event that the player wins by switching is:[1]

$$\begin{aligned}
\Pr[\text{switching wins}] &= \Pr[(A, B, C)] + \Pr[(A, C, B)] + \Pr[(B, A, C)] + \\
&\quad \Pr[(B, C, A)] + \Pr[(C, A, B)] + \Pr[(C, B, A)] \\
&= \frac{1}{9} + \frac{1}{9} + \frac{1}{9} + \frac{1}{9} + \frac{1}{9} + \frac{1}{9} \\
&= \frac{2}{3}.
\end{aligned}$$

It seems Marilyn's answer is correct! A player who switches doors wins the car with probability $2/3$. In contrast, a player who stays with his or her original door wins with probability $1/3$, since staying wins if and only if switching loses.

We're done with the problem! We didn't need any appeals to intuition or ingenious analogies. In fact, no mathematics more difficult than adding and multiplying fractions was required. The only hard part was resisting the temptation to leap to an "intuitively obvious" answer.

14.2.5 An Alternative Interpretation of the Monty Hall Problem

Was Marilyn really right? Our analysis indicates that she was. But a more accurate conclusion is that her answer is correct *provided we accept her interpretation of the*

[1] "Switching wins" is shorthand for the set of outcomes where switching wins; namely, $\{(A, B, C), (A, C, B), (B, A, C), (B, C, A), (C, A, B), (C, B, A)\}$. We will frequently use such shorthand to denote events.

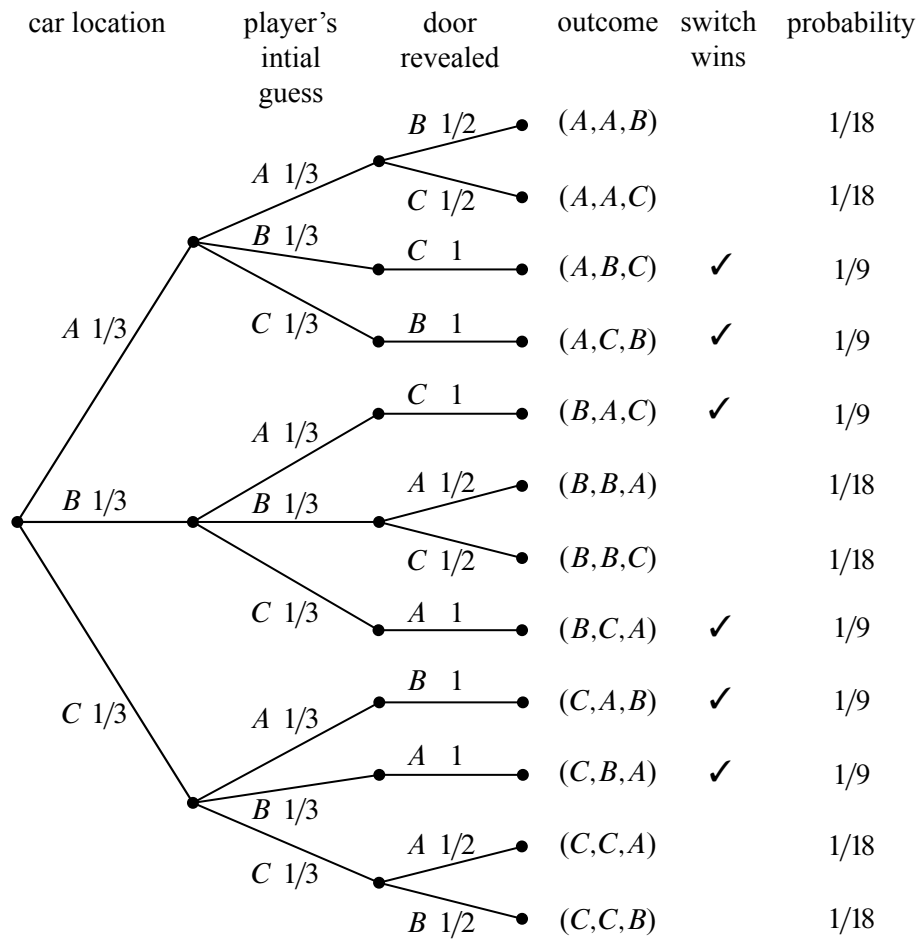
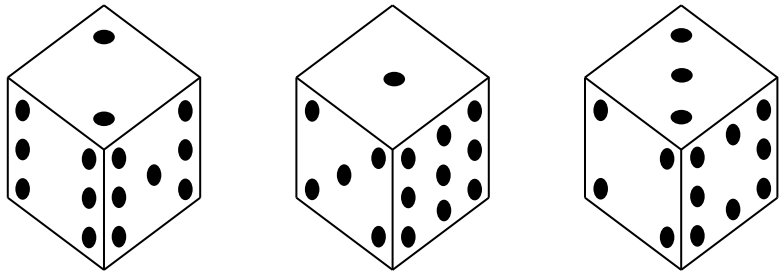


Figure 14.6 The rightmost column shows the outcome probabilities for the Monty Hall Problem. Each outcome probability is simply the product of the probabilities on the branches on the path from the root to the leaf for that outcome.



a *b* *c*

Figure 14.7 The strange dice. The number of pips on each concealed face is the same as the number on the opposite face. For example, when you roll die A, the probabilities of getting a 2, 6, or 7 are each $1/3$.

question. There is an equally plausible interpretation in which Marilyn's answer is wrong. Notice that Craig Whitaker's original letter does not say that the host is *required* to reveal a goat and offer the player the option to switch, merely that he *did* these things. In fact, on the *Let's Make a Deal* show, Monty Hall sometimes simply opened the door that the contestant picked initially. Therefore, if he wanted to, Monty could give the option of switching only to contestants who picked the correct door initially. In this case, switching never works!

14.3 Strange Dice

The four-step method is surprisingly powerful. Let's get some more practice with it. Imagine, if you will, the following scenario.

It's a typical Saturday night. You're at your favorite pub, contemplating the true meaning of infinite cardinalities, when a burly-looking biker plops down on the stool next to you. Just as you are about to get your mind around $\mathcal{P}(\mathcal{P}(\mathbb{R}))$, biker dude slaps three strange-looking dice on the bar and challenges you to a \$100 wager.

The rules are simple. Each player selects one die and rolls it once. The player with the lower value pays the other player \$100.

Naturally, you are skeptical. A quick inspection reveals that these are not ordinary dice. They each have six sides, but the numbers on the dice are different, as shown in Figure 14.7.

Biker dude notices your hesitation and so he offers to let you pick a die first, and

then he will choose his die from the two that are left. That seals the deal since you figure that you now have an advantage.

But which of the dice should you choose? Die B is appealing because it has a 9, which is a sure winner if it comes up. Then again, die A has two fairly large numbers and die B has an 8 and no really small values.

In the end, you choose die B because it has a 9, and then biker dude selects die A. Let's see what the probability is that you will win.[2] Not surprisingly, we will use the four-step method to compute this probability.

14.3.1 Die A versus Die B

Step 1: Find the sample space.
The sample space for this experiment is worked out in the tree diagram shown in Figure 14.8.[3]

For this experiment, the sample space is a set of nine outcomes:

$$\mathcal{S} = \{ (2,1),\ (2,5),\ (2,9),\ (6,1),\ (6,5),\ (6,9),\ (7,1),\ (7,5),\ (7,9) \}.$$

Step 2: Define events of interest.
We are interested in the event that the number on die A is greater than the number on die B. This event is a set of five outcomes:

$$\{ (2,1),\ (6,1),\ (6,5),\ (7,1),\ (7,5) \}.$$

These outcomes are marked A in the tree diagram in Figure 14.8.

Step 3: Determine outcome probabilities.
To find outcome probabilities, we first assign probabilities to edges in the tree diagram. Each number on each die comes up with probability $1/3$, regardless of the value of the other die. Therefore, we assign all edges probability $1/3$. The probability of an outcome is the product of the probabilities on the corresponding root-to-leaf path, which means that every outcome has probability $1/9$. These probabilities are recorded on the right side of the tree diagram in Figure 14.8.

Step 4: Compute event probabilities.
The probability of an event is the sum of the probabilities of the outcomes in that event. In this case, all the outcome probabilities are the same. In general, when the probability of every outcome is the same, we say that the sample space is *uniform*. Computing event probabilities for uniform sample spaces is particularly easy since

[2] Of course, you probably should have done this before picking die B in the first place.

[3] Actually, the whole probability space is worked out in this one picture. But pretend that each component sort of fades in—nyyrrroom!—as you read about the corresponding step below.

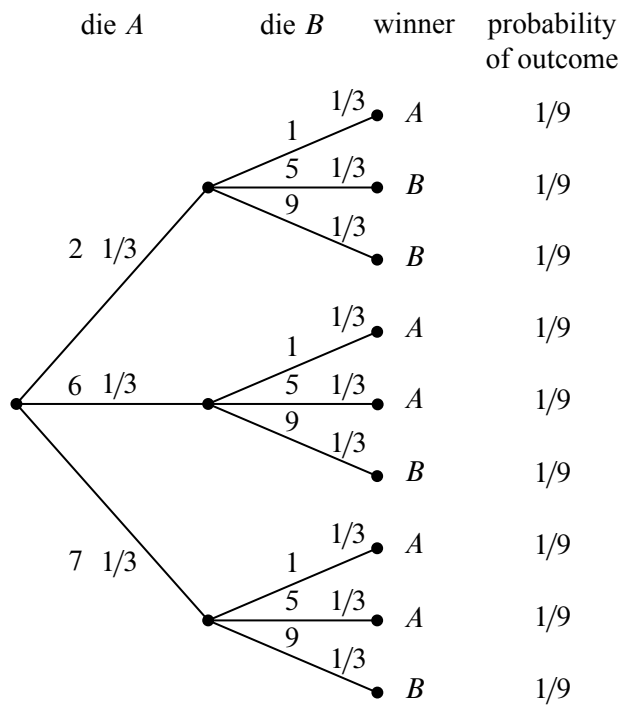


Figure 14.8 The tree diagram for one roll of die A versus die B. Die A wins with probability $5/9$.

you just have to compute the number of outcomes in the event. In particular, for any event E in a uniform sample space $\mathcal{S}$,

$$\Pr[E] = \frac{|E|}{|\mathcal{S}|}. \tag{14.1}$$

In this case, E is the event that die A beats die B, so $|E| = 5$, $|\mathcal{S}| = 9$, and

$$\Pr[E] = 5/9.$$

This is bad news for you. Die A beats die B more than half the time and, not surprisingly, you just lost \$100.

Biker dude consoles you on your "bad luck" and, given that he's a sensitive guy beneath all that leather, he offers to go double or nothing.[4] Given that your wallet only has \$25 in it, this sounds like a good plan. Plus, you figure that choosing die A will give *you* the advantage.

So you choose A, and then biker dude chooses C. Can you guess who is more likely to win? (Hint: it is generally not a good idea to gamble with someone you don't know in a bar, especially when you are gambling with strange dice.)

14.3.2 Die A versus Die C

We can construct the three diagram and outcome probabilities as before. The result is shown in Figure 14.9 and there is bad news again. Die C will beat die A with probability $5/9$, and you lose once again.

You now owe the biker dude \$200 and he asks for his money. You reply that you need to go to the bathroom.

Being a sensitive guy, biker dude nods understandingly and offers yet another wager. This time, he'll let you have die C. He'll even let you raise the wager to \$200 so you can win your money back.

This is too good a deal to pass up. You know that die C is likely to beat die A and that die A is likely to beat die B, and so die C is *surely* the best. Whether biker dude picks A or B, the odds are *surely* in your favor this time. Biker dude must really be a nice guy.

So you pick C, and then biker dude picks B. Let's use the tree method to figure out the probability that you win.

[4] *Double or nothing* is slang for doing another wager after you have lost the first. If you lose again, you will owe biker dude *double* what you owed him before. If you win, you will now be even and you will owe him *nothing*.

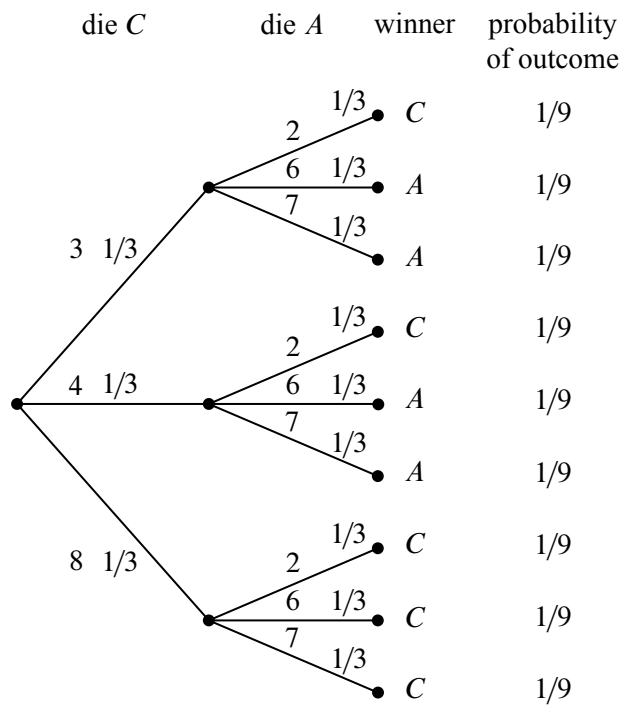


Figure 14.9 The tree diagram for one roll of die C versus die A. Die C wins with probability $5/9$.

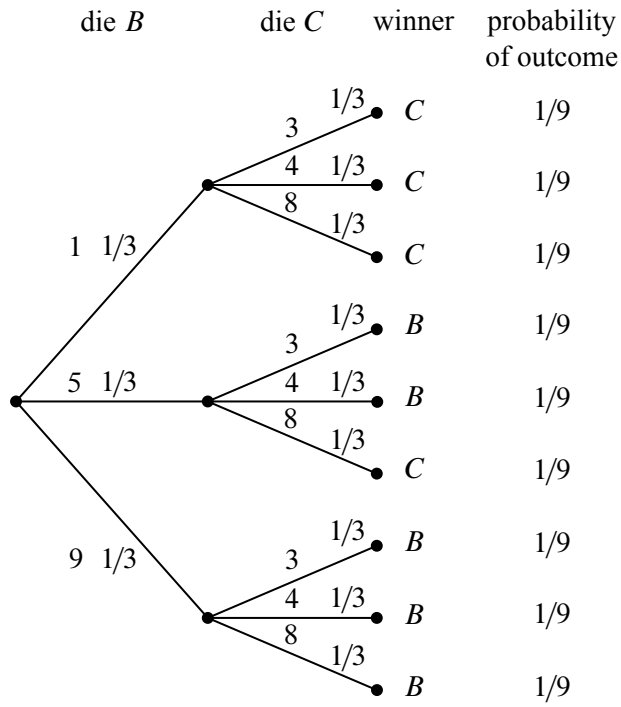


Figure 14.10 The tree diagram for one roll of die B versus die C. Die B wins with probability $5/9$.

14.3.3 Die B versus Die C

The tree diagram and outcome probabilities for B versus C are shown in Figure 14.10. But surely there is a mistake! The data in Figure 14.10 shows that *die* B wins with probability $5/9$. How is it possible that

> C beats A with probability $5/9$,
>
> A beats B with probability $5/9$, and
>
> B beats C with probability $5/9$?

The problem is not with the math, but with your intuition. It *seems* that the "likely-to-beat" relation should be transitive. But it is not, and whatever die you pick, biker dude can pick one of the others and be likely to win. So picking first is a big disadvantage and you now owe biker dude $400.

Just when you think matters can't get worse, biker dude offers you one final wager for $1,000. This time, you demand to choose second. Biker dude agrees,

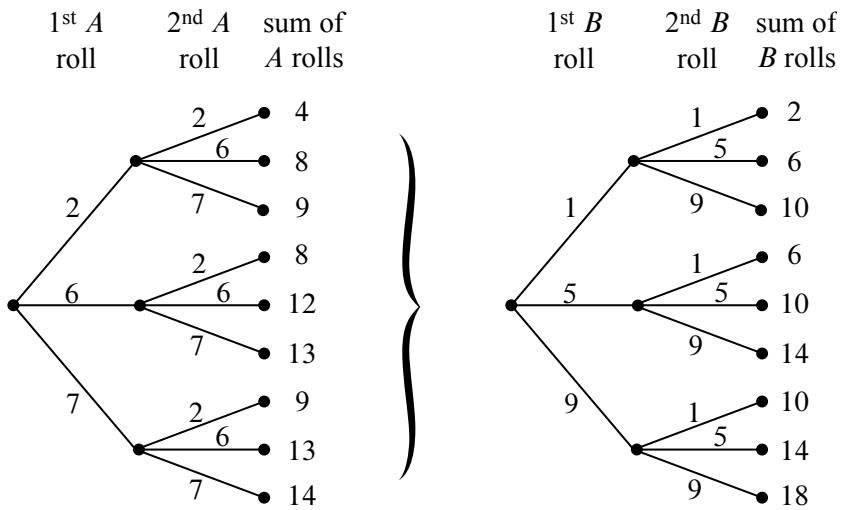


Figure 14.11 Parts of the tree diagram for die B versus die A where each die is rolled twice. The first two levels are shown in (a). The last two levels consist of nine copies of the tree in (b).

but with the condition that instead of rolling each die once, you each roll your die twice and your score is the sum of your rolls.

Believing that you finally have a winning wager, you agree.[5] Biker dude chooses die B and, of course, you grab die A. That's because you know that die A will beat die B with probability $5/9$ on one roll and so *surely* two rolls of die A are likely to beat two rolls of die B, right?

Wrong!

14.3.4 Rolling Twice

If each player rolls twice, the tree diagram will have four levels and $3^4 = 81$ outcomes. This means that it will take a while to write down the entire tree diagram. We can, however, easily write down the first two levels (as we have done in Figure 14.11(a)) and then notice that the remaining two levels consist of nine identical copies of the tree in Figure 14.11(b).

The probability of each outcome is $(1/3)^4 = 1/81$ and so, once again, we have a uniform probability space. By Equation 14.1, this means that the probability that A wins is the number of outcomes where A beats B divided by 81.

To compute the number of outcomes where A beats B, we observe that the sum

[5] Did we mention that playing strange gambling games with strangers in a bar is a bad idea?

of the two rolls of die A is equally likely to be any element of the following multiset:

$$\mathcal{S}_A = \{4, 8, 8, 9, 9, 12, 13, 13, 14\}.$$

The sum of two rolls of die B is equally likely to be any element of the following multiset:

$$\mathcal{S}_B = \{2, 6, 6, 10, 10, 10, 14, 14, 18\}.$$

We can treat each outcome as a pair $(x, y) \in \mathcal{S}_A \times \mathcal{S}_B$, where A wins iff $x > y$. If $x = 4$, there is only one y (namely $y = 2$) for which $x > y$. If $x = 8$, there are three values of y for which $x > y$. Continuing the count in this way, the number of pairs for which $x > y$ is

$$1 + 3 + 3 + 3 + 3 + 6 + 6 + 6 + 6 = 37.$$

A similar count shows that there are 42 pairs for which $x > y$, and there are two pairs ((14, 14), (14, 14)) which result in ties. This means that A *loses* to B with probability $42/81 > 1/2$ and ties with probability $2/81$. Die A wins with probability only $37/81$.

How can it be that A is more likely than B to win with 1 roll, but B is more likely to win with 2 rolls?!? Well, why not? The only reason we'd think otherwise is our (faulty) intuition. In fact, the die strength reverses no matter which two die we picked. So for 1 roll,

$$A \succ B \succ C \succ A,$$

but for two rolls,

$$A \prec B \prec C \prec A,$$

where we have used the symbols $\succ$ and $\prec$ to denote which die is more likely to result in the larger value. This is surprising even to us, but at least we don't owe biker dude $1400.

14.3.5 Even Stranger Dice

Now that we know that strange things can happen with strange dice, it is natural, at least for mathematicians, to ask how strange things can get. It turns out that things can get very strange. In fact, mathematicians[6] recently made the following discovery:

Theorem 14.3.1. *For any $n \geq 2$, there is a set of n dice $D_1, D_2, \ldots, D_n$ such that for any n-node tournament graph[7] G, there is a number of rolls k such that if each*

[6]Reference Ron Graham paper.

[7]Recall that a tournament graph is a directed graph for which there is precisely one directed edge between any two distinct nodes. In other words, for every pair of distinct nodes u and v, either u beats v or v beats u, but not both.

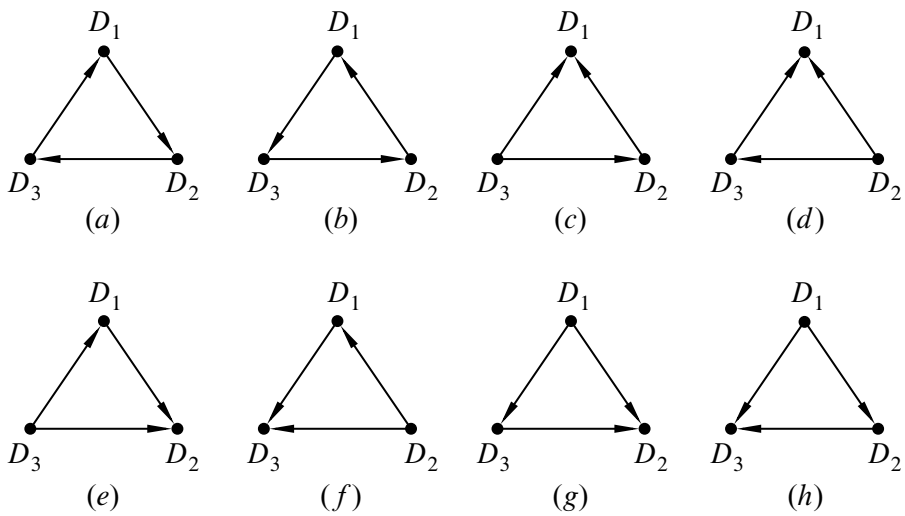


Figure 14.12 All possible relative strengths for three dice D_1, D_2, and D_3. The edge $D_i \to D_j$ denotes that the sum of rolls for D_i is likely to be greater than the sum of rolls for D_j.

die is rolled k times, then for all $i \neq j$, the sum of the k rolls for D_i will exceed the sum for D_j with probability greater than $1/2$ iff $D_i \to D_j$ is in G.

It will probably take a few attempts at reading Theorem 14.3.1 to understand what it is saying. The idea is that for some sets of dice, by rolling them different numbers of times, the dice have varying strengths relative to each other. (This is what we observed for the dice in Figure 14.7.) Theorem 14.3.1 says that there is a set of (very) strange dice where *every* possible collection of relative strengths can be observed by varying the number of rolls. For example, the eight possible relative strengths for $n = 3$ dice are shown in Figure 14.12.

Our analysis for the dice in Figure 14.7 showed that for 1 roll, we have the relative strengths shown in Figure 14.12(a), and for two rolls, we have the (reverse) relative strengths shown in Figure 14.12(b). Can you figure out what other relative strengths are possible for the dice in Figure 14.7 by using more rolls? This might be worth doing if you are prone to gambling with strangers in bars.

14.4 Set Theory and Probability

The study of probability is very closely tied to set theory. That is because any set can be a sample space and any subset can be an event. This means that most of the rules and identities that we have developed for sets extend very naturally to probability. We'll cover several examples in this section, but first let's review some definitions that should already be familiar.

14.4.1 Probability Spaces

Definition 14.4.1. A countable[8] *sample space* $\mathcal{S}$ is a nonempty countable set. An element $w \in \mathcal{S}$ is called an *outcome*. A subset of $\mathcal{S}$ is called an *event*.

Definition 14.4.2. A *probability function* on a sample space $\mathcal{S}$ is a total function $\Pr : \mathcal{S} \to \mathbb{R}$ such that

- $\Pr[w] \geq 0$ for all $w \in \mathcal{S}$, and
- $\sum_{w \in \mathcal{S}} \Pr[w] = 1$.

A sample space together with a probability function is called a *probability space*. For any event $E \subseteq \mathcal{S}$, the *probability of* E is defined to be the sum of the probabilities of the outcomes in E:

$$\Pr[E] ::= \sum_{w \in E} \Pr[w].$$

14.4.2 Probability Rules from Set Theory

An immediate consequence of the definition of event probability is that for *disjoint* events E and F,

$$\Pr[E \cup F] = \Pr[E] + \Pr[F].$$

This generalizes to a countable number of events, as follows.

Rule 14.4.3 (Sum Rule)**.** *If* $\{E_0, E_1, \ldots\}$ *is collection of disjoint events, then*

$$\Pr\left[\bigcup_{n \in \mathbb{N}} E_n\right] = \sum_{n \in \mathbb{N}} \Pr[E_n].$$

[8] Yes, sample spaces can be infinite. We'll see some examples shortly. If you did not read Chapter 13, don't worry—*countable* means that you can list the elements of the sample space as w_1, w_2, $w_3, \ldots$.

The Sum Rule lets us analyze a complicated event by breaking it down into simpler cases. For example, if the probability that a randomly chosen MIT student is native to the United States is 60%, to Canada is 5%, and to Mexico is 5%, then the probability that a random MIT student is native to North America is 70%.

Another consequence of the Sum Rule is that $\Pr[A] + \Pr[\overline{A}] = 1$, which follows because $\Pr[\mathcal{S}] = 1$ and $\mathcal{S}$ is the union of the disjoint sets A and $\overline{A}$. This equation often comes up in the form:

Rule 14.4.4 (Complement Rule)**.**

$$\Pr[\overline{A}] = 1 - \Pr[A].$$

Sometimes the easiest way to compute the probability of an event is to compute the probability of its complement and then apply this formula.

Some further basic facts about probability parallel facts about cardinalities of finite sets. In particular:

$$\begin{aligned}
&\Pr[B - A] = \Pr[B] - \Pr[A \cap B], && \text{(Difference Rule)}\\
&\Pr[A \cup B] = \Pr[A] + \Pr[B] - \Pr[A \cap B], && \text{(Inclusion-Exclusion)}\\
&\Pr[A \cup B] \leq \Pr[A] + \Pr[B], && \text{(Boole's Inequality}\\
&\text{If } A \subseteq B, \text{ then } \Pr[A] \leq \Pr[B]. && \text{(Monotonicity)}
\end{aligned}$$

The Difference Rule follows from the Sum Rule because B is the union of the disjoint sets $B - A$ and $A \cap B$. Inclusion-Exclusion then follows from the Sum and Difference Rules, because $A \cup B$ is the union of the disjoint sets A and $B - A$. Boole's inequality is an immediate consequence of Inclusion-Exclusion since probabilities are nonnegative. Monotonicity follows from the definition of event probability and the fact that outcome probabilities are nonnegative.

The two-event Inclusion-Exclusion equation above generalizes to n events in the same way as the corresponding Inclusion-Exclusion rule for n sets. Boole's inequality also generalizes to

$$\Pr[E_1 \cup \cdots \cup E_n] \leq \Pr[E_1] + \cdots + \Pr[E_n]. \qquad \text{(Union Bound)}$$

This simple Union Bound is useful in many calculations. For example, suppose that E_i is the event that the i-th critical component in a spacecraft fails. Then $E_1 \cup \cdots \cup E_n$ is the event that *some* critical component fails. If $\sum_{i=1}^{n} \Pr[E_i]$ is small, then the Union Bound can give an adequate upper bound on this vital probability.

14.4.3 Uniform Probability Spaces

Definition 14.4.5. A finite probability space $\mathcal{S}$, Pr is said to be *uniform* if $\Pr[w]$ is the same for every outcome $w \in \mathcal{S}$.

As we saw in the strange dice problem, uniform sample spaces are particularly easy to work with. That's because for any event $E \subseteq \mathcal{S}$,

$$\Pr[E] = \frac{|E|}{|\mathcal{S}|}. \tag{14.2}$$

This means that once we know the cardinality of E and $\mathcal{S}$, we can immediately obtain $\Pr[E]$. That's great news because we developed lots of tools for computing the cardinality of a set in Part III.

For example, suppose that you select five cards at random from a standard deck of 52 cards. What is the probability of having a full house? Normally, this question would take some effort to answer. But from the analysis in Section 11.7.2, we know that

$$|\mathcal{S}| = \binom{13}{5}$$

and

$$|E| = 13 \cdot \binom{4}{3} \cdot 12 \cdot \binom{4}{2}$$

where E is the event that we have a full house. Since every five-card hand is equally likely, we can apply Equation 14.2 to find that

$$\begin{aligned}
\Pr[E] &= \frac{13 \cdot 12 \cdot \binom{4}{3} \cdot \binom{4}{2}}{\binom{13}{5}} \\
&= \frac{13 \cdot 12 \cdot 4 \cdot 6 \cdot 5 \cdot 4 \cdot 3 \cdot 2}{52 \cdot 51 \cdot 50 \cdot 49 \cdot 48} \\
&= \frac{18}{12495} \\
&\approx \frac{1}{694}.
\end{aligned}$$

14.5 Infinite Probability Spaces

General probability theory deals with uncountable sets like $\mathbb{R}$, but in computer science, it is usually sufficient to restrict our attention to countable probability spaces.

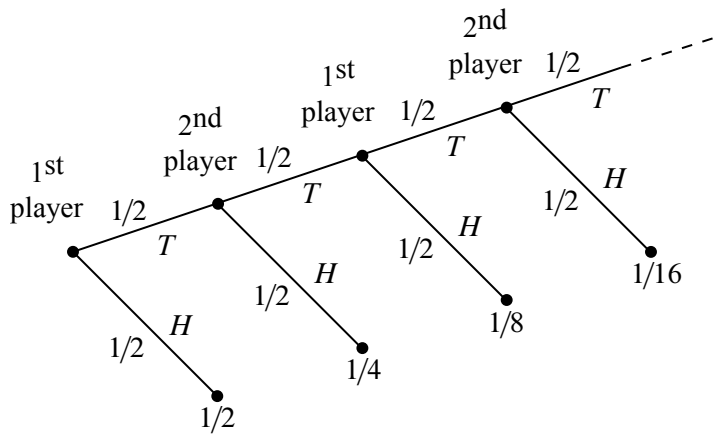


Figure 14.13 The tree diagram for the game where players take turns flipping a fair coin. The first player to flip heads wins.

It's also a lot easier—infinite sample spaces are hard enough to work with without having to deal with uncountable spaces.

Infinite probability spaces are fairly common. For example, two players take turns flipping a fair coin. Whoever flips heads first is declared the winner. What is the probability that the first player wins? A tree diagram for this problem is shown in Figure 14.13.

The event that the first player wins contains an infinite number of outcomes, but we can still sum their probabilities:

$$\begin{aligned}\Pr[\text{first player wins}] &= \frac{1}{2} + \frac{1}{8} + \frac{1}{32} + \frac{1}{128} + \cdots \\ &= \frac{1}{2}\sum_{n=0}^{\infty}\left(\frac{1}{4}\right)^n \\ &= \frac{1}{2}\left(\frac{1}{1-1/4}\right) = \frac{2}{3}.\end{aligned}$$

Similarly, we can compute the probability that the second player wins:

$$\Pr[\text{second player wins}] = \frac{1}{4} + \frac{1}{16} + \frac{1}{64} + \frac{1}{256} + \cdots = \frac{1}{3}.$$

In this case, the sample space is the infinite set

$$\mathcal{S} ::= \{\, \mathrm{T}^n\mathrm{H} \mid n \in \mathbb{N} \,\},$$

where T^n stands for a length n string of T's. The probability function is

$$\Pr[T^n H] ::= \frac{1}{2^{n+1}}.$$

To verify that this is a probability space, we just have to check that all the probabilities are nonnegative and that they sum to 1. Nonnegativity is obvious, and applying the formula for the sum of a geometric series, we find that

$$\sum_{n \in \mathbb{N}} \Pr[T^n H] = \sum_{n \in \mathbb{N}} \frac{1}{2^{n+1}} = 1.$$

Notice that this model does not have an outcome corresponding to the possibility that both players keep flipping tails forever.[9] That's because the probability of flipping forever would be

$$\lim_{n \to \infty} \frac{1}{2^{n+1}} = 0,$$

and outcomes with probability zero will have no impact on our calculations.

[9]In the diagram, flipping forever corresponds to following the infinite path in the tree without ever reaching a leaf or outcome. Some texts deal with this case by adding a special "infinite" sample point w_{forever} to the sample space, but we will follow the more traditional approach of excluding such sample points, as long as they collectively have probability 0.

Chapter 13

MFCS 15: Conditional Probability

15 Conditional Probability

15.1 Definition

Suppose that we pick a random person in the world. Everyone has an equal chance of being selected. Let A be the event that the person is an MIT student, and let B be the event that the person lives in Cambridge. What are the probabilities of these events? Intuitively, we're picking a random point in the big ellipse shown in Figure 15.1 and asking how likely that point is to fall into region A or B.

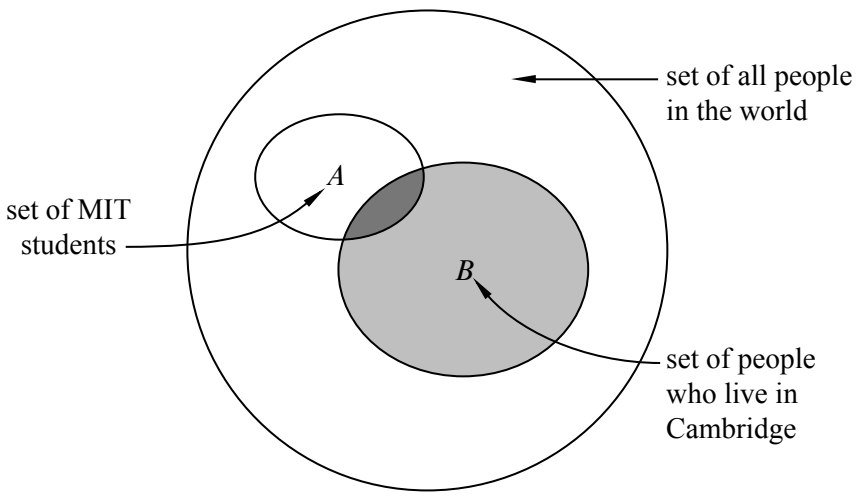


Figure 15.1 Selecting a random person. A is the event that the person is an MIT student. B is the even that the person lives in Cambridge.

The vast majority of people in the world neither live in Cambridge nor are MIT students, so events A and B both have low probability. But what about the probability that a person is an MIT student, *given* that the person lives in Cambridge? This should be much greater—but what is it exactly?

What we're asking for is called a *conditional probability*; that is, the probability that one event happens, given that some other event definitely happens. Questions about conditional probabilities come up all the time:

- What is the probability that it will rain this afternoon, given that it is cloudy this morning?

- What is the probability that two rolled dice sum to 10, given that both are odd?
- What is the probability that I'll get four-of-a-kind in Texas No Limit Hold 'Em Poker, given that I'm initially dealt two queens?

There is a special notation for conditional probabilities. In general, $\Pr\left[A \mid B\right]$ denotes the probability of event A, given that event B happens. So, in our example, $\Pr\left[A \mid B\right]$ is the probability that a random person is an MIT student, given that he or she is a Cambridge resident.

How do we compute $\Pr\left[A \mid B\right]$? Since we are *given* that the person lives in Cambridge, we can forget about everyone in the world who does not. Thus, all outcomes outside event B are irrelevant. So, intuitively, $\Pr\left[A \mid B\right]$ should be the fraction of Cambridge residents that are also MIT students; that is, the answer should be the probability that the person is in set $A \cap B$ (the darkly shaded region in Figure 15.1) divided by the probability that the person is in set B (the lightly shaded region). This motivates the definition of conditional probability:

Definition 15.1.1.

$$\Pr\left[A \mid B\right] ::= \frac{\Pr[A \cap B]}{\Pr[B]}$$

If $\Pr[B] = 0$, then the conditional probability $\Pr\left[A \mid B\right]$ is undefined.

Pure probability is often counterintuitive, but conditional probability is even worse! Conditioning can subtly alter probabilities and produce unexpected results in randomized algorithms and computer systems as well as in betting games. Yet, the mathematical definition of conditional probability given above is very simple and should give you no trouble—provided that you rely on formal reasoning and not intuition. The four-step method will also be very helpful as we will see in the next examples.

15.2 Using the Four-Step Method to Determine Conditional Probability

15.2.1 The "Halting Problem"

The *Halting Problem* was the first example of a property that could not be tested by any program. It was introduced by Alan Turing in his seminal 1936 paper. The problem is to determine whether a Turing machine halts on a given . . . yadda yadda

yadda . . . more importantly, it was the name of the MIT EECS department's famed C-league hockey team.

In a best-of-three tournament, the Halting Problem wins the first game with probability $1/2$. In subsequent games, their probability of winning is determined by the outcome of the previous game. If the Halting Problem won the previous game, then they are invigorated by victory and win the current game with probability $2/3$. If they lost the previous game, then they are demoralized by defeat and win the current game with probability only $1/3$. What is the probability that the Halting Problem wins the tournament, given that they win the first game?

This is a question about a conditional probability. Let A be the event that the Halting Problem wins the tournament, and let B be the event that they win the first game. Our goal is then to determine the conditional probability $\Pr\left[A \mid B\right]$.

We can tackle conditional probability questions just like ordinary probability problems: using a tree diagram and the four step method. A complete tree diagram is shown in Figure 15.2.

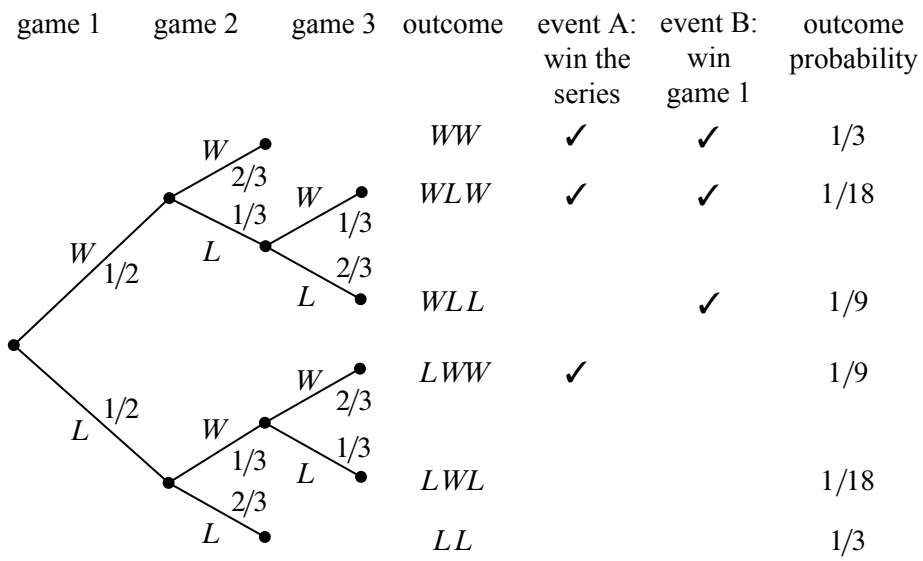


Figure 15.2 The tree diagram for computing the probability that the "Halting Problem" wins two out of three games given that they won the first game.

Step 1: Find the Sample Space

Each internal vertex in the tree diagram has two children, one corresponding to a win for the Halting Problem (labeled W) and one corresponding to a loss (la-

beled L). The complete sample space is:

$$\mathcal{S} = \{WW,\ WLW,\ WLL,\ LWW,\ LWL,\ LL\}.$$

Step 2: Define Events of Interest
The event that the Halting Problem wins the whole tournament is:

$$T = \{WW,\ WLW,\ LWW\}.$$

And the event that the Halting Problem wins the first game is:

$$F = \{WW,\ WLW,\ WLL\}.$$

The outcomes in these events are indicated with check marks in the tree diagram in Figure 15.2.

Step 3: Determine Outcome Probabilities
Next, we must assign a probability to each outcome. We begin by labeling edges as specified in the problem statement. Specifically, The Halting Problem has a $1/2$ chance of winning the first game, so the two edges leaving the root are each assigned probability $1/2$. Other edges are labeled $1/3$ or $2/3$ based on the outcome of the preceding game. We then find the probability of each outcome by multiplying all probabilities along the corresponding root-to-leaf path. For example, the probability of outcome WLL is:

$$\frac{1}{2} \cdot \frac{1}{3} \cdot \frac{2}{3} = \frac{1}{9}.$$

Step 4: Compute Event Probabilities

We can now compute the probability that The Halting Problem wins the tournament, given that they win the first game:

$$\begin{aligned}
\Pr\left[A \mid B\right] &= \frac{\Pr[A \cap B]}{\Pr[B]} \\
&= \frac{\Pr[\{WW, WLW\}]}{\Pr[\{WW, WLW, WLL\}]} \\
&= \frac{1/3 + 1/18}{1/3 + 1/18 + 1/9} \\
&= \frac{7}{9}.
\end{aligned}$$

We're done! If the Halting Problem wins the first game, then they win the whole tournament with probability $7/9$.

15.2.2 Why Tree Diagrams Work

We've now settled into a routine of solving probability problems using tree diagrams. But we've left a big question unaddressed: what is the mathematical justification behind those funny little pictures? Why do they work?

The answer involves conditional probabilities. In fact, the probabilities that we've been recording on the edges of tree diagrams *are* conditional probabilities. For example, consider the uppermost path in the tree diagram for the Halting Problem, which corresponds to the outcome WW. The first edge is labeled $1/2$, which is the probability that the Halting Problem wins the first game. The second edge is labeled $2/3$, which is the probability that the Halting Problem wins the second game, *given* that they won the first—that's a conditional probability! More generally, on each edge of a tree diagram, we record the probability that the experiment proceeds along that path, given that it reaches the parent vertex.

So we've been using conditional probabilities all along. But why can we multiply edge probabilities to get outcome probabilities? For example, we concluded that:

$$\Pr[WW] = \frac{1}{2} \cdot \frac{2}{3} = \frac{1}{3}.$$

Why is this correct?

The answer goes back to Definition 15.1.1 of conditional probability which could be written in a form called the *Product Rule* for probabilities:

Rule (Product Rule for 2 Events)**.** *If* $\Pr[E_1] \neq 0$*, then:*

$$\Pr[E_1 \cap E_2] = \Pr[E_1] \cdot \Pr\left[E_2 \mid E_1\right].$$

Multiplying edge probabilities in a tree diagram amounts to evaluating the right side of this equation. For example:

$$\begin{aligned}
&\Pr[\text{win first game} \cap \text{win second game}] \\
&\qquad = \Pr[\text{win first game}] \cdot \Pr\left[\text{win second game} \mid \text{win first game}\right] \\
&\qquad = \frac{1}{2} \cdot \frac{2}{3}.
\end{aligned}$$

So the Product Rule is the formal justification for multiplying edge probabilities to get outcome probabilities! Of course to justify multiplying edge probabilities along longer paths, we need a Product Rule for n events.

Rule (Product Rule for n Events)**.**

$$\begin{aligned}
\Pr[E_1 \cap E_2 \cap \ldots \cap E_n] = \Pr[E_1] \cdot \Pr\left[E_2 \mid E_1\right] \cdot \Pr\left[E_3 \mid E_1 \cap E_2\right] \cdots \\
\cdot \Pr\left[E_n \mid E_1 \cap E_2 \cap \ldots \cap E_{n-1}\right]
\end{aligned}$$

provided that

$$\Pr[E_1 \cap E_2 \cap \cdots \cap E_{n-1}] \neq 0.$$

This rule follows from the definition of conditional probability and induction on n.

15.2.3 Medical Testing

There is an unpleasant condition called *BO* suffered by 10% of the population. There are no prior symptoms; victims just suddenly start to stink. Fortunately, there is a test for latent *BO* before things start to smell. The test is not perfect, however:

- If you have the condition, there is a 10% chance that the test will say you do not. These are called "false negatives".
- If you do not have the condition, there is a 30% chance that the test will say you do. These are "false positives".

Suppose a random person is tested for latent *BO*. If the test is positive, then what is the probability that the person has the condition?

Step 1: Find the Sample Space

The sample space is found with the tree diagram in Figure 15.3.

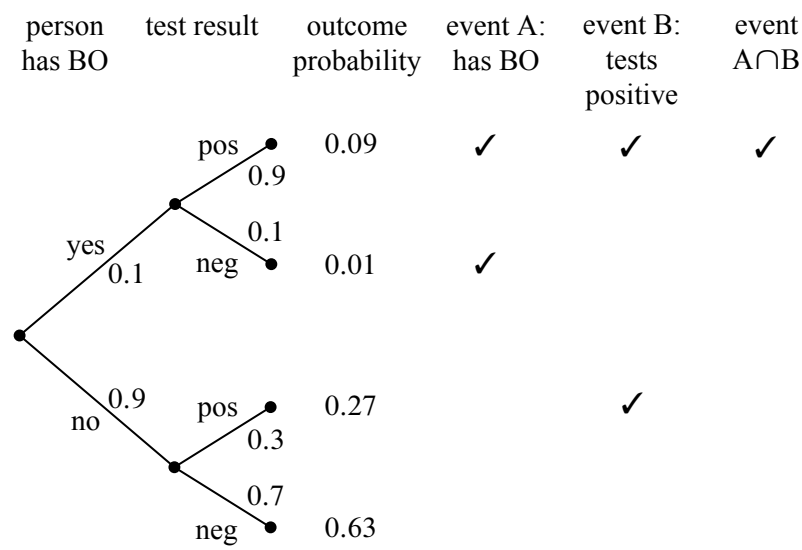


Figure 15.3 The tree diagram for the BO problem.

Step 2: Define Events of Interest

Let A be the event that the person has BO. Let B be the event that the test was positive. The outcomes in each event are marked in the tree diagram. We want to find $\Pr\left[A \mid B\right]$, the probability that a person has BO, given that the test was positive.

Step 3: Find Outcome Probabilities

First, we assign probabilities to edges. These probabilities are drawn directly from the problem statement. By the Product Rule, the probability of an outcome is the product of the probabilities on the corresponding root-to-leaf path. All probabilities are shown in Figure 15.3.

Step 4: Compute Event Probabilities

From Definition 15.1.1, we have

$$\Pr\left[A \mid B\right] = \frac{\Pr[A \cap B]}{\Pr[B]} = \frac{0.09}{0.09 + 0.27} = \frac{1}{4}.$$

So, if you test positive, then there is only a 25% chance that you have the condition!

This answer is initially surprising, but makes sense on reflection. There are two ways you could test positive. First, it could be that you have the condition and the test is correct. Second, it could be that you are healthy and the test is incorrect. The problem is that almost everyone is healthy; therefore, most of the positive results arise from incorrect tests of healthy people!

We can also compute the probability that the test is correct for a random person. This event consists of two outcomes. The person could have the condition and test positive (probability 0.09), or the person could be healthy and test negative (probability 0.63). Therefore, the test is correct with probability $0.09 + 0.63 = 0.72$. This is a relief; the test is correct almost three-quarters of the time.

But wait! There is a simple way to make the test correct 90% of the time: always return a negative result! This "test" gives the right answer for all healthy people and the wrong answer only for the 10% that actually have the condition. So a better strategy by this measure is to completely ignore the test result!

There is a similar paradox in weather forecasting. During winter, almost all days in Boston are wet and overcast. Predicting miserable weather every day may be more accurate than really trying to get it right!

Chapter 14

MFCS 16: Independence

16 Independence

16.1 Definitions

Suppose that we flip two fair coins simultaneously on opposite sides of a room. Intuitively, the way one coin lands does not affect the way the other coin lands. The mathematical concept that captures this intuition is called *independence*:

Definition 16.1.1. Events A and B are independent if $\Pr[B] = 0$ or if

$$\Pr\left[A \mid B\right] = \Pr[A]. \tag{16.1}$$

In other words, A and B are independent if knowing that B happens does not alter the probability that A happens, as is the case with flipping two coins on opposite sides of a room.

16.1.1 Potential Pitfall

Students sometimes get the idea that disjoint events are independent. The *opposite* is true: if $A \cap B = \emptyset$, then knowing that A happens means you know that B does not happen. So disjoint events are *never* independent—unless one of them has probability zero.

16.1.2 Alternative Formulation

Sometimes it is useful to express independence in an alternate form:

Theorem 16.1.2. *A and B are independent if and only if*

$$\Pr[A \cap B] = \Pr[A] \cdot \Pr[B]. \tag{16.2}$$

Proof. There are two cases to consider depending on whether or not $\Pr[B] = 0$.

Case 1 ($\Pr[B] = 0$)**:** If $\Pr[B] = 0$, A and B are independent by Definition 16.1.1. In addition, Equation 16.2 holds since both sides are 0. Hence, the theorem is true in this case.

Case 2 ($\Pr[B] > 0$)**:** By Definition 15.1.1,

$$\Pr[A \cap B] = \Pr\left[A \mid B\right] \Pr[B].$$

So Equation 16.2 holds if

$$\Pr\left[A \mid B\right] = \Pr[A],$$

which, by Definition 16.1.1, is true iff A and B are independent. Hence, the theorem is true in this case as well. ■

16.2 Independence Is an Assumption

Generally, independence is something that you *assume* in modeling a phenomenon. For example, consider the experiment of flipping two fair coins. Let A be the event that the first coin comes up heads, and let B be the event that the second coin is heads. If we assume that A and B are independent, then the probability that both coins come up heads is:

$$\Pr[A \cap B] = \Pr[A] \cdot \Pr[B] = \frac{1}{2} \cdot \frac{1}{2} = \frac{1}{4}.$$

In this example, the assumption of independence is reasonable. The result of one coin toss should have negligible impact on the outcome of the other coin toss. And if we were to repeat the experiment many times, we would be likely to have $A \cap B$ about 1/4 of the time.

There are, of course, many examples of events where assuming independence is *not* justified, For example, let C be the event that tomorrow is cloudy and R be the event that tomorrow is rainy. Perhaps $\Pr[C] = 1/5$ and $\Pr[R] = 1/10$ in Boston. If these events were independent, then we could conclude that the probability of a rainy, cloudy day was quite small:

$$\Pr[R \cap C] = \Pr[R] \cdot \Pr[C] = \frac{1}{5} \cdot \frac{1}{10} = \frac{1}{50}.$$

Unfortunately, these events are definitely not independent; in particular, every rainy day is cloudy. Thus, the probability of a rainy, cloudy day is actually $1/10$.

Deciding when to *assume* that events are independent is a tricky business. In practice, there are strong motivations to assume independence since many useful formulas (such as Equation 16.2) only hold if the events are independent. But you need to be careful lest you end up deriving false conclusions. We'll see several famous examples where (false) assumptions of independence led to trouble over the next several chapters. This problem gets even trickier when there are more than two events in play.

16.3 Mutual Independence

16.3.1 Definition

We have defined what it means for two events to be independent. What if there are more than two events? For example, how can we say that the flips of n coins are all independent of one another?

Events $E_1, \ldots, E_n$ are said to be *mutually independent* if and only if the probability of any event E_i is unaffected by knowledge of the other events. More formally:

Definition 16.3.1. A set of events $E_1, E_2, \ldots, E_n$, is *mutually independent* if $\forall i \in [1, n]$ and $\forall S \subseteq [1, n] - \{i\}$, either

$$\Pr\left[\bigcap_{j \in S} E_j\right] = 0 \quad \text{or} \quad \Pr[E_i] = \Pr\left[E_i \;\middle|\; \bigcap_{j \in S} E_j\right].$$

In other words, no matter which other events are known to occur, the probability that E_i occurs is unchanged for any i.

For example, if we toss 100 fair coins at different times, we might reasonably assume that the tosses are mutually independent since the probability that the ith coin is heads should be $1/2$, no matter which other coin tosses came out heads.

16.3.2 Alternative Formulation

Just as Theorem 16.1.2 provided an alternative definition of independence for two events, there is an alternative definition for mutual independence.

Theorem 16.3.2. *A set of events $E_1, E_2, \ldots, E_n$ is mutually independent iff $\forall S \subseteq [1, n]$,*

$$\Pr\left[\bigcap_{j \in S} E_j\right] = \prod_{j \in S} \Pr[E_j].$$

The proof of Theorem 16.3.2 uses induction and reasoning similar to the proof of Theorem 16.1.2. We will not include the details here.

Theorem 16.3.2 says that $E_1, E_2, \ldots, E_n$ are mutually independent if and only

if all of the following equations hold for all distinct i, j, k, and l:

$$\begin{aligned}
\Pr[E_i \cap E_j] &= \Pr[E_i] \cdot \Pr[E_j] \\
\Pr[E_i \cap E_j \cap E_k] &= \Pr[E_i] \cdot \Pr[E_j] \cdot \Pr[E_k] \\
\Pr[E_i \cap E_j \cap E_k \cap E_l] &= \Pr[E_i] \cdot \Pr[E_j] \cdot \Pr[E_k] \cdot \Pr[E_l] \\
&\vdots \\
\Pr[E_1 \cap \cdots \cap E_n] &= \Pr[E_1] \cdots \Pr[E_n].
\end{aligned}$$

For example, if we toss n fair coins, the tosses are mutually independent iff for all $m \in [1, n]$ and every subset of m coins, the probability that every coin in the subset comes up heads is 2^{-m}.

16.3.3 DNA Testing

Assumptions about independence are routinely made in practice. Frequently, such assumptions are quite reasonable. Sometimes, however, the reasonableness of an independence assumption is not so clear, and the consequences of a faulty assumption can be severe.

For example, consider the following testimony from the O. J. Simpson murder trial on May 15, 1995:

Mr. Clarke: When you make these estimations of frequency—and I believe you touched a little bit on a concept called independence?

Dr. Cotton: Yes, I did.

Mr. Clarke: And what is that again?

Dr. Cotton: It means whether or not you inherit one allele that you have is not—does not affect the second allele that you might get. That is, if you inherit a band at 5,000 base pairs, that doesn't mean you'll automatically or with some probability inherit one at 6,000. What you inherit from one parent is what you inherit from the other.

Mr. Clarke: Why is that important?

Dr. Cotton: Mathematically that's important because if that were not the case, it would be improper to multiply the frequencies between the different genetic locations.

Mr. Clarke: How do you—well, first of all, are these markers independent that you've described in your testing in this case?

Presumably, this dialogue was as confusing to you as it was for the jury. Essentially, the jury was told that genetic markers in blood found at the crime scene matched Simpson's. Furthermore, they were told that the probability that the markers would be found in a randomly-selected person was at most 1 in 170 million. This astronomical figure was derived from statistics such as:

- 1 person in 100 has marker A.
- 1 person in 50 marker B.
- 1 person in 40 has marker C.
- 1 person in 5 has marker D.
- 1 person in 170 has marker E.

Then these numbers were multiplied to give the probability that a randomly-selected person would have all five markers:

$$\begin{aligned}\Pr[A \cap B \cap C \cap D \cap E] &= \Pr[A] \cdot \Pr[B] \cdot \Pr[C] \cdot \Pr[D] \cdot \Pr[E] \\ &= \frac{1}{100} \cdot \frac{1}{50} \cdot \frac{1}{40} \cdot \frac{1}{5} \cdot \frac{1}{170} \\ &= \frac{1}{170{,}000{,}000}.\end{aligned}$$

The defense pointed out that this assumes that the markers appear mutually independently. Furthermore, all the statistics were based on just a few hundred blood samples.

After the trial, the jury was widely mocked for failing to "understand" the DNA evidence. If you were a juror, would *you* accept the 1 in 170 million calculation?

16.4 Pairwise Independence

The definition of mutual independence seems awfully complicated—there are so many subsets of events to consider! Here's an example that illustrates the subtlety of independence when more than two events are involved. Suppose that we flip three fair, mutually-independent coins. Define the following events:

- A_1 is the event that coin 1 matches coin 2.
- A_2 is the event that coin 2 matches coin 3.

Chapter 15

MFCS 17: Random Variables and Distributions

17 Random Variables and Distributions

Thus far, we have focused on probabilities of events. For example, we computed the probability that you win the Monty Hall game, or that you have a rare medical condition given that you tested positive. But, in many cases we would like to more more. For example, *how many* contestants must play the Monty Hall game until one of them finally wins? *How long* will this condition last? *How much* will I lose gambling with strange dice all night? To answer such questions, we need to work with random variables.

17.1 Definitions and Examples

Definition 17.1.1. A *random variable* R on a probability space is a total function whose domain is the sample space.

The codomain of R can be anything, but will usually be a subset of the real numbers. Notice that the name "random variable" is a misnomer; random variables are actually functions!

For example, suppose we toss three independent[1], unbiased coins. Let C be the number of heads that appear. Let $M = 1$ if the three coins come up all heads or all tails, and let $M = 0$ otherwise. Every outcome of the three coin flips uniquely determines the values of C and M. For example, if we flip heads, tails, heads, then $C = 2$ and $M = 0$. If we flip tails, tails, tails, then $C = 0$ and $M = 1$. In effect, C counts the number of heads, and M indicates whether all the coins match.

Since each outcome uniquely determines C and M, we can regard them as functions mapping outcomes to numbers. For this experiment, the sample space is

$$\mathcal{S} = \{HHH, HHT, HTH, HTT, THH, THT, TTH, TTT\}$$

and C is a function that maps each outcome in the sample space to a number as

[1]Going forward, when we talk about flipping independent coins, we will assume that they are *mutually* independent.

follows:

$$\begin{aligned} C(HHH) &= 3 & C(THH) &= 2 \\ C(HHT) &= 2 & C(THT) &= 1 \\ C(HTH) &= 2 & C(TTH) &= 1 \\ C(HTT) &= 1 & C(TTT) &= 0. \end{aligned}$$

Similarly, M is a function mapping each outcome another way:

$$\begin{aligned} M(HHH) &= 1 & M(THH) &= 0 \\ M(HHT) &= 0 & M(THT) &= 0 \\ M(HTH) &= 0 & M(TTH) &= 0 \\ M(HTT) &= 0 & M(TTT) &= 1. \end{aligned}$$

So C and M are random variables.

17.1.1 Indicator Random Variables

An *indicator random variable* is a random variable that maps every outcome to either 0 or 1. Indicator random variables are also called *Bernoulli variables*. The random variable M is an example. If all three coins match, then $M = 1$; otherwise, $M = 0$.

Indicator random variables are closely related to events. In particular, an indicator random variable partitions the sample space into those outcomes mapped to 1 and those outcomes mapped to 0. For example, the indicator M partitions the sample space into two blocks as follows:

$$\underbrace{HHH \quad TTT}_{M=1} \quad \underbrace{HHT \quad HTH \quad HTT \quad THH \quad THT \quad TTH}_{M=0}.$$

In the same way, an event E partitions the sample space into those outcomes in E and those not in E. So E is naturally associated with an indicator random variable, I_E, where $I_E(w) = 1$ for outcomes $w \in E$ and $I_E(w) = 0$ for outcomes $w \notin E$. Thus, $M = I_E$ where E is the event that all three coins match.

17.1.2 Random Variables and Events

There is a strong relationship between events and more general random variables as well. A random variable that takes on several values partitions the sample space into several blocks. For example, C partitions the sample space as follows:

$$\underbrace{TTT}_{C=0} \quad \underbrace{TTH \quad THT \quad HTT}_{C=1} \quad \underbrace{THH \quad HTH \quad HHT}_{C=2} \quad \underbrace{HHH}_{C=3}.$$

Each block is a subset of the sample space and is therefore an event. Thus, we can regard an equation or inequality involving a random variable as an event. For example, the event that $C = 2$ consists of the outcomes THH, HTH, and HHT. The event $C \leq 1$ consists of the outcomes TTT, TTH, THT, and HTT.

Naturally enough, we can talk about the probability of events defined by properties of random variables. For example,

$$\begin{aligned}\Pr[C = 2] &= \Pr[THH] + \Pr[HTH] + \Pr[HHT] \\ &= \frac{1}{8} + \frac{1}{8} + \frac{1}{8} \\ &= \frac{3}{8}.\end{aligned}$$

As another example:

$$\begin{aligned}\Pr[M = 1] &= \Pr[TTT] + \Pr[HHH] \\ &= \frac{1}{8} + \frac{1}{8} \\ &= \frac{1}{4}.\end{aligned}$$

17.1.3 Functions of Random Variables

Random variables can be combined to form other random variables. For example, suppose that you roll two unbiased, independent 6-sided dice. Let D_i be the random variable denoting the outcome of the ith die for $i = 1, 2$. For example,

$$\Pr[D_1 = 3] = 1/6.$$

Then let $T = D_1 + D_2$. T is also a random variable and it denotes the sum of the two dice. For example,

$$\Pr[T = 2] = 1/36$$

and

$$\Pr[T = 7] = 1/6.$$

Random variables can be combined in complicated ways, as we will see in Chapter 19. For example,

$$Y = e^T$$

is also a random variable. In this case,

$$\Pr[Y = e^2] = 1/36$$

and