

Discrete Mathematics

Thomas Goller

January 2013

Contents

1	Mathematics	1
1.1	Axioms	1
1.2	Definitions	2
1.3	Theorems	3
1.4	Statements	4
2	Proofs	7
2.1	What is a Proof?	7
2.2	Direct Proof	8
2.3	Proof by Contraposition	9
2.4	Proof by Contradiction	10
2.5	Proving Logical Equivalence	11
2.6	Practice	12
3	Sets and Functions	14
3.1	Sets	14
3.2	Functions	19
3.3	Injectivity	21
3.4	Surjectivity	23
3.5	Bijectivity and Inverses	24
4	Number Theory	29
4.1	Divisibility and Primes	29
4.2	Euclidean Algorithm	32
4.3	Modular Arithmetic	36
4.4	Linear Congruences	41
4.5	Cryptography	43
5	Induction	49
5.1	Proof by Induction	49
5.2	Strong Induction and the FTA	52

6	Combinatorics	54
6.1	Counting	54
6.2	Permutations	56
6.3	Combinations	57
6.4	Binomial Theorem	59
6.5	Pascal's Triangle	60
6.6	Application of Combinations: 5-Card Poker Hands	63
7	Graph Theory	66
7.1	Graphs	67
7.2	Isomorphism	68
7.3	Some Types of Graphs	71
7.4	Connected Graphs	74
7.5	Induction on Connected Graphs	78
7.6	Planar Graphs and Euler's Formula	80

Acknowledgements

Much of the material in these notes is based on Kenneth Rosen's *Discrete Mathematics and Its Applications, Seventh Edition*. His encyclopedia of discrete mathematics covers far more than these few pages will allow. Edward Scheinerman's *Mathematics: A Discrete Introduction, Third Edition* is an inspiring model of a textbook written for the learner of discrete mathematics, rather than the teacher. I've tried (and surely failed) to make my prose and organization of the material as reader-friendly as his. Lastly, the internet is incredibly useful for learning mathematics: Wikipedia and Google are handy if you want to quickly find information on a topic, and Wolfram Alpha is fantastic for computations, finding roots of functions, and graphing.

Note to the Reader

Proofs, not computations, form the core of real mathematics. Gone are the days of churning through pages and pages of derivatives. Instead of focusing on techniques for solving problems, mathematicians focus on understanding abstract concepts well enough to prove theorems about them. As an ongoing student of mathematics, I assure you that real mathematics is hard, so our goals in this course are modest. If you work hard, you should become familiar with the rigorous language of mathematics and be able to write short proofs of your own. This course should prepare you for more advanced mathematics courses, such as foundations of analysis courses that many undergraduate mathematics students find incredibly challenging.

The topics we will cover – sets, functions, number theory, combinatorics, and graph theory – are a sampling of the many topics lumped together as “discrete mathematics”. For us, these topics will be primarily a means to practice proof-based mathematics, though sets and functions are fundamental to all areas of mathematics.

Chapter 1

Mathematics

1.1 Axioms

Mathematics begins with statements assumed to be true, called **axioms** or **postulates**. For example, the axioms of plane geometry, which you can look up online. Our starting point is the following two axioms:

1.1.1 Axiom. There is a set $\mathbb{Z} = \{\dots, -2, -1, 0, 1, 2, \dots\}$, called the **integers**, with the operations addition (+) and multiplication (\cdot), such that the following properties hold for all integers a, b, c :

$(a + b) + c = a + (b + c)$	associativity of addition
$a + b = b + a$	commutativity of addition
$a + (-a) = 0$	additive inverses
$a + 0 = a$	additive identity
$(a \cdot b) \cdot c = a \cdot (b \cdot c)$	associativity of multiplication
$a \cdot b = b \cdot a$	commutativity of multiplication
$a \cdot 1 = a$	multiplicative identity
$a \cdot (b + c) = a \cdot b + a \cdot c$	distributive property.

1.1.2 Axiom. There is a set \mathbb{R} , called the **real numbers**, with the operations addition (+) and multiplication (\cdot), satisfying the same properties as the integers plus one additional property: for any nonzero real number a , there is a real number $\frac{1}{a}$ such that $a \cdot \frac{1}{a} = 1$. We say that “multiplicative inverses” exist.

1.1.3 Note. These properties ensure that addition and multiplication in \mathbb{Z} and \mathbb{R} work in the way we are used to them working. Note that for both \mathbb{Z} and \mathbb{R} , subtraction is just addition of a negative number. Similarly, division in \mathbb{R} is just multiplication by a multiplicative inverse, but there is no division in \mathbb{Z} because there are no multiplicative inverses (except for 1 and -1).

1.1.4 Note. We could spend an entire semester studying how to construct the integers from much more basic (and not very enlightening) axioms. Then we could spend another semester creating the real numbers.

1.2 Definitions

All we have so far are two sets of “numbers” with some operations and properties. We now use **definitions** to introduce new mathematical concepts. In order to be useful, definitions have to be unambiguous, so they have to be stated very precisely.

Let’s try to define what it means for an integer to be “even”. The way we normally state a definition is:

Definition. An integer n is **even** if ...

We could finish the sentence with “ n is in the set $\{\dots, -4, -2, 0, 2, 4, \dots\}$ ”, but that would not be very enlightening. A better definition would describe “evenness” in terms of a property, such as “ n is divisible by 2”. But we have to be careful: we don’t know what it means to be “divisible by 2”, since that is not in our axioms, so we would have to first define “divisible by”. Although we could do that, we can give a definition purely in terms of multiplication:

1.2.1 Definition. An integer n is **even** if there exists an integer k such that $n = 2k$.

While we’re at it, we might as well define “odd” too. One option would be to say “an integer n is odd if it is not even”, but there is a more illuminating definition:

1.2.2 Definition. An integer n is **odd** if there exists an integer k such that $n = 2k + 1$.

These definitions of even and odd are useful because they tell us something about how the integer decomposes into a sum and product of other integers. But it is not clear from our definitions that an integer cannot be both even and odd, so that is a fact we will have to prove in the next chapter.

A familiar concept we have not yet defined is “rational number”. If we try to define “rational number” using just the integers – “a rational number is one integer divided by another integer” – we run into problems because division isn’t defined for integers. We would have to build the rational numbers as a new set of numbers, define operations, and do a lot of work to show that the usual properties hold. But since we are already assuming the existence of the real numbers, we can cheat by defining a rational number as a special kind of real number. Then we can use division since it is defined for nonzero real numbers.

1.2.3 Definition. A real number r is **rational** if there exist integers a and b with $b \neq 0$ such that $r = a/b$.

We’d also like to define “irrational”. Can you think of a useful property that characterizes when a real number is irrational? There does not seem to be such a property, so we have to settle for:

1.2.4 Definition. A real number that is not rational is called **irrational**.

We'll go crazy with new definitions in later chapters. But first we will introduce some other key notions in mathematics.

1.3 Theorems

Theorems are mathematical statements that have been proven to be true. For instance, here's a famous theorem you have surely encountered:

Theorem (Pythagoras). *Let a and b denote the lengths of the legs of a right triangle, and let c be the length of the hypotenuse. Then*

$$a^2 + b^2 = c^2.$$

We'll focus on proofs in the next chapter. For now, let's try to understand how theorems are stated:

- The “let ...” sentence establishes the context and notation: we're dealing with a right triangle with side lengths a , b , and c . In fact, since the sentence does not specify a particular right triangle, the theorem applies to *every* right triangle.
- The “then ...” sentence makes a non-obvious statement about an equation those lengths satisfy. It is remarkable that this equation holds for *every* right triangle. This is the part of the theorem that requires a proof.

The word “theorem” is usually reserved for the most important theorems. Less important theorems are called **propositions**, **results**, and **facts**. A **lemma** is a theorem whose main purpose is in the proof of a more important theorem. A **corollary** is a theorem whose proof follows almost immediately from a previous theorem.

Mathematical statements that have not yet been proven or disproven are called **claims** or **conjectures**. We will call less important mathematical statements “claims”, even if we then go on to prove them, in which case we can refer to them as **true claims**. Here is a claim:

Claim. *Let n be an integer. If n is odd, then n^2 is odd.*

Again, let's figure out what the claim is saying:

- The “let ...” sentence establishes the notation: we are using n to represent a fixed but *arbitrary* integer.
- The second sentence is in the form “if ... then ...”. Roughly speaking, the “if” part imposes an additional condition on n , namely that “ n is odd”. Given this additional condition, the claim says that n^2 must be odd.

Another common sentence type is “... if and only if ...”:

Claim. *Let n be an integer. Then n is odd if and only if n^2 is odd.*

We'll analyze what these sentences mean in the next section. To do so, we will take a step back and define “statement”, which we have already used to define “theorem”, above.

1.4 Statements

A (mathematical) **statement** is a meaningful sentence about mathematics that is either true or false, not both. We often denote statements by lower-case letters like p and q . The **truth value** of a statement is either true (T) or false (F).

1.4.1 Example.

- “2 is even” is a true statement.
- “ $1 + 2 = -7$ ” is a false statement.
- “ $1 = + + 3 -$ ” is not a statement because it is gibberish.

1.4.2 Exercise. Come up with your own examples of a true statement, a false statement, and a sentence that is not a statement.

Given two statements p and q , there are several ways to combine them to get a new statement.

- ▷ **“Not”**: The statement “it is not true that p ”, often shortened to “not p ”, is true if p is false and false if p is true.
- ▷ **“And”**: The statement “ p and q ” is true when both p and q are true and false otherwise.
- ▷ **“Or”**: The statement “ p or q ” is true when p or q (possibly both) are true and false when both p and q are false.
- ▷ **“If, then”**: The statement “if p , then q ” or “ p implies q ” is often denoted $p \implies q$. It is trivially true when p is false, since then the “if p ” condition is not met, and true when both p and q are true. It is false only when p is true and q is false.
- ▷ **“If and only if”**: The statement “ p if and only if q ”, which means “(if p , then q) and (if q , then p)”, is often denoted $p \iff q$. It is true when p and q have the same truth value and false otherwise.

1.4.3 Example.

- “It is not true that 2 is even”, which can be shortened to “2 is not even”, is a false statement. “It is not true that $1 + 2 \neq -7$ ”, which can be shortened to “ $1 + 2 = -7$ ”, is a true statement.

- “2 is even and $1 + 2 = -7$ ” is a false statement because “ $1 + 2 = -7$ ” is false.
- “2 is even or $1 + 2 = -7$ ” is a true statement because “2 is even” is true. “2 is even or $1 + 2 \neq -7$ ” is also a true statement.
- “If 2 is even, then $1 + 2 = -7$ ” is false, because “2 is even” is true and $1 + 2 = -7$ is false.
- “If 2 is even, then $1 + 2 \neq -7$ ” is true.
- “If 2 is not even, then $1 + 2 = 7$ ” is (trivially) true because “2 is not even” is false!
- “2 is even if and only if $1 + 2 = -7$ ” is a false statement.

1.4.4 Note. Here are some comments about how our notions of “or”, “if, then”, and “if and only if” relate to everyday usage.

- In everyday language, “or” can be ambiguous. I might say “I will buy you a lollipop or I will buy you a Honeycrisp apple”, with the implication that I will buy you one or the other but not both. This is *not* how our “or” works! The statement “ p or q ” is considered to be *true* when both p and q are true!
- A father might say to his daughter: “My dear, if you get an ‘A’ in astrophysics, then I will buy you a bicycle”. When has the father upheld his word? If the daughter gets an ‘A’ and father buys the bike, then all is well. If the daughter fails to get an ‘A’, then father need not buy a bike, though he might get it anyway as consolation. But if the daughter gets the ‘A’ and father does not buy her the bicycle, then the daughter has a reason to be upset, and we would say daddy has been dishonest.
- Suppose I tell you “I will move to India if and only if you will move to India”. I am claiming that we will either move together or not at all, and I will be wrong if one of us goes while the other stays.

1.4.5 Example. We can use a **truth table** to summarize the truth values of more complicated statements in terms of the truth values of their building blocks.

p	q	“not p ”	“ p and q ”	“ p or q ”	“if p , then q ”	“ p if and only if q ”
T	T	F	T	T	T	T
T	F	F	F	T	F	F
F	T	T	F	T	T	F
F	F	T	F	F	T	T

The four rows correspond to the four possible combinations of truth values for p and q , which appear in the first two columns. To determine the truth values in the other five columns, use the truth values for p and q and the definition of “not”, “and”, “or”, and so on.

Let p and q be statements. The **converse** of the statement “if p , then q ” is the statement “if q , then p ”. The **inverse** of “if p , then q ” is the statement “if not p , then not q ”. The **contrapositive** of “if p , then q ” is “if not q , then not p ”.

1.4.6 Exercise. Let p and q be statements. Use truth tables to show that:

- (a) “If p , then q ” may not have the same truth value as its converse.
- (b) “If p , then q ” may not have the same truth value as its inverse.
- (c) “If p , then q ” always has the same truth value as its contrapositive.
- (d) “Not (p and q)” always has the same truth value as “not p or not q ”. (The parentheses indicate that the “not” applies to the whole statement “ p and q ”.)
- (e) “Not (p or q)” always has the same truth value as “not p and not q ”.

Parts (d) and (e) are called De Morgan’s laws. They say that you can distribute “not” over an “and” or an “or”, but you have to flip “and” to “or” and vice versa.

1.4.7 Note. Because of (c), we say that an “if ... then ...” statement and its contrapositive are **logically equivalent**. Logical equivalence is the same notion as “if and only if”. Another way of saying “ p if and only if q ” is saying p is equivalent to q .

1.4.8 Example. Now we are ready to analyze some claims.

- **Claim.** *Let n be an integer. If n is odd, then n^2 is odd.*

Recall that the “let ...” sentence establishes that n is a fixed but arbitrary integer. When n is not odd, the “if ... then ...” sentence is trivially true since it does not promise anything. The only way the “if ... then ...” sentence can be false is if there is an integer n such that n is odd and n^2 is not odd.

- **Claim.** *Let n be an integer. Then n is odd if and only if n^2 is odd.*

By the definition of “if and only if”, the claim says that when n is odd, so is n^2 , and when n^2 is odd, so is n . Since the statement “if p , then q ” is equivalent to its contrapositive, “if n^2 is odd, then n is odd” is equivalent to “if n is not odd, then n^2 is not odd”.

- **Claim.** *Let m and n be integers. If mn is even, then m is even or n is even.*

By definition of “if, then”, the claim is true unless mn is even and “ m is even or n is even” is false, which happens when “ m is not even and n is not even” is true.

Chapter 2

Proofs

2.1 What is a Proof?

As we saw in Chapter 1, axioms and definitions establish terminology, so they cannot be true or false. A theorem or claim, on the other hand, makes an assertion, often indicated by a “then” statement, which we call its **conclusion**. For instance, the conclusion of the Pythagorean theorem is the statement that $a^2 + b^2 = c^2$.

A (mathematical) **proof** is a short essay used to show that a theorem, proposition, result, fact, claim, conjecture, lemma, or corollary is true. Starting with the assumptions of the theorem, one uses axioms, definitions, and previously proved theorems to build a chain of implications ending with the conclusion.

Claims or conjectures may be false. To **disprove** a claim or conjecture, namely to show it is false, find a **counterexample**, which is a case where the assumptions of the claim are true, but the conclusion is false. This is usually much easier than proving a claim is true!

2.1.1 Note. For a claim to be true, it has to be true in every possible case. A single counterexample makes a claim false, even if the claim is true in infinitely many other cases!

2.1.2 Example.

- **Claim.** *Let n be a real number. Then $n^2 > 0$.*

The claim is false because it fails when $n = 0$.

- **Claim.** *Let n be a nonzero real number. Then $n^2 > 0$.*

This claim is true because the only counterexample of the previous claim has been removed by adding the assumption “nonzero”. As you can see, claims must be stated very precisely! Some claims have a lot of assumptions because they are needed to rule out counterexamples.

2.1.3 Note (Outline for Proving or Disproving Claims).

- **Step 1:** Before you attempt a proof, convince yourself that the claim is correct! Checking easy cases is often a good idea. See if you can find a counterexample. If you can, just write “This claim is false!”, state your counterexample, briefly explain why it is a counterexample, and you’re done!
- **Step 2:** On scratch paper, write down the assumptions (“given”) at the top of the page, and the conclusion (“want”) at the bottom of the page. Make sure you know what you want to show before you try to show it!
- **Step 3:** Fill in the (both physical and logical) space between the “given” and “want”. Use definitions, axioms (mainly arithmetic in our case), and previously proved theorems to deduce the “want” from the “given”. (This is an attempt at a direct proof: if you get stuck, try a proof by contraposition or contradiction instead.)
- **Step 4:** Once you have an outline of the proof on your scratch paper, convert it into precise, crisp English sentences, possibly with some mathematical symbols mixed in. Label it “proof”, draw your favorite symbol at the end, and you have yourself a proof!

2.2 Direct Proof

Recall our definition of even and odd from Chapter 1:

Definition. An integer n is **even** if there exists an integer k such that $n = 2k$. An integer n is **odd** if there exists an integer k such that $n = 2k + 1$.

Let’s take a look at our first easy result!

2.2.1 Claim. *Let n be an integer. If n is odd, then n^2 is odd.*

In proving a claim involving an “if ... then ...”, we always assume the statement following the “if” to be true, since otherwise the claim is trivially true. Checking the claim for some odd integers, we see that $(-3)^2 = 9$, $(-1)^2 = 1$, $1^2 = 1$, $3^2 = 9$, $5^2 = 25$, and so on, so the claim seems plausible.

Our “given” is not only that n is an integer but also that n is odd, namely that there exists an integer k such that $n = 2k + 1$. Our “want” is to show that n^2 is odd, which means *finding* an integer j such that $n^2 = 2j + 1$. How do we find this integer j ? We need some way to use our information about n to deduce something about n^2 . We have an equation for n , so square it! This gives us $n^2 = (2k + 1)^2 = 4k^2 + 4k + 1 = 2(2k^2 + 2k) + 1$. Since $2k^2 + 2k$ is an integer, setting $j = 2k^2 + 2k$ yields $n^2 = 2j + 1$, so n^2 is odd. We’ve finished our scratch work, so here is the proof:

Proof. Since n is odd, there exists an integer k such that $n = 2k + 1$. Then $n^2 = (2k + 1)^2 = 4k^2 + 4k + 1 = 2(2k^2 + 2k) + 1$, so n^2 is odd. \square

Notice how short and precise the finished product is! Since the claim tells us that n is an integer, we don't have to repeat that in the proof. Moreover, it is clear that we will be assuming the "if" statement, so we don't have to make a big fuss about assuming that n is odd. We do write "since n is odd" because that is the reason why k exists.

Also note how we incorporate the computation into the second sentence. We don't talk about an integer j because there is no need: the equation $n^2 = 2(2k^2 + 2k) + 1$ is clearly of the form 2 times an integer plus 1. In general, it is better to avoid introducing new variables unless there is a good reason for doing so.

Finally, note that we used the definition of "odd" twice: once to extract useful information about n (there exists k such that $n = 2k + 1$), and again to tell us what we need to prove (that there exists j such that $n^2 = 2j + 1$). If you don't know the definition of odd, you have no chance of proving a claim about odd integers. You have to know the main definitions so that you can prove claims involving them!

The preceding proof is called a **direct proof**. We started with the assumptions (n is an odd integer) and were able to directly deduce the conclusion (n^2 is odd).

2.2.2 Exercise. Let n be an integer. Prove the following claims:

- (a) If n is even, then n^2 is even.
- (b) If n is odd, then n^3 is odd.

2.3 Proof by Contraposition

Sometimes direct proofs are difficult:

2.3.1 Claim. *Let n be an integer. If n^2 is odd, then n is odd.*

Let's try to work out a direct proof. Our "given" is that n is an integer and n^2 is odd, so there exists k such that $n^2 = 2k + 1$. Our "want" is to find an integer j such that $n = 2j + 1$. But how can we use information about n^2 to deduce something about n ? We might try taking the square root of both sides of the equation $n^2 = 2k + 1$, which yields the equation $n = \pm\sqrt{2k + 1}$. Unfortunately, this equation does not seem to help in finding j . What next?

This claim is easy to prove if we use a **proof by contraposition**. Recall that a statement of the form "if p , then q " is equivalent to its contrapositive "if not q , then not p ". So we can prove the contrapositive instead!

In our case, the contrapositive is the claim "Let n be an integer. If n is not odd, then n^2 is not odd." We'll prove shortly that any integer that is not odd is even. Granting that fact, the claim we want to prove is "Let n be an integer. If n is even, then n^2 is even." The proof is similar to our proof of Claim 2.2.1 above.

Proof. Proof by contraposition. Since n is even, there exists an integer k such that $n = 2k$. Then $n^2 = (2k)^2 = 4k^2 = 2(2k^2)$, so n^2 is even. \square

Note that we started the proof by pointing out that we are using proof by contraposition. Then we started proving the contrapositive, without even stating what it is. Of course, you should write down the contrapositive on your scratch paper!

2.3.2 Exercise. Let n be an integer. Prove the following claims by contraposition:

- (a) If n^2 is even, then n is even.
- (b) If n^3 is odd, then n is odd.

2.4 Proof by Contradiction

Another useful type of proof is **proof by contradiction**. To set up a proof by contradiction, use all the usual assumptions in the claim, but also suppose the conclusion is false. Then try to derive something false, which is called a **contradiction**. If this is achieved, then the conclusion couldn't be false, hence it must be true.

2.4.1 Claim. *Let n be an integer. Then n is even or odd, but not both.*

This claim is more complicated. The conclusion really means “(n is even or odd) and (n is not even and odd)”. Since this is an “and” statement, we have to prove each part. We'll use a direct proof to show “ n is even or odd” and a proof by contradiction to show “ n is not even and odd”. Here we go!

Proof. First we prove that n is even or odd. Since we can write the set of integers as

$$\{\dots, -2, -1, 0, 1, 2, \dots\} = \{\dots, 2 \cdot (-1), 2 \cdot (-1) + 1, 2 \cdot 0, 2 \cdot 0 + 1, 2 \cdot 1, \dots\},$$

the integers alternate between even and odd, so n must be even or odd.

Next we prove that n cannot be both even and odd. Assume for contradiction that n is even and odd. Since n is even, $n = 2k$ for some integer k . Since n is odd, $n = 2j + 1$ for some integer j . Then $2k = 2j + 1$, so $2(k - j) = 1$. There are three possibilities for $k - j$:

- (1) $k - j \geq 1$. But then $2(k - j) \geq 2$, contradicting $2(k - j) = 1$.
- (2) $k - j = 0$. But then $2(k - j) = 0 \neq 1$.
- (3) $k - j \leq -1$. But then $2(k - j) \leq -2$, contradicting $2(k - j) = 1$.

In each case, there is a contradiction. Thus our assumption must be false. So n cannot be both even and odd. □

Note that we made it very clear which part of the conclusion we were proving. The first part was a direct proof, which is always assumed unless otherwise specified. For the second part, we clearly stated the assumption we were making.

This proof was much more difficult! It was not enough to just apply definitions; we had to use some ingenuity too! Don't worry: most of the proofs you will be writing will look more like the proofs of Claims 2.2.1 and 2.3.1.

2.4.2 Exercise. Let r be irrational. Prove the following claims by contradiction:

(a) $r + \frac{1}{2}$ is irrational.

(b) $2r$ is irrational.

(*Hint:* Recall the definitions of rational and irrational!)

2.4.3 Note. We could also write the claims in the preceding exercise as “if ... then ...” statements, for instance “Let r be a real number. If r is irrational, then $r + \frac{1}{2}$ is irrational.” Then to do a proof by contradiction, you assume the “then” statement is false. This is very similar to proof by contraposition!

2.5 Proving Logical Equivalence

Here is a claim involving logical equivalence.

2.5.1 Claim. *Let n be an integer. Then n is odd if and only if n^2 is odd.*

In other words, the claim is saying that the statements “ n is odd” and “ n^2 is odd” are logically equivalent. As we mentioned in Section 1.4, an “if and only if” statement is really just two “if ... then ...” statements combined. Another way of stating the above claim is “Let n be an integer. If n is odd, then n^2 is odd, and if n^2 is odd, then n is odd.” As usual, to prove an “and” statement we need to prove both parts. We’ve already proved both parts in this chapter, so all we have to do is combine the two proofs and clarify which part we are proving:

Proof. First we prove that if n is odd, then n^2 is odd. Since n is odd, there exists an integer k such that $n = 2k + 1$. Then $n^2 = (2k + 1)^2 = 4k^2 + 4k + 1 = 2(2k^2 + 2k) + 1$, so n^2 is odd.

Next we use contraposition to prove that if n^2 is odd, then n is odd. Since n is even, there is an integer k such that $n = 2k$. Then $n^2 = 4k^2 = 2(2k^2)$, so n is even. \square

2.5.2 Exercise. Let n be an integer. Prove the following claims:

(a) n is odd if and only if $n + 2$ is odd.

(b) n is odd if and only if $n + 1$ is even.

There are also more complicated statements of equivalence.

2.5.3 Claim. *Let n be an integer. Then the following are equivalent:*

(i) n is even;

(ii) $n + 1$ is odd;

(iii) $n + 2$ is even.

A statement of the form **the following are equivalent**, which is often abbreviated “TFAE”, is followed by two or more statements (in our case, (i), (ii), and (iii)). When dealing with so many statements, writing “if ... then ...” becomes cumbersome, so we often use the notation “ \implies ” and the word “implies” instead. TFAE means that each of the following statements implies all of the others. In other words: (i) \implies (ii) and (iii); (ii) \implies (i) and (iii), and (iii) \implies (i) and (ii). Another way of making the same claim is to say (i) \iff (ii), (ii) \iff (iii), and (iii) \iff (i). Proving three “if and only if” statements would usually involve 6 separate proofs, but luckily we don’t have to do that much work. Instead, it suffices to prove a circular chain of implications involving all the statements. For us, this means proving (i) \implies (ii) \implies (iii) \implies (i) or proving (i) \implies (iii) \implies (ii) \implies (i). In either case, any one of the three statements implies both other statements. We’ll give a proof using the first chain of implications:

Proof. (i) \implies (ii): Since n is even, there exists an integer k such that $n = 2k$. Then $n + 1 = 2k + 1$ is odd.

(ii) \implies (iii): Since $n + 1$ is odd, there is an integer j such that $n + 1 = 2j + 1$. Then $n + 2 = 2j + 2 = 2(j + 1)$ is even.

(iii) \implies (i): Since $n + 2$ is even, there is an integer l such that $n + 2 = 2l$. Then $n = 2l - 2 = 2(l - 1)$ is even. \square

That wasn’t so bad! We only had to prove three easy “if ... then ...” statements. Here’s a similar claim for you to prove:

2.5.4 Exercise. Let n be an integer. Prove that the following are equivalent:

- (i) n is odd;
- (ii) $n - 3$ is even;
- (iii) $n - 2$ is odd.

2.6 Practice

In Sections 2.3 to 2.5, we let down our guard and just blindly proved everything in our path. Now it is time to reactivate your brain and use the 4-step procedure in Note 2.1.3 before plunging into a proof. Many of the claims below are false, and it is impossible to prove a false claim! If you think a claim is true, it is up to you to figure out whether to use a direct proof or a proof by contraposition or contradiction.

Don’t turn in your scratch work! If a claim is false, your submitted solution should consist of “this claim is false!”, a counterexample, and a brief explanation of why it’s a counterexample. If a claim is true, write a precise, crisp proof. In most cases, your proof shouldn’t be more than a few sentences!

2.6.1 Exercise. Let m and n be integers. Prove or disprove each of the following claims.

- (a) If m is even and n is even, then $m + n$ is even.

- (b) If m is even and n is odd, then mn is odd.
- (c) n is odd if and only if n^3 is odd.
- (d) n is even if and only if $7n + 4$ is even.
- (e) If mn is even, then m is even or n is even.
- (f) If n is odd, then n can be written as the sum of the squares of two integers.
(*Note:* The two integers need not be distinct!)
- (g) If $n \geq 0$, then n can be written as the sum of the squares of three integers.
(*Note:* The three integers need not be distinct!)

2.6.2 Exercise. Let r and s be real numbers. Prove or disprove each of the following claims.

- (a) If r and s are rational, then $r + s$ and rs are rational.
- (b) If r is rational and s is irrational, then $r + s$ is irrational.
- (c) If r and s are irrational, then rs is irrational.
- (d) r is rational if and only if r^2 is rational.

Chapter 3

Sets and Functions

Now that we have some experience writing proofs, we can begin a rigorous study of mathematics. The natural place to start is with the most basic and important mathematical objects, sets and functions.

3.1 Sets

3.1.1 Definition. A **set** is an unordered collection of objects, called **elements** of the set. A set is said to **contain** its elements. If A is a set, we write $a \in A$ to denote that a is an element of A , and $a \notin A$ if a is not an element of A .

Two ways to describe a set are:

- (1) List all of its elements between a pair of curly braces ($\{$ and $\}$). If the elements follow an obvious pattern, then you can use an ellipsis (\dots) to indicate that the pattern continues.
- (2) Given a set A , you can create a new set B by defining properties that an element of the original set must satisfy to be in the new set. This is done by writing $B = \{a \in A \mid \text{properties}\}$. The vertical line (\mid) means “such that”.

3.1.2 Example. Here are some sets.

- The set with no elements, called the **empty set** and denoted \emptyset or $\{\}$;
- $\{\text{kiwi, dragon, shark, turtle, penguin, radish}\}$;
- $\mathbb{Z} = \{\dots, -2, -1, 0, 1, 2, \dots\}$, the set of **integers**;
- $\mathbb{Z}_{\geq 1} = \{1, 2, 3, \dots\} = \{n \in \mathbb{Z} \mid n \geq 1\}$, the set of **positive integers**;
- $\mathbb{Z}_{\leq -1} = \{-1, -2, -3, \dots\} = \{n \in \mathbb{Z} \mid n \leq -1\}$, the set of **negative integers**;
- $\mathbb{Z}_{\geq 0} = \{0, 1, 2, \dots\} = \{n \in \mathbb{Z} \mid n \geq 0\}$, the set of **non-negative integers**;

- $\mathbb{Z}_{\leq 0} = \{0, -1, -2, \dots\} = \{n \in \mathbb{Z} \mid n \leq 0\}$, the set of **non-positive integers**;
- \mathbb{R} , the set of **real numbers**;
- $(0, 1] = \{x \in \mathbb{R} \mid 0 < x \leq 1\}$;

3.1.3 Note. Changing the order in which you list elements of a set does not change the set. For instance, $\{1, 0\} = \{0, 1\}$. Listing an element repeatedly also does not change the set. For instance, $\{1, 1, 0, 1\} = \{0, 1\}$.

3.1.4 Note. Both $\mathbb{Z}_{\geq 0}$ and $\mathbb{Z}_{\geq 1}$ are sometimes referred to as the set of “natural numbers”. We will avoid ambiguity by saying “non-negative integers” and “positive integers”, respectively.

3.1.5 Note. In Chapters 1 and 2, we often started a claim with the sentence “Let n be an integer.” Now, with our new notation, we can shorten such a sentence to “Let $n \in \mathbb{Z}$.”

3.1.6 Definition. Let A and B be sets. We say A is a **subset** of B , written $A \subseteq B$ or $B \supseteq A$, if every element of A is an element of B . Two sets A and B are **equal**, written $A = B$, if A and B have the exactly same elements. If two sets are not equal we write $A \neq B$. We say A is a **proper subset** of B , written $A \subsetneq B$, if A is a subset of B and $A \neq B$.

3.1.7 Example.

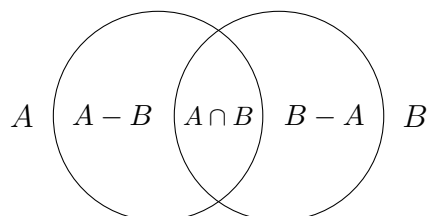
- The empty set is a subset of every set.
- $\mathbb{Z}_{\geq 1} \subsetneq \mathbb{Z}_{\geq 0} \subsetneq \mathbb{Z} \subsetneq \mathbb{R}$ and $(0, 1] \subsetneq \mathbb{R}$.
- The second set in Example 3.1.2 is a subset of the set that contains all vegetables and all mythical and real animals that are not mammals (“kiwi” refers to the bird, of course!).

3.1.8 Definition. Let A and B be sets. The **union** of A and B is the set $A \cup B = \{x \mid x \in A \text{ or } x \in B\}$. The **intersection** of A and B is the set $A \cap B = \{x \mid x \in A \text{ and } x \in B\}$. A and B are called **disjoint** if $A \cap B = \emptyset$. The union of two disjoint sets A and B is called the **disjoint union** of A and B and is denoted $A \sqcup B$. The **difference** of A and B (or the **complement** of B in A) is the set $A - B = \{x \in A \mid x \notin B\}$.

3.1.9 Example.

- Let $A = \{0, 1, 2\}$ and $B = \{2, 3\}$. Then $A \cup B = \{0, 1, 2, 3\}$, $A \cap B = \{2\}$, $A - B = \{0, 1\}$, and $B - A = \{3\}$. It wouldn’t make sense to write $A \sqcup B$ since A and B are not disjoint.
- We can write $\{a, b\} \sqcup \{c, d\} = \{a, b, c, d\}$ since $\{a, b\} \cap \{c, d\} = \emptyset$. It is also correct to write $\{a, b\} \cup \{c, d\} = \{a, b, c, d\}$.

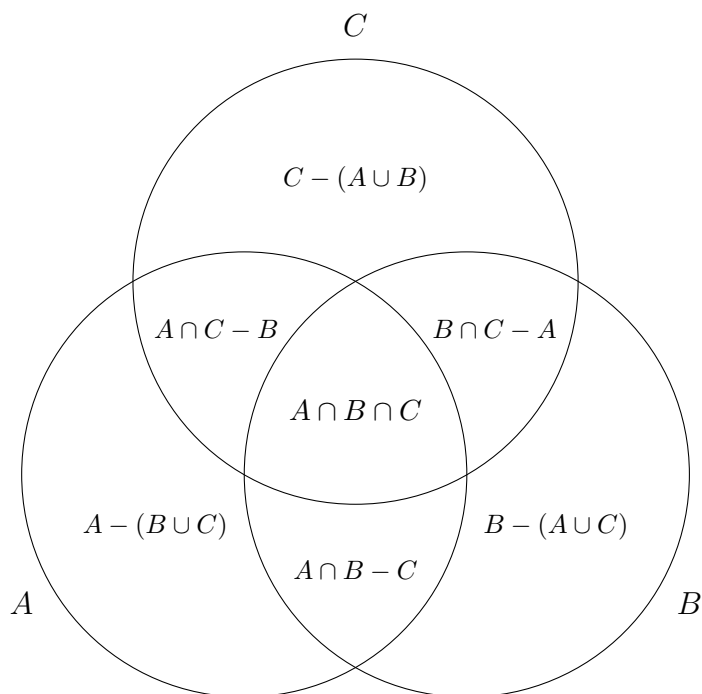
- Draw Venn diagrams to visualize union, intersection, disjoint union, and complement. For instance, visualize two general sets A and B as the interiors of two overlapping circles:



The labels “A” and “B” distinguish the two circles. Each region in the drawing is described using the terminology of the definition. As the diagram suggests, $(A - B) \sqcup (A \cap B) = A$, $(A \cap B) \sqcup (B - A) = B$, and $(A - B) \sqcup (A \cap B) \sqcup (B - A) = A \cup B$, where we have used “ \sqcup ” because the unions are disjoint.

A Venn diagram is only useful as a tool for distinguishing various intersections, unions, and complements. It should not be taken too literally. For instance, the labeled regions are disjoint, but it is unclear how many elements are in each labeled region. For all we know, A could be empty, in which case the interior of the A -circle contains no elements. The sets A and B are disjoint if $A \cap B$ is empty.

- For three sets, the Venn diagram looks like



Make sure you understand why the labels on each region are correct. The diagram suggests the identity $(A \cap B - C) \sqcup (A \cap B \cap C) = A \cap B$ among many others. As before, the diagram says nothing about how many elements are in each region.

- Let E be the even integers and O be the odd integers. Then $E \cap O = \emptyset$, so E and O are disjoint. $E \sqcup O = \mathbb{Z}$, $\mathbb{Z} - E = O$, and $E - \mathbb{Z} = \emptyset$.

3.1.10 Definition. If A is a set containing finitely many elements, then we call A **finite**. When A is finite, the **cardinality** of A , denoted $|A|$, is the number of elements of A . Note that $|A| \in \mathbb{Z}_{\geq 0}$.

3.1.11 Example.

- $|\emptyset| = 0$;
- $|\{17, \text{apple}\}| = 2$.

3.1.12 Exercise. Let A and B be finite sets. Prove or disprove the following claims:

- $|A \cup B| = |A| + |B|$.
- $|A - B| = |A| - |B|$.
- If A and B are disjoint, then $|A \sqcup B| = |A| + |B|$.
- $|A \cup B| = |A| + |B| - |A \cap B|$.

Here are some more complicated claims involving sets. The first claim is essential for proving two sets are equal.

3.1.13 Claim. *Let A and B be sets. Then $A = B$ if and only if $A \subseteq B$ and $B \subseteq A$.*

Proof. (\implies): Since $A = B$, if $a \in A$ then $a \in B$, so $A \subseteq B$. Similarly, if $b \in B$, then $b \in A$, so $B \subseteq A$.

(\impliedby): If $a \in A$, then $a \in B$ since $A \subseteq B$. If $b \in B$, then $b \in A$ since $B \subseteq A$. Thus $A = B$. \square

This claim is so useful that we will almost always use it when proving set equality. For instance, consider:

3.1.14 Claim. *Let A , B , and C be sets. Then $A \cap (B \cup C) = (A \cap B) \cup (A \cap C)$.*

Before even thinking about a proof, check the claim on a Venn diagram. Venn diagrams are incredibly useful for checking whether a claim involving several sets, unions, intersections, and complements is true. In this case, you should discover that the claim is correct.

By Claim 3.1.13, we need to show that $A \cap (B \cup C) \subseteq (A \cap B) \cup (A \cap C)$ and that $A \cap (B \cup C) \supseteq (A \cap B) \cup (A \cap C)$. These inequalities are tedious to write out, so instead of writing “First we will prove that $A \cap (B \cup C) \subseteq (A \cap B) \cup (A \cap C)$ ” we will simply write “ \subseteq : ” in the proof, and similarly we will write “ \supseteq : ” for the second inequality. This isn’t too ambiguous because it is easy to see what is going on in a well-written proof.

Proof. \subseteq : If $a \in A \cap (B \cup C)$, then $a \in A$ and $a \in B \cup C$. Thus $a \in B$ or $a \in C$, so $a \in A \cap B$ or $a \in A \cap C$, namely $a \in (A \cap B) \cup (A \cap C)$.

\supseteq : If $a \in (A \cap B) \cup (A \cap C)$, then $a \in A \cap B$ or $a \in A \cap C$, namely $a \in A$ and $a \in B$, or $a \in A$ and $a \in C$. In either case $a \in A$, and $a \in B$ or $a \in C$, namely $a \in A$ and $a \in B \cup C$. Thus $a \in A \cap (B \cup C)$. \square

Proofs like these are not pleasant, and it is easy to get confused. Still, it is important to be comfortable working with sets in such a technical manner. Keeping a Venn diagram handy can help.

3.1.15 Exercise. Let A , B , and C be sets. Prove or disprove each of the following claims:

- (a) If $A \subseteq B$ if and only if $A \cap B = A$.
- (b) $A \cup (B \cap C) = (A \cup B) \cap (A \cup C)$.
- (c) Let X be a set containing A and B (for instance $X = A \cup B$). Then $X - (A \cup B) = (X - A) \cap (X - B)$.
- (d) Let X be a set containing A and B (for instance $X = A \cup B$). Then $X - (A \cap B) = (X - A) \cup (X - B)$.
- (e) If $A \cup C = B \cup C$, then $A = B$.
- (f) If $A \cap C = B \cap C$, then $A = B$.
- (g) If $A \cup C = B \cup C$ and $A \cap C = B \cap C$, then $A = B$.

Here's a more advanced set construction which will be useful in the next section.

3.1.16 Definition. Let A and B be sets. The **(Cartesian) product** of A and B is the set of ordered pairs $A \times B := \{(a, b) \mid a \in A, b \in B\}$.

3.1.17 Example.

- Let $A = \{1, 2, 3\}$ and $B = \{u, v\}$. Then

$$A \times B = \{(1, u), (1, v), (2, u), (2, v), (3, u), (3, v)\}.$$

- Let A be any set. Then $A \times \emptyset = \emptyset$.

3.1.18 Exercise. What is the product $\mathbb{R} \times \mathbb{R}$?

3.1.19 Exercise. Let A and B be finite sets. How many elements does $A \times B$ have?

3.1.20 Definition. Let A be a set and let $n \in \mathbb{Z}_{\geq 1}$. Then let $A^n = A \times A \times \cdots \times A$ be the product of n copies A . We call the elements of A^n **sequences in A of length n** . We can also define $A^\infty = A \times A \times \cdots$ to be the product of infinitely many copies of A . The elements of A^∞ are called **infinite sequences in A** .

3.1.21 Example. The set \mathbb{R}^n is the set of ordered n -tuples (r_1, \dots, r_n) of real numbers. The r_i need not be distinct. The set \mathbb{R}^∞ contains infinite sequences (r_1, r_2, \dots) .

3.1.22 Exercise. Let A be a finite set and let $n \in \mathbb{Z}_{\geq 1}$. How many elements does A^n have?

3.1.23 Exercise. Let A be a finite set with $|A| = 1$. How many elements does A^∞ have?

3.2 Functions

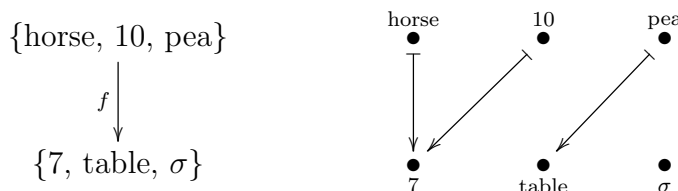
Now for what may be the most important object in all of mathematics:

3.2.1 Definition. Let A and B be sets. A **function** (or **map**) f from A to B , written $f: A \rightarrow B$ or $A \xrightarrow{f} B$, is an assignment of exactly one element of B to each element of A . We call A the **domain** of f and B the **codomain** of f . For every $a \in A$, we denote the element of B assigned to a by $f(a)$ and say $f(a)$ is the **image** of a and a is the **preimage** of $f(a)$. We also say “ a maps to $f(a)$ ” and denote this by $a \mapsto f(a)$. The set of all images of f is called the **image** (or **range**) of f ; it is a subset of B .

3.2.2 Note. To describe a function, either state explicitly which element of B is being assigned to each element of A , or give a “rule” (or several rules) that specifies the assignment. Functions between sets with few elements can be visualized by representing each element of the two sets as a dot and drawing arrows to depict the mapping.

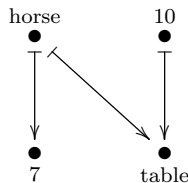
3.2.3 Example.

- $\{\text{horse}, 10, \text{pea}\} \xrightarrow{f} \{7, \text{table}, \sigma\}$ defined by $\text{horse} \mapsto 7$, $10 \mapsto 7$, $\text{pea} \mapsto \sigma$. We can visualize this function as



The image of f is $\{7, \text{table}\}$.

- The assignment



is *not* a function because it assigns more than one element (7 and “table”) to the element “horse”.

- For every set A , there is a unique function $\emptyset \xrightarrow{f} A$.
- For every set A , the **identity** function $A \xrightarrow{\text{id}_A} A$ is defined by $a \mapsto a$ for all $a \in A$.
- Let B be a set containing exactly one element. Then for every set A , there is a unique function $A \xrightarrow{f} B$, which must map every element of A to the only element of B .
- If $A \subseteq B$, then there is a natural “inclusion” function $A \xrightarrow{\iota} B$ with rule $\iota(a) = a$. The identity function id_A is the special case when $A = B$.
- $\mathbb{R} \xrightarrow{f} \mathbb{R}$ given by $x \mapsto x^2$. Or $\mathbb{R} \xrightarrow{g} \mathbb{R}_{\geq 0}$ with the same rule.

3.2.4 Exercise. Let A and B be finite sets. How many functions $A \longrightarrow B$ are there?

3.2.5 Definition. Two functions $A \xrightarrow{f} B$ and $C \xrightarrow{g} D$ are **equal** if $A = C$, $B = D$, and $f(a) = g(a)$ for all $a \in A = C$.

3.2.6 Example. The functions $\mathbb{R} \xrightarrow{f} \mathbb{R}$ given by $x \mapsto x^2$ and $\mathbb{R} \xrightarrow{g} \mathbb{R}_{\geq 0}$ given by $x \mapsto x^2$ are *not* equal because the codomains are different ($\mathbb{R} \neq \mathbb{R}_{\geq 0}$).

Here is a set you have used many times, but probably never defined:

3.2.7 Definition. The **graph** of a function $A \xrightarrow{f} B$ is the set

$$\Gamma_f = \{(a, f(a)) \in A \times B \mid a \in A\}.$$

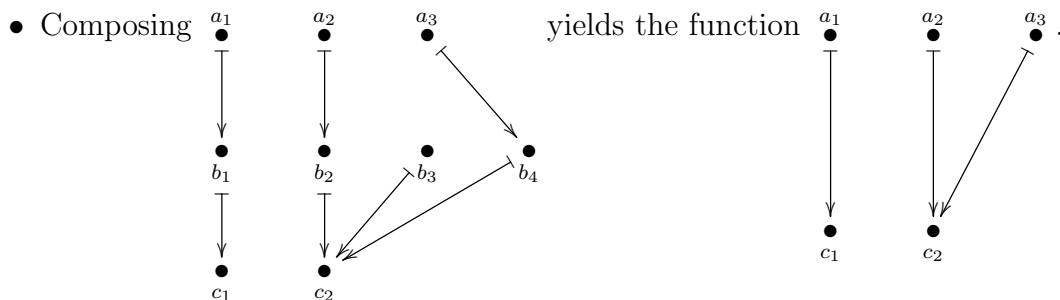
Think about what this definition is saying in the case when $A = B = \mathbb{R}$ and f is a familiar function from calculus like a polynomial (such as x^2), exponential (such as e^x), or trigonometric function (such as $\sin(x)$). You are used to drawing a “real axis” for the x -values, a “real axis” for the y -values, and then plotting the values $y = f(x)$ for all $x \in \mathbb{R}$. This is *exactly the same thing* as the graph!

3.2.8 Exercise. Why does the graph of a function from \mathbb{R} to \mathbb{R} always satisfy the “vertical line test”?

One of the most useful properties of functions is the possibility of composing them.

3.2.9 Definition. Given functions $A \xrightarrow{f} B$ and $B \xrightarrow{g} C$, the **composition** of f and g is the function $A \xrightarrow{g \circ f} C$ given by $(g \circ f)(a) = g(f(a))$.

3.2.10 Example.



- For any set A , $\text{id}_A \circ \text{id}_A = \text{id}_A$.
- Given $\mathbb{Z} \xrightarrow{f} \mathbb{Z}$ defined by $f(n) = n + 3$ and $\mathbb{Z} \xrightarrow{g} \mathbb{Z}$ defined by $g(n) = 2n$, $\mathbb{Z} \xrightarrow{g \circ f} \mathbb{Z}$ is the map $(g \circ f)(n) = g(f(n)) = g(n + 3) = 2(n + 3) = 2n + 6$. Composing in the other order, $\mathbb{Z} \xrightarrow{f \circ g} \mathbb{Z}$ is the map $(f \circ g)(n) = f(g(n)) = f(2n) = 2n + 3$. Note that $f \circ g \neq g \circ f$!

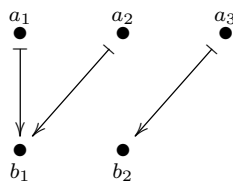
3.2.11 Exercise. Let $\mathbb{R} \xrightarrow{f} \mathbb{R}$ have rule $f(x) = x + 2$ and $\mathbb{R} \xrightarrow{g} \mathbb{R}$ have rule $g(x) = \frac{x}{3}$. Compute the rule for $g \circ f$ and the rule for $f \circ g$. Does $g \circ f = f \circ g$?

3.3 Injectivity

3.3.1 Definition. A function $A \xrightarrow{f} B$ is **injective** (**one-to-one**) if f maps no two distinct elements of A to the same element of B . Expressed in notation, this means that if $a_1, a_2 \in A$ and $a_1 \neq a_2$, then $f(a_1) \neq f(a_2)$. Another common way of expressing this is by the contrapositive: if $f(a_1) = f(a_2)$, then $a_1 = a_2$. Yet another way to define injectivity of f is to require that every element of B has 0 or 1 preimage.

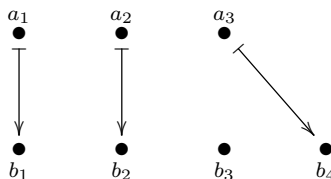
3.3.2 Example.

- The function



is not injective because b_1 has two preimages, namely a_1 and a_2 .

- The function



is injective because every b_i has 0 or 1 preimage.

3.3.3 Exercise. How can you determine whether a function $\mathbb{R} \xrightarrow{f} \mathbb{R}$ is injective by looking at its graph? Is $f(x) = 2x$ injective? How about $f(x) = x^2$, $f(x) = x^3 - x$, and $f(x) = \sin(x)$?

3.3.4 Note. To prove a function $A \xrightarrow{f} B$ is not injective, find two distinct elements of A that have the same image. In other words, find $a_1, a_2 \in A$ with $a_1 \neq a_2$ such that $f(a_1) = f(a_2)$. To prove f is injective, you have to check that for arbitrary $a_1, a_2 \in A$ with $a_1 \neq a_2$, you must have $f(a_1) \neq f(a_2)$. Or, equivalently, you can prove that for any $a_1, a_2 \in A$, if $f(a_1) = f(a_2)$, then $a_1 = a_2$.

3.3.5 Example.

- To prove that $\mathbb{R} \xrightarrow{f} \mathbb{R}$ defined by $f(x) = x^2$ is not injective, we simply note that $f(1) = 1 = f(-1)$.
- To prove that $\mathbb{R} \xrightarrow{f} \mathbb{R}$ defined by $f(x) = 3x - 4$ is injective, we suppose that there are $x_1, x_2 \in \mathbb{R}$ such that $f(x_1) = f(x_2)$, namely that $3x_1 - 4 = 3x_2 - 4$. Adding 4 to both sides and dividing by 3, we see that $x_1 = x_2$. Thus f is injective.

3.3.6 Mistake. A function $A \xrightarrow{f} B$ being injective does *not* mean that the function assigns one and only one element of B to each element of A . Every function does that!

3.3.7 Exercise. Let $A \xrightarrow{f} B$ be a function. Prove or disprove: there is a subset $A' \subseteq A$ such that restricting the domain of f to A' yields an injective function $A' \xrightarrow{f} B$.

An important question is: how does injectivity interact with composition of functions?

3.3.8 Claim. Let A, B, C be sets, let $A \xrightarrow{f} B$ and $B \xrightarrow{g} C$ be functions, and let $A \xrightarrow{g \circ f} C$ be the composition of f and g . If f and g are injective, then $g \circ f$ is injective.

We'll give two proofs, one for each common way of expressing injectivity.

Proof #1. Suppose $a_1, a_2 \in A$ satisfy $a_1 \neq a_2$. Then $f(a_1) \neq f(a_2)$ by the injectivity of f , so $g(f(a_1)) \neq g(f(a_2))$ by the injectivity of g . But this means $(g \circ f)(a_1) \neq (g \circ f)(a_2)$, so $g \circ f$ is injective. \square

Proof #2. Suppose $a_1, a_2 \in A$ satisfy $(g \circ f)(a_1) = (g \circ f)(a_2)$, namely $g(f(a_1)) = g(f(a_2))$. Then $f(a_1) = f(a_2)$ by the injectivity of g , from which it follows that $a_1 = a_2$ by the injectivity of f . Thus $g \circ f$ is injective. \square

3.3.9 Exercise. Let $A \xrightarrow{f} B$ and $B \xrightarrow{g} C$ be functions, and let $A \xrightarrow{g \circ f} C$ be the composition of f and g . Prove or disprove the following claims:

- (a) If $g \circ f$ is injective, then f is injective.

(b) If $g \circ f$ is injective, then g is injective.

3.3.10 Exercise. Let $A \xrightarrow{g} B$ and $A \xrightarrow{h} B$ be functions and let $B \xrightarrow{f} C$ be an injective function. Prove or disprove: if $f \circ g = f \circ h$, then $g = h$.

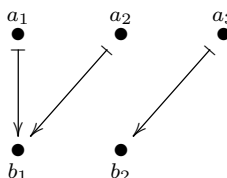
3.3.11 Exercise. Let $B \xrightarrow{g} C$ and $B \xrightarrow{h} C$ be functions and let $A \xrightarrow{f} B$ be an injective function. Prove or disprove: if $g \circ f = h \circ f$, then $g = h$.

3.4 Surjectivity

3.4.1 Definition. A function $A \xrightarrow{f} B$ is **surjective (onto)** if the image of f is the entire codomain B . This means that for every $b \in B$ there is an element $a \in A$ such that $f(a) = b$.

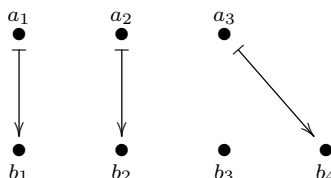
3.4.2 Example.

- The function



is surjective because both b_1 and b_2 have at least one preimage. More precisely, b_1 has two preimages and b_2 has one preimage.

- The function



is not surjective because b_3 has no preimage.

3.4.3 Exercise. How can you tell whether a function $\mathbb{R} \xrightarrow{f} \mathbb{R}$ is surjective by looking at its graph? Is $f(x) = 2x$ surjective? How about $f(x) = x^2$, $f(x) = x^3 - x$, and $f(x) = \sin(x)$?

3.4.4 Note. To prove that a function $A \xrightarrow{f} B$ is not surjective, you have to find an element of the codomain B that has no preimage, and prove that it has no preimage, namely that for every $a \in A$, $f(a) \neq b$. To prove that f is surjective, you have to show that an arbitrary element $b \in B$ has a preimage, namely that there is an $a \in A$ such that $f(a) = b$.

3.4.5 Example.

- To prove that $\mathbb{R} \xrightarrow{f} \mathbb{R}$ defined by $f(x) = x^2$ is not surjective, we note that for any x , $x^2 \geq 0$. Thus the negative element $-1 \in \mathbb{R}$ has no preimage.
- To prove that $\mathbb{R} \xrightarrow{f} \mathbb{R}$ defined by $f(x) = 3x - 4$ is surjective, choose an arbitrary element $y \in \mathbb{R}$ of the codomain. We want to find $x \in \mathbb{R}$ such that $f(x) = y$. To do this, we simply solve the equation $3x - 4 = y$ for x , which yields $x = \frac{y+4}{3}$. Since $f(\frac{y+4}{3}) = y$, every element of the codomain has at least one preimage.

3.4.6 Exercise. Let $A \xrightarrow{f} B$ be a function. Prove or disprove: there is a subset $B' \subseteq B$ such that restricting the codomain of f to B' yields a surjective function $A \xrightarrow{f} B'$.

Like injectivity, surjectivity is preserved under composition of functions.

3.4.7 Claim. Let $A \xrightarrow{f} B$ and $B \xrightarrow{g} C$ be functions. If f and g are surjective, then $g \circ f$ is surjective.

Proof. Let $c \in C$ be arbitrary. Since g is surjective, there is some $b \in B$ such that $g(b) = c$. Since f is surjective, there is some $a \in A$ such that $f(a) = b$. Thus $(g \circ f)(a) = g(f(a)) = g(b) = c$, so $g \circ f$ is surjective. \square

3.4.8 Exercise. Let $A \xrightarrow{f} B$ and $B \xrightarrow{g} C$ be functions, and let $A \xrightarrow{g \circ f} C$ be the composition of f and g . Prove or disprove the following claims:

- If $g \circ f$ is surjective, then f is surjective.
- If $g \circ f$ is surjective, then g is surjective.
- If $g \circ f$ is injective and f is surjective, then g is injective.
- If $g \circ f$ is surjective and g is injective, then f is surjective.

3.4.9 Exercise. Let $A \xrightarrow{g} B$ and $A \xrightarrow{h} B$ be functions and let $B \xrightarrow{f} C$ be a surjective function. Prove or disprove: if $f \circ g = f \circ h$, then $g = h$.

3.4.10 Exercise. Let $B \xrightarrow{g} C$ and $B \xrightarrow{h} C$ be functions and let $A \xrightarrow{f} B$ be a surjective function. Prove or disprove: if $g \circ f = h \circ f$, then $g = h$.

3.5 Bijectivity and Inverses

3.5.1 Definition. A function $A \xrightarrow{f} B$ is **bijective** if it is both injective and surjective. This means that f establishes a perfect correspondence of the elements of A with the elements of B .

3.5.2 Example.

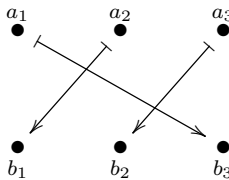
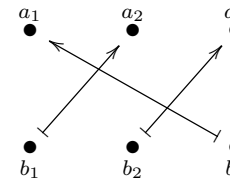
- The function $\mathbb{R} \xrightarrow{f} \mathbb{R}$ with rule $f(x) = 3x - 4$ is bijective. We proved above that f is both injective and surjective.
- For any set A , the identity function $A \xrightarrow{\text{id}_A} A$ is clearly bijective.
- The function $\mathbb{Z} \xrightarrow{f} \mathbb{Z}$ given by $f(n) = n + 3$ is bijective.
- The exponential function $\mathbb{R} \xrightarrow{g} \mathbb{R}_{>0}$ given by $x \mapsto 2^x$ is bijective. We'll prove this below.

3.5.3 Note. Proving a function is bijective by checking both injectivity and surjectivity can be tedious. An easier way to prove bijectivity is to find an inverse function, as we will discuss below.

3.5.4 Exercise. How can you use the graph of a function $\mathbb{R} \xrightarrow{f} \mathbb{R}$ to determine whether it is bijective?

3.5.5 Definition. Let $A \xrightarrow{f} B$ and $B \xrightarrow{g} A$ be functions such that $g \circ f = \text{id}_A$ and $f \circ g = \text{id}_B$. Then we say f and g are **inverse** functions. We also say f is **invertible** and that f has g as an **inverse**, and often denote that inverse by f^{-1} .

3.5.6 Example.

- The inverse of  is .
- The inverse of id_A is id_A .

3.5.7 Note. If a function $\mathbb{R} \xrightarrow{f} \mathbb{R}$ is given by a rule $f(x)$, then the way to find an inverse of f , if it exists, is to set $y = f(x)$ and solve for x in terms of y . If this is possible and the result is a function $\mathbb{R} \xrightarrow{g} \mathbb{R}$, then check that the compositions $g \circ f$ and $f \circ g$ are equal to the identity function $\text{id}_{\mathbb{R}}$. If they are, then you have found an inverse! The same procedure works if the domain and codomain allow you to “solve for x ”; usually you will need sets in which you can at least do addition and multiplication. When we work over \mathbb{Z} , we usually use ‘ n ’ instead of ‘ x ’ and ‘ m ’ instead of ‘ y ’.

3.5.8 Example.

- The function $\mathbb{R} \xrightarrow{f} \mathbb{R}$ with rule $f(x) = 2x$ has an inverse. To find it, set $y = 2x$ and solve for x , which yields $x = \frac{y}{2}$. So $\mathbb{R} \xrightarrow{g} \mathbb{R}$ with rule $g(y) = \frac{y}{2}$ is the inverse function. Check that both compositions give the identity!

- The function $\mathbb{Z} \xrightarrow{f} \mathbb{Z}$ with rule $f(n) = 2n$ does not have an inverse. When we try to solve $m = 2n$ for n , we cannot divide by 2 since we are working over \mathbb{Z} . Indeed, $g(m) = \frac{m}{2}$ will not yield a function $\mathbb{Z} \xrightarrow{g} \mathbb{Z}$ because the image is not always an integer.
- The inverse of the function $\mathbb{Z} \xrightarrow{f} \mathbb{Z}$ with rule $f(n) = n + 3$ is the function $\mathbb{Z} \xrightarrow{f^{-1}} \mathbb{Z}$ with rule $f^{-1}(m) = m - 3$.
- The exponential function $\mathbb{R} \xrightarrow{g} \mathbb{R}_{>0}$ with rule $x \mapsto 2^x$ has an inverse! Set $y = 2^x$ and take \log_2 of both sides to obtain $x = \log_2 y$. Thus the inverse function is $\mathbb{R}_{>0} \xrightarrow{g^{-1}} \mathbb{R}$ with rule $g^{-1}(y) = \log_2 y$.

It turns out that invertible functions and bijective functions are one and the same!

3.5.9 Proposition. *A function $A \xrightarrow{f} B$ has an inverse if and only if it is bijective.*

Proof. (\implies): Let $B \xrightarrow{g} A$ be the inverse of f . Suppose $f(a_1) = f(a_2)$ for $a_1, a_2 \in A$. Then

$$a_1 = \text{id}_A(a_1) = (g \circ f)(a_1) = g(f(a_1)) = g(f(a_2)) = (g \circ f)(a_2) = \text{id}_A(a_2) = a_2,$$

so f is injective. To see that f is surjective, suppose $b \in B$. Then $g(b) \in A$ and

$$f(g(b)) = (f \circ g)(b) = \text{id}_B(b) = b,$$

so f is surjective. Thus f is bijective.

(\impliedby): Define $B \xrightarrow{g} A$ to be the map that assigns to each element $b \in B$ the unique element $a \in A$ such that $f(a) = b$. Such an element a exists since f is surjective, and is unique since f is injective. Now we use this definition of g to compute

$$(g \circ f)(a) = g(f(a)) = a, \quad (f \circ g)(b) = f(g(b)) = b.$$

Thus $g \circ f = \text{id}_A$ and $f \circ g = \text{id}_B$, so g is the inverse of f . □

We can use Proposition 3.5.9 to prove that functions are bijective. For instance:

3.5.10 Claim. *The function $\mathbb{R} \xrightarrow{f} \mathbb{R}$ with rule $f(x) = x^3$ is bijective.*

Proof. We will construct a function that is the inverse of f . Since an invertible function must be bijective, this will prove that f is bijective. Consider the function $\mathbb{R} \xrightarrow{g} \mathbb{R}$ mapping $x \mapsto x^{1/3}$, which is a function since the cube root of any real number is uniquely defined. We check the compositions. $\mathbb{R} \xrightarrow{g \circ f} \mathbb{R}$ is the map $x \mapsto (g \circ f)(x) = g(f(x)) = g(x^3) = (x^3)^{1/3} = x$, so $g \circ f = \text{id}_{\mathbb{R}}$. Similarly, $\mathbb{R} \xrightarrow{f \circ g} \mathbb{R}$ is the map $(f \circ g)(x) = f(g(x)) = f(x^{1/3}) = (x^{1/3})^3 = x$, so $f \circ g = \text{id}_{\mathbb{R}}$. □

3.5.11 Note (How to determine whether a function $\mathbb{R} \xrightarrow{f} \mathbb{R}$ with a “nice” rule is bijective). First, determine whether the function is most likely injective, surjective, and bijective by looking at the graph Γ_f of f :

- f is injective if and only if every horizontal line in \mathbb{R}^2 intersects Γ_f in *at most one* point (we say Γ_f satisfies the “horizontal line test”).
- f is surjective if and only if every horizontal line in \mathbb{R}^2 intersects Γ_f in *at least one* point.
- Thus f is bijective if and only if every horizontal line in \mathbb{R}^2 intersects Γ_f in *exactly one* point.

Studying the graph usually not a good way to *prove* that a function is bijective because functions with complicated rules may have small wiggles you can’t see or surprising behavior for large values of the domain. So the graph is best used only as a guide, showing you what you should be trying to prove! We do proofs as described above:

- To prove f is not injective, find two distinct elements of the domain that have the same image under f . If f is not injective, then it is certainly not bijective.
- To prove f is not surjective, find an element of the codomain that is not in the image of f . If f is not surjective, then it is not bijective.
- To prove f is bijective, construct the inverse function as explained in Note 3.5.7 and check the compositions. Or prove injectivity and surjectivity directly as explained in Notes 3.3.4 and 3.4.4.

3.5.12 Example. Determine whether each of these functions from \mathbb{R} to \mathbb{R} is bijective:

- $f(x) = 3x - 4$.

The graph Γ_f is a (non-horizontal, non-vertical) line, so every horizontal line intersects Γ_f in exactly one point. Thus f appears to be bijective. Indeed, we proved above that f is both injective and surjective, hence bijective. An easier way to prove f is bijective is to show that the function $\mathbb{R} \xrightarrow{g} \mathbb{R}$ with rule $g(y) = \frac{y+4}{3}$ is an inverse of f . The rule for g makes sense for all $y \in \mathbb{R}$, so g is a function. We now check the compositions:

$$(g \circ f)(x) = g(f(x)) = g(3x - 4) = \frac{(3x - 4) + 4}{3} = x, \quad \text{so } g \circ f = \text{id}_{\mathbb{R}};$$

$$(f \circ g)(y) = f(g(y)) = f\left(\frac{y+4}{3}\right) = 3 \cdot \frac{y+4}{3} - 4 = y, \quad \text{so } f \circ g = \text{id}_{\mathbb{R}}.$$

Thus g is an inverse of f , so f must be bijective.

- $f(x) = -3x^2 + 7$.

Γ_f is a parabola, which looks neither injective (it is symmetric about the y -axis) nor surjective (it has a maximum value of 7). Since $-3x^2 \leq 0$ for all x , $f(x) \leq 7$ for all x . Thus the image of f does not contain any real number > 7 , so f is not surjective, hence f is not bijective. In fact, f is also not injective. To prove this, it is enough to give two distinct values of \mathbb{R} that map to the same element, and this is quite easy: $f(1) = f(-1)$.

- $f(x) = (x + 1)/(x + 2)$.

f is not a function from \mathbb{R} to \mathbb{R} because it is not defined at -2 ! So f certainly cannot be a bijective function.

- $f(x) = x^5 + 1$.

From the graph, f appears to be bijective. Let's prove f is bijective by finding the inverse g . Setting $y = x^5 + 1$, solving for x yields $x = (y - 1)^{\frac{1}{5}}$. The rule $g(y) = (y - 1)^{\frac{1}{5}}$ makes sense for all $y \in \mathbb{R}$, so $\mathbb{R} \xrightarrow{g} \mathbb{R}$ is a function. Prove that g is an inverse of f by checking the compositions.

3.5.13 Exercise. Prove or disprove that each of the following functions is bijective:

(a) $\mathbb{Z} \xrightarrow{f} \mathbb{Z}$ with rule $f(n) = 4 - n$.

(b) $\mathbb{Z} \xrightarrow{f} \mathbb{Z}$ with rule $f(n) = n^3$.

(c) $\mathbb{R} \xrightarrow{f} \mathbb{R}_{\geq 1}$ with rule $f(x) = x^2 + 1$.

(d) $\mathbb{R}_{>0} \xrightarrow{f} \mathbb{R}_{>0}$ with rule $f(x) = \frac{1}{x}$.

(e) $\mathbb{R} \xrightarrow{f} \mathbb{R}$ with rule $f(x) = \sin x$.

(f) $\left[-\frac{\pi}{2}, \frac{\pi}{2}\right] \xrightarrow{f} [-1, 1]$ with rule $f(x) = \sin x$.

Chapter 4

Number Theory

Number theory is the study of the integers, \mathbb{Z} . We will start with the basic definitions and end by learning how number theory is used to encrypt data online.

4.1 Divisibility and Primes

4.1.1 Definition. If $a, b \in \mathbb{Z}$ with $a \neq 0$, then a **divides** b , written $a \mid b$, if there is an integer c such that $b = ac$. We then call a a **factor** or **divisor** of b , and b a **multiple** of a .

4.1.2 Mistake. “ $a \mid b$ ” is the statement, which can be true or false, that a and b have a certain relationship. It is not the same thing as dividing a/b or $\frac{a}{b}$, which is a rational number! A statement like “ $2 \mid 6 = 3$ ” is nonsense. The “division” you’re used to is not used in number theory, since it is not a valid operation in \mathbb{Z} . Thus when we talk about a number “dividing” another number, we will mean the definition above, unless otherwise specified.

4.1.3 Example.

- 6 divides 0, 6, -6 , 12, and -12 because these integers can be written as $6 \cdot 0$, $6 \cdot 1$, $6 \cdot (-1)$, $6 \cdot 2$, and $6 \cdot (-2)$.
- Let $n \in \mathbb{Z}$. Then $2 \mid n$ if and only if n is even.
- $1 \mid a$ for every $a \in \mathbb{Z}$.
- $a \mid 0$ for every nonzero $a \in \mathbb{Z}$.

Some easy but critical properties of divisibility are:

4.1.4 Proposition. Let $a, b, c \in \mathbb{Z}$, with $a \neq 0$. Then

- (i) if $a \mid b$ and $a \mid c$, then $a \mid (b + c)$;
- (ii) if $a \mid b$, then $a \mid bc$ for all integers c ;

(iii) if $a \mid b$ and $b \mid c$, then $a \mid c$.

Proof. For (i), since $a \mid b$, there is $s \in \mathbb{Z}$ such that $b = as$. Since $a \mid c$, there is $t \in \mathbb{Z}$ such that $c = at$. Then $b + c = as + at = a(s + t)$, so $a \mid (b + c)$.

(ii) and (iii) are left to you as an exercise! \square

4.1.5 Exercise. Prove (ii) and (iii) of Proposition 4.1.4.

4.1.6 Corollary. Let $a, b, c \in \mathbb{Z}$ with $a \neq 0$. If $a \mid b$ and $a \mid c$, then $a \mid mb + nc$ for any $m, n \in \mathbb{Z}$.

Now let's get reacquainted with our old friends from elementary school, the prime numbers.

4.1.7 Definition. An integer $p > 1$ is **prime** if its only positive factors are 1 and p . An integer greater than 1 is called **composite** if it is not prime.

Note: 1 is not prime! We'll see in a moment why we want to exclude 1.

4.1.8 Example. The set of primes less than 100 is

$\{2, 3, 5, 7, 11, 13, 17, 19, 23, 29, 31, 37, 41, 43, 47, 53, 59, 61, 67, 71, 73, 79, 83, 89, 97\}$.

Why are primes important? They are the building blocks of all positive integers!

4.1.9 Theorem (Fundamental theorem of arithmetic). Let $n \in \mathbb{Z}_{\geq 2}$. Then there is a unique sequence of primes $p_1 \leq p_2 \leq \cdots \leq p_r$ such that $n = p_1 p_2 \cdots p_r$.

We call the expression $n = p_1 p_2 \cdots p_r$ the **prime factorization** of n . Note that if 1 were a prime, then the prime factorization would not be unique since we could take any number of 1's as factors!

Proof. Sadly we have to postpone this crucial proof until we have learned proof by induction. For the proof, see Section 5.2. \square

To find the prime factorization of a number n , check whether any of the primes less than n divide n . As soon as you find a prime p such that $p \mid n$, add p to your factorization, and repeat the process for $\frac{n}{p}$.

4.1.10 Example. $2 = 2$ is prime! $6 = 2 \cdot 3$. $63 = 3 \cdot 3 \cdot 7$. $100 = 2 \cdot 2 \cdot 5 \cdot 5$.

4.1.11 Note. An integer > 1 is composite if and only if its prime factorization has at least two primes (possibly the same prime twice)!

4.1.12 Exercise. Find the prime factorizations of

(a) 88;

(b) 126;

(c) 1111;

(d) 909,090.

How many primes are there? Plenty!

4.1.13 Theorem. *There are infinitely many primes.*

Proof. Assume for contradiction that there are only finitely many primes, say $\{p_1, p_2, \dots, p_n\}$. Then consider

$$N = p_1 p_2 \cdots p_n + 1.$$

Since N has a prime factorization, there is a p_i that divides N . But p_i also divides $p_1 p_2 \cdots p_n$, so p_i divides $N - p_1 p_2 \cdots p_n = 1$ by Corollary 4.1.6. This is a contradiction since no prime can divide 1. \square

4.1.14 Exercise. We can use the **sieve of Eratosthenes** to compute all the primes less than a given number. For instance, to find all primes ≤ 30 :

- **Step 1:** Write the integers from 2 to 30 in ascending order.
- **Step 2:** Circle the lowest unmarked number (it is prime!) and cross out all other multiples of that number (they are composite!).
- **Step 3:** Repeat Step 2 until all numbers are circled or crossed out. The circled numbers are the primes!

4.1.15 Exercise. Let $n \in \mathbb{Z}_{\geq 2}$. Prove or disprove the following claim: If n is composite, then it has a prime factor p such that $p \leq \sqrt{n}$.

4.1.16 Exercise. Let $m, n \in \mathbb{Z}$, let $d \in \mathbb{Z}_{\geq 1}$, and let $p \in \mathbb{Z}_{\geq 2}$ be prime. Prove or disprove each of the following claims:

- (a) If $d \nmid n$, then $d \nmid mn$.
- (b) If $d \mid n$, then $d \mid mn$.
- (c) If $dn \mid mn$ and $n \neq 0$, then $d \mid m$.
- (d) If $d \mid mn$, then $d \mid m$ or $d \mid n$.
- (e) If $p \mid mn$, then $p \mid m$ or $p \mid n$.

4.2 Euclidean Algorithm

In this section, we will introduce a crucial technical tool called the Euclidean algorithm. As usual, we begin with definitions.

4.2.1 Definition. For $a, b \in \mathbb{Z}$, the largest integer d such that $d \mid a$ and $d \mid b$ is called the **greatest common divisor** of a and b and is denoted by $\gcd(a, b)$. We say a and b are **relatively prime** if $\gcd(a, b) = 1$.

4.2.2 Note. Since all nonzero integers divide 0, $\gcd(0, 0)$ is not defined, but $\gcd(a, 0) = a$ for any $a \in \mathbb{Z}_{\geq 1}$.

4.2.3 Note. You can compute $\gcd(a, b)$ by taking the “intersection” of the prime factorizations of a and b . For example, $36 = 2 \cdot 2 \cdot 3 \cdot 3$ and $84 = 2 \cdot 2 \cdot 3 \cdot 7$ have $\gcd(36, 84) = 2 \cdot 2 \cdot 3 = 12$.

4.2.4 Exercise. List the positive integers less than 12 that are relatively prime to 12.

4.2.5 Definition. The **least common multiple** of $a, b \in \mathbb{Z}_{\geq 1}$, denoted $\text{lcm}(a, b)$, is the smallest positive integer m such that $a \mid m$ and $b \mid m$.

4.2.6 Note. You can also compute $\text{lcm}(a, b)$ from the prime factorizations: take the “union” of the prime factorizations for a and b . For instance, $\text{lcm}(36, 84) = 2 \cdot 2 \cdot 3 \cdot 3 \cdot 7$.

4.2.7 Proposition. Let $a, b \in \mathbb{Z}_{\geq 1}$. Then $ab = \gcd(a, b) \cdot \text{lcm}(a, b)$.

Idea of proof. Use the description of $\gcd(a, b)$ and $\text{lcm}(a, b)$ in terms of prime factorizations. \square

4.2.8 Exercise. For each of the following pairs of integers, find the gcd and lcm and check that Proposition 4.2.7 holds:

- (a) 12, 24;
- (b) 15, 25;
- (c) 100, 100;
- (d) 1, 22.

Now let’s take a journey back in time and recall division with remainder from elementary school:

$$\frac{23}{4} = 5 + \frac{3}{4} = 5 \text{ R}3.$$

Since we don’t have fractions when working with just \mathbb{Z} , we multiply through by 4 to get

$$23 = 4 \cdot 5 + 3.$$

This is the notion of “division” that we will use!

4.2.9 Theorem (Division algorithm). *Let $a, d \in \mathbb{Z}$ with $d > 0$. Then there are unique integers q, r with $0 \leq r < d$ such that $a = qd + r$.*

Idea of proof. Keep adding or subtracting d from a until you end up in the set $\{0, 1, \dots, d-1\}$. Think of dividing a by d and finding the remainder! \square

4.2.10 Definition. In the equality $a = qd + r$ in the division algorithm, d is the **divisor**, a is the **dividend**, q is the **quotient**, and r is the **remainder**.

The remainder is so important that all of Section 4.3 will be devoted to it!

4.2.11 Example. Note that the remainder must be non-negative! So applying the division algorithm to $a = -7$ and $d = 6$, we get $q = -2$ and $r = 5$, namely $-7 = -2 \cdot 6 + 5$.

Given a large integer n , it is hard to find its prime factorization. The obvious method is to test whether each prime smaller than n is a divisor of n , but there are too many primes to test! However, if we are more modest and only want to know the gcd of two positive integers a, b , then there is an efficient method for finding $\gcd(a, b)$, called the **Euclidean algorithm**. The idea is to repeatedly use the division algorithm, first on a and b to get $a = qb + r$, then on the previous divisor b and remainder r , then again on the previous divisor and remainder, and so on, until the remainder is 0. It turns out that the last nonzero remainder is the gcd of a and b . Here is an example:

4.2.12 Example. Find $\gcd(123, 45)$ using the Euclidean algorithm. Running the algorithm, we compute:

$$\begin{array}{ll} 123 = 2 \cdot 45 + 33 & \text{div. alg. for 123 and 45} \\ 45 = 1 \cdot 33 + 12 & \text{div. alg. for divisor (45) and rem. (33) of prev. equation} \\ 33 = 2 \cdot 12 + 9 & \text{div. alg. for divisor (33) and rem. (12) of prev. equation} \\ 12 = 1 \cdot 9 + \boxed{3} & \text{div. alg. for divisor (12) and rem. (9) of prev. equation} \\ 9 = 3 \cdot 3 + 0 & \text{div. alg. for divisor (9) and rem. (3) of prev. equation.} \end{array}$$

The last nonzero remainder, boxed above, is 3, so $\gcd(123, 45) = 3$.

Here are the formal steps in the algorithm:

4.2.13 Algorithm (Euclidean algorithm). Find the gcd of two positive integers a and b .

- **Step 1:** Use the division algorithm on a and b to obtain an equation $a = q_0b + r_0$, where $0 \leq r_0 < b$.
- **Step 2:** Use the division algorithm on b and r_0 to obtain an equation $b = q_1r_0 + r_1$, where $0 \leq r_1 < r_0$.
- **Step 3:** For each $i \geq 0$, use the division algorithm on r_i and r_{i+1} to obtain an equation $r_i = q_{i+2}r_{i+1} + r_{i+2}$, where $0 \leq r_{i+2} < r_{i+1}$. Stop when the remainder becomes 0, which is guaranteed to happen after finitely many steps since the r_i 's are strictly decreasing.

- **Step 4:** The last nonzero remainder equals $\gcd(a, b)$.

The reason that the Euclidean algorithm works is that the gcd is preserved from one step to the next. In the above example, this means that $\gcd(123, 45)$ equals the divisor and remainder in each equation, namely

$$\gcd(123, 45) = \gcd(45, 33) = \gcd(33, 12) = \gcd(12, 9) = \gcd(9, 3) = \gcd(3, 0).$$

But the gcd any any positive integer and 0 is that integer, so all of these gcd's equal 3. The reason all of these gcd's are equal is because we can repeatedly apply:

4.2.14 Lemma. *Let $a = qb + r$, where $a, b, q, r \in \mathbb{Z}$. Then $\gcd(a, b) = \gcd(b, r)$.*

Proof. It suffices to show that the common divisors of a and b are the same as the common divisors of b and r . If $d \mid a$ and $d \mid b$, then $d \mid a - qb$ by Corollary 4.1.6, namely $d \mid r$. Similarly, if $d \mid b$ and $d \mid r$, then $d \mid qb + r$, namely $d \mid a$. \square

The Euclidean algorithm actually gives us more than just the gcd of a and b . It allows us to express that gcd as a sum of multiples of a and b , which will be very, very useful for solving linear congruences in the next section!

4.2.15 Theorem (Bézout's identity). *For $a, b \in \mathbb{Z}_{\geq 1}$, there exist $s, t \in \mathbb{Z}$ such that $\gcd(a, b) = sa + tb$.*

We say that $\gcd(a, b)$ can be expressed as a **linear combination** of a and b , with integer coefficients.

Idea of proof. Run the Euclidean algorithm on a and b and use the output to find s and t . Let's see how this is done for the above example, namely $a = 123$ and $b = 45$. The output of the Euclidean algorithm is

$$123 = 2 \cdot 45 + 33 \qquad 33 = 123 - 2 \cdot 45 \qquad (4)$$

$$45 = 1 \cdot 33 + 12 \qquad 12 = 45 - 1 \cdot 33 \qquad (3)$$

$$33 = 2 \cdot 12 + 9 \qquad 9 = 33 - 2 \cdot 12 \qquad (2)$$

$$12 = 1 \cdot 9 + \boxed{3} \qquad \boxed{3} = 12 - 1 \cdot 9 \qquad (1)$$

$$9 = 3 \cdot 3 + 0,$$

where in the right column we have solved for the remainders in the equations on the left.

With these equations in front of us, we use a method called **back substitution**. We start with equation (1), which expresses 3 as a linear combination of 12 and 9. We want 3 as a linear combination of 123 and 45, so we use equations (2), (3), and (4) once each, in order, to substitute for 9, 12, and 33 until we get 3 as a linear combination of 123 and

45. Here is the computation:

$$\begin{array}{ll}
 3 = 12 - 1 \cdot 9 & \text{copy (1)} \\
 = 12 - 1 \cdot (33 - 2 \cdot 12) & \text{substitute the 9 using (2)} \\
 = -33 + 3 \cdot 12 & \text{group multiples of 33 and 12} \\
 = -33 + 3 \cdot (45 - 1 \cdot 33) & \text{substitute the 12 using (3)} \\
 = 3 \cdot 45 - 4 \cdot 33 & \text{group multiples of 45 and 33} \\
 = 3 \cdot 45 - 4 \cdot (123 - 2 \cdot 45) & \text{substitute the 33 using (4)} \\
 = \boxed{-4} \cdot 123 + \boxed{9} \cdot 45 & \text{group multiples of 123 and 45.}
 \end{array}$$

Then we just read off the coefficients of 123 and 45, namely $s = -4$ and $t = 9$.

The process is similar for general a and b . Take the output of the Euclidean algorithm, ignore the last equation, and solve all other equations for their (nonzero) remainders. Start with the equation for the last nonzero remainder, namely $\gcd(a, b)$, and substitute using the previous equation. Group like terms, substitute, and repeat, until you get $\gcd(a, b)$ as a linear combination of a and b . \square

4.2.16 Example. Use the Euclidean algorithm with back substitution to express $\gcd(1260, 41)$ as a linear combination of 1266 and 41.

The Euclidean algorithm yields

$$\begin{array}{ll}
 1266 = 30 \cdot 41 + 36 & 36 = 1266 - 30 \cdot 41 \quad (3) \\
 41 = 1 \cdot 36 + 5 & 5 = 41 - 1 \cdot 36 \quad (2) \\
 36 = 7 \cdot 5 + \boxed{1} & 1 = 36 - 7 \cdot 5 \quad (1) \\
 7 = 7 \cdot 1 + 0 &
 \end{array}$$

Now back substitution yields

$$\begin{array}{ll}
 1 = 36 - 7 \cdot 5 & \text{copy (1)} \\
 = 36 - 7 \cdot (41 - 1 \cdot 36) & \text{substitute the 5 using (2)} \\
 = -7 \cdot 41 + 8 \cdot 36 & \text{group multiples of 41 and 36} \\
 = -7 \cdot 41 + 8 \cdot (1266 - 30 \cdot 41) & \text{substitute the 36 using (3)} \\
 = \boxed{8} \cdot 1266 \boxed{-247} \cdot 41 & \text{group multiples of 1266 and 41}
 \end{array}$$

In terms of Bézout's identity, $s = 8$ and $t = -247$.

4.2.17 Exercise. Use the Euclidean algorithm and back substitution to express $\gcd(a, b)$ as a linear combination of a and b , where the values of a and b are:

- (a) 10, 11
- (b) 9, 11
- (c) 21, 44

- (d) 33, 44
- (e) 36, 48
- (f) 34, 55
- (g) 35, 78
- (h) 117, 213

The Euclidean algorithm and back substitution are essential. We will rely heavily on both of them in the next section, so make sure they are a well-honed tool in your arsenal!

4.3 Modular Arithmetic

In some situations we care only about remainders. For instance, what time will it be 50 hours from now? One way to find the answer is to compute the remainder of 50 divided by 24, and add that to the current time. Working with remainders is surprisingly interesting mathematically, and has amazingly powerful applications to cryptography, as we will see in Section 4.5.

Recall the division algorithm 4.2.9:

Theorem (Division algorithm). *Let $a, d \in \mathbb{Z}$ with $d > 0$. Then there are unique integers q, r with $0 \leq r < d$ such that $a = qd + r$.*

Recall that r is the unique “remainder” when a is divided by d and that $0 \leq r < d$. We will often say that r is the **remainder of a modulo d** .

4.3.1 Definition. If $a, b \in \mathbb{Z}$ and $m \in \mathbb{Z}_{\geq 1}$, then a is **congruent to b modulo m** , written $a \equiv b \pmod{m}$, if a and b have the same remainder when divided by m (modulo m). We say that $a \equiv b \pmod{m}$ is a **congruence** (rather than an *equation*) and that m is its **modulus**. If a and b are not congruent modulo m , we write $a \not\equiv b \pmod{m}$.

4.3.2 Example. $11 \equiv 7 \equiv 3 \equiv -1 \equiv -5 \pmod{4}$ because all these numbers have remainder 3 when divided by 4. For example, $-1 = -1 \cdot 4 + 3$ (recall that by definition the remainder is non-negative!).

4.3.3 Note. As in the previous example, we write “ \pmod{m} ” only once per line of congruences. Writing “ \pmod{m} ” simply means that one is working in a special setting where numbers are considered “congruent” if they have the same remainders when they are divided by m .

4.3.4 Note. Congruence modulo m divides \mathbb{Z} into m **congruence classes**, one for each possible remainder $r \in \{0, 1, 2, \dots, m-1\}$. The class corresponding to the remainder r , written as a set, is $\{r + mk \mid k \in \mathbb{Z}\}$, and we sometimes denote it by \bar{r} (“ r bar”). For example, if $m = 2$, the congruence classes are the even numbers ($\bar{0}$) and the odd numbers ($\bar{1}$).

4.3.5 Exercise. What are the congruence classes modulo 1?

4.3.6 Proposition. Let $a, b \in \mathbb{Z}$ and $m \in \mathbb{Z}_{\geq 1}$. Then the following are equivalent:

(a) $a \equiv b \pmod{m}$;

(b) $a = b + km$ for some $k \in \mathbb{Z}$;

(c) $m \mid a - b$.

Proof. It suffices to prove (a) \implies (b) \implies (c) \implies (a).

For (a) \implies (b), suppose $a \equiv b \pmod{m}$, namely that a and b have the same remainder when divided by m . Thus there are integers q_1 and q_2 such that $a = q_1m + r$ and $b = q_2m + r$, so $a - q_1m = r = b - q_2m$. Adding q_1m to both sides, we see that $a = b + (q_1 - q_2)m$, which proves (b).

For (b) \implies (c), suppose $a = b + km$ for some $k \in \mathbb{Z}$. Then $a - b = km$, so $m \mid a - b$.

For (c) \implies (a), suppose $m \mid a - b$, namely there is a $k \in \mathbb{Z}$ such that $a - b = mk$. Applying the division algorithm on a and m and again on b and m yields equations $a = q_1m + r_1$ and $b = q_2m + r_2$, and we will be done if we can show that $r_1 = r_2$. Note that

$$mk = a - b = (q_1 - q_2)m + (r_1 - r_2),$$

which implies $r_1 - r_2 = (m - q_1 + q_2)m$. Thus m divides $r_1 - r_2$. But r_1 and r_2 are both between 0 and $m - 1$, so $-(m - 1) \leq r_1 - r_2 \leq m - 1$. The only multiple of m in this interval is 0, so we must have $r_1 - r_2 = 0$, namely $r_1 = r_2$. \square

4.3.7 Note. The proof of the proposition is long, but don't worry: you only need to remember the statement of the proposition! Part (b) in particular is much more useful than our definition of $a \equiv b \pmod{m}$ in terms of remainders, so use it whenever you need to prove something about congruences. Another way to think about (b) is that adding a multiple of m to a number doesn't change its remainder when divided by m .

4.3.8 Example. Using part (b) of the proposition, we can quickly prove that $11 \equiv 7 \pmod{4}$, simply by noting that $11 = 7 - 4$.

4.3.9 Exercise. Let $a \in \mathbb{Z}$ and $d \in \mathbb{Z}_{\geq 1}$. Use Proposition 4.3.6 to show that $d \mid a$ if and only if $a \equiv 0 \pmod{d}$. Thus we can express any statement about divisibility in terms of congruences.

A useful result that makes computations modulo m easy is the following:

4.3.10 Proposition. Let $a, b \in \mathbb{Z}$ and $m \in \mathbb{Z}_{\geq 1}$. If $a \equiv b \pmod{m}$ and c is any integer, then

$$a + c \equiv b + c \pmod{m} \quad \text{and} \quad ac \equiv bc \pmod{m}.$$

Proof. Since $a \equiv b \pmod{m}$, there is an integer s such that $a = b + sm$. Then $a + c = b + c + sm$, so $a + c \equiv b + c \pmod{m}$, and $ac = (b + sm)c = bc + (sc)m$, so $ac \equiv bc \pmod{m}$. \square

4.3.11 Note. What the theorem says is this: if we are working modulo m , then we can replace any number in a sum or product by any other number with the same remainder (in the same congruence class). This makes computations easy, because we can replace big numbers by small numbers, such as those between 0 and m , or sometimes small negative numbers.

4.3.12 Example. To find the remainder of an expression a modulo m , you could evaluate a and then divide by m in the elementary school sense to find the remainder. But sometimes a is hard to evaluate because it may be hundreds of digits long! A better solution is to repeatedly use Proposition 4.3.10 to simplify the expression until you get a result between 0 and $m - 1$.

- Find the remainder of $3 \cdot 10^2 + 15$ modulo 7. In elementary school, you would compute $3 \cdot 10^2 + 15 = 315$, and then do long division to see that $315 = 45 \cdot 7$, namely the remainder is 0. Using congruences, we note that $10 \equiv 3$, $15 \equiv 1$, and $9 \equiv 2$ modulo 7, whence

$$3 \cdot 10^2 + 15 \equiv 3 \cdot 3^2 + 1 = 3 \cdot 9 + 1 \equiv 3 \cdot 2 + 1 = 6 + 1 \equiv 0 \pmod{7}.$$

Thus the remainder of $3 \cdot 10^2 + 15$ modulo 7 is 0.

- Find the remainder of $-2 \cdot 11 - 4$ modulo 8. Since $-2 \cdot 11 - 4 = -26$, long division yields $-26 = -4 \cdot 8 + 6$. Note that the remainder needs to be ≥ 0 . Using congruences:

$$-2 \cdot 11 - 4 \equiv -2 \cdot 3 - 4 = -10 \equiv 6 \pmod{8}.$$

- Find the remainder of 12^{100} modulo 11. Now the elementary school method is much less appealing, while congruences make life easy. Simply note that $12 \equiv 1 \pmod{11}$, so that

$$12^{100} \equiv 1^{100} = 1 \pmod{11},$$

so the remainder is 1.

- Find the remainder of 10^{2013} modulo 11. The easiest way to find the remainder is to see that $10 \equiv -1 \pmod{11}$. Then we compute

$$10^{2013} \equiv (-1)^{2013} = -1 \equiv 10 \pmod{11},$$

so the remainder is 10.

- Find the remainder of 3^{100} modulo 7. There's no trick quite as easy as in the previous examples, but the small modulus still makes things easy. Start by finding a power of 3 that is particularly easy to work with modulo 7: note that $3^3 = 27 \equiv -1 \pmod{7}$. Then compute

$$3^{100} = 3 \cdot (3^3)^{33} \equiv 3 \cdot (-1)^{33} = -3 \equiv 4 \pmod{7},$$

so the remainder is 4.

4.3.13 Note. In the computations of the previous example, we use the notation “ \equiv ” when we are using congruence and the usual “ $=$ ” when we are simplifying in the usual sense without using congruence. This makes the computation easier to follow because it pinpoints where congruence is being used. For example, we wrote $3 \cdot 2 + 1 = 6 + 1 \equiv 0 \pmod{7}$. Since two integers that are equal are *always* congruent for any modulus, it would be correct but less precise to write $3 \cdot 2 + 1 \equiv 6 + 1 \pmod{7}$.

4.3.14 Note. To summarize, arithmetic modulo m works just like normal arithmetic, except that you can replace numbers being added or multiplied (but not exponents!) by congruent numbers modulo m . This makes modular arithmetic much *easier*, because you can keep the numbers smaller than the modulus!

4.3.15 Exercise. Find the remainder of each of the following using congruences:

- (a) $3 \cdot 4 \cdot 5 + 11$ modulo 6;
- (b) $-2^3 \cdot 11 - 71$ modulo 7;
- (c) $5^{999} + 16$ modulo 4;
- (d) 3^{2200} modulo 8.

4.3.16 Exercise. Let $n \in \mathbb{Z}$. Prove or disprove each of the following:

- (a) $n^2 \equiv 0$ or $1 \pmod{4}$.
- (b) If n is odd, then $n^2 \equiv 1 \pmod{8}$.

Instead of resorting to tricks to compute large powers of integers modulo m , the following theorem takes care of most such problems when the modulus m is prime:

4.3.17 Theorem (Fermat’s little theorem, abbreviated FLT). *If p is prime and a is an integer not divisible by p , then*

$$a^{p-1} \equiv 1 \pmod{p}.$$

The theorem is saying that for any a not divisible by p , p divides $a^{p-1} - 1$. For instance, if $3 \nmid a$, then $3 \mid (a^2 - 1)$. Indeed, 3 divides $1^2 - 1 = 0$, $2^2 - 1 = 3$, $4^2 - 1 = 15$, $5^2 - 1 = 24$, and so on.

Outline of proof.

- Step 1: Show that if a is not divisible by p , then no two of the integers $1 \cdot a, 2 \cdot a, \dots, (p-1) \cdot a$ are congruent modulo p .
- Step 2: Let n be the product $1 \cdot 2 \cdots (p-1)$. Use Step 1 to show that $n \equiv a^{p-1}n \pmod{p}$.
- Step 3: Show that $\gcd(n, p) = 1$. Therefore n has a multiplicative inverse modulo p by Proposition 4.4.3.

Step 4: Use Steps 2 and 3 to deduce that $1 \equiv a^{p-1} \pmod{p}$.

□

4.3.18 Exercise. Fill in the details in the proof of Fermat's little theorem 4.3.17.

4.3.19 Example. Find the remainder of 3^{100} modulo 7. Since $\gcd(3, 7) = 1$, we can apply Fermat's little theorem to compute

$$3^{100} = 3^4 \cdot (3^6)^{16} \equiv 3^4 \cdot (1)^{16} = 3^4 = 9^2 \equiv 2^2 \equiv 4 \pmod{7},$$

which is the same answer we got earlier!

4.3.20 Exercise. Use Fermat's little theorem to find the remainder of:

(a) 7^{121} modulo 13;

(b) 22^{100} modulo 11;

(c) 24^{39} modulo 7.

Congruences can be used to give easy proofs of criteria for divisibility.

4.3.21 Claim. Let $a \in \mathbb{Z}$ and let D be the sum of the digits of a . Then $3 \mid a$ if and only if $3 \mid D$.

Proof. Write $a = a_n a_{n-1} \dots a_0$, where the a_i denote the digits of a . Then since $10 \equiv 1 \pmod{3}$,

$$a = a_0 + 10a_1 + 10^2a_2 + \dots + 10^na_n \equiv a_0 + a_1 + \dots + a_n = D \pmod{3}.$$

Thus $a \equiv 0 \pmod{3}$ if and only if $D \equiv 0 \pmod{3}$.

□

4.3.22 Exercise. Let $a \in \mathbb{Z}_{\geq 1}$ and let D denote the sum of the digits of a . Prove or disprove the following claims:

(a) $5 \mid a$ if and only if $5 \mid D$.

(b) $9 \mid a$ if and only if $9 \mid D$.

4.3.23 Exercise. Let $a \in \mathbb{Z}_{\geq 1}$. Come up with a criterion involving the digits of a for when a is divisible by 5 and a criterion for when a is divisible by 11. Prove your criteria!

4.4 Linear Congruences

One of the things you learn in your first algebra class is how to solve linear equations like $3x + 4 = 0$ for x . One first subtracts 4 (adds -4), to obtain $3x = -4$. Then one divides by 3, which really means multiplying by $\frac{1}{3}$. The key is that the coefficient of x , namely 3, has a multiplicative inverse, namely $\frac{1}{3}$, which is a number such that $\frac{1}{3} \cdot 3 = 1$. The multiplicative inverse allows you change the coefficient of x to 1.

A natural question is whether we can solve the congruence analog of linear equations.

4.4.1 Definition. A **linear congruence** is a congruence of the form $ax + b \equiv 0 \pmod{m}$, where $a, b \in \mathbb{Z}$, $m \in \mathbb{Z}_{\geq 1}$, and x is a variable.

We want to solve for all integer values of x that satisfy the congruence.

4.4.2 Example. Solve the linear congruence $3x + 1 \equiv 0 \pmod{5}$. First, we add -1 (which is congruent to 4) to both sides, to get $3x \equiv 4 \pmod{5}$. Now we want to remove the coefficient of x . For this, we need to find a multiplicative inverse of 3 modulo 5, namely some $c \in \mathbb{Z}$ such that $c \cdot 3 \equiv 1 \pmod{5}$. Guess and check reveals that 2 works: $2 \cdot 3 \equiv 1 \pmod{5}$. So we multiply our equation by 2 on both sides, to get $2 \cdot 3x \equiv 2 \cdot 4 \pmod{5}$, which simplifies to $x \equiv 3 \pmod{5}$. Now we can read off the integer solutions to our linear congruence: x can be anything with remainder 3 modulo 5, namely any integer of the form $3 + 5k$ for $k \in \mathbb{Z}$. So we can write the integer solution set as $\{3 + 5k \mid k \in \mathbb{Z}\}$. Instead of just one value for x , our solution is a whole congruence class modulo 5.

The only tricky part of solving the congruence in the example was the existence of a multiplicative inverse for 3. The natural question to ask is: When does a have a multiplicative inverse modulo m , and how can we find it if it exists?

Here is the answer:

4.4.3 Proposition. Let $a \in \mathbb{Z}$ and $m \in \mathbb{Z}_{\geq 2}$. If $\gcd(a, m) = 1$, then a has an inverse modulo m .

We've done all the hard work already by studying the Euclidean algorithm and back substitution, leading up to Bézout's identity. Bézout's identity will give the multiplicative inverse. It is time to reap the rewards!

Proof. Since $\gcd(a, m) = 1$, by Bézout's identity there exist $s, t \in \mathbb{Z}$ such that $1 = sa + tm$. Thus $sa \equiv 1 \pmod{m}$, so s is a multiplicative inverse of a . \square

So we can solve any linear congruence $ax + b \equiv 0 \pmod{m}$ for which $\gcd(a, m) = 1$. For instance:

4.4.4 Example. Solve $3x + 6 \equiv 0 \pmod{10}$. We run the Euclidean algorithm on 3 and 10:

$$\begin{aligned} 10 &= 3 \cdot 3 + \boxed{1} \\ 3 &= 3 \cdot 1 + 0. \end{aligned}$$

So $\gcd(3, 10) = 1$. Moreover, $1 = 10 - 3 \cdot 3$, so the multiplicative inverse of 3 modulo 10 is -3 , which is congruent to 7. So we add 4 to both sides of the original congruence to get $3x \equiv 4 \pmod{10}$, then multiply by 7 to get

$$x \equiv 7 \cdot 4 \equiv 8 \pmod{10},$$

so the set of integer solutions is $\{8 + 10k \mid k \in \mathbb{Z}\}$.

But what if $\gcd(a, m) = d > 1$? For instance:

4.4.5 Example. Solve $2x \equiv 1 \pmod{6}$. Here $\gcd(2, 6) = 2$, so we cannot find an inverse for 2 modulo 6. Test some numbers for x , and you'll find none of them work! That's because if x were a solution, then $2x = 1 + 6k$ for some k , but $2 \mid 2x$ and $2 \mid 6k$, so necessarily $2 \mid (2x - 6k)$, which is a contradiction since $2 \nmid 1$. So there are no solutions!

The problem in the above example was that $\gcd(a, m) \nmid b$, which makes it impossible to find an integer solution. So what happens if $\gcd(a, m) \mid b$?

4.4.6 Example. Solve $4x + 2 \equiv 0 \pmod{6}$. As before, $\gcd(4, 6) = 2$, and this time $b = 2$ and $2 \mid 2$. x is a solution if $4x + 2 = 6k$, which implies that $2x + 1 = 3k$. So we see that the integer solutions to $4x + 2 \equiv 0 \pmod{6}$ are the same as the integer solutions to $2x + 1 \equiv 0 \pmod{3}$, which we can solve since $\gcd(2, 3) = 1$.

We summarize all of this:

4.4.7 Theorem. Let $a, b \in \mathbb{Z}$ and $m \in \mathbb{Z}_{\geq 2}$ and set $d = \gcd(a, m)$. Then the linear congruence $ax + b \equiv 0 \pmod{m}$ has integer solutions if and only if $d \mid b$, in which case the integer solutions are the same as the integer solutions of the congruence $\frac{a}{d}x + \frac{b}{d} \equiv 0 \pmod{\frac{m}{d}}$. Since $\gcd(\frac{a}{d}, \frac{m}{d}) = 1$, this latter system can be solved by finding a multiplicative inverse for $\frac{a}{d}$ modulo $\frac{m}{d}$.

4.4.8 Exercise. Solve the following linear congruences. Feel free to use your results from Exercise 4.2.17 as long as you state the results you are using. Show your work!

- (a) $10x + 3 \equiv 0 \pmod{11}$.
- (b) $9x \equiv 0 \pmod{11}$.
- (c) $21x + 40 \equiv 0 \pmod{44}$.
- (d) $33x + 1 \equiv 0 \pmod{44}$.
- (e) $36x + 24 \equiv 0 \pmod{48}$.

4.4.9 Exercise. Give an example for each of the following or argue that no example exists:

- A linear congruence modulo 5 with no solutions.
- A linear congruence modulo 6 with no solutions.
- A linear congruence modulo 6 with solution set $\{1 + 3k \mid k \in \mathbb{Z}\}$.

4.5 Cryptography

Have you ever solved a cryptogram? You see an encrypted phrase like

XBZZA, LBALZB!

which is the result of taking a phrase and performing a letter substitution, such as replacing A by T , B by L , and so on. In this case, analyzing the punctuation and letter patterns could lead you to the solution

HELLO, PEOPLE!

To make a simple cryptogram, choose a short message, let's call it \mathcal{M} , written in capital letters with spaces but no punctuation, and a bijective (one-to-one) function $\{A, B, C, \dots, Z\} \xrightarrow{\phi_E} \{A, B, C, \dots, Z\}$. Then “encrypt” the message \mathcal{M} by applying the function to each letter in \mathcal{M} , to obtain an encrypted message $\phi_E(\mathcal{M})$. To “decrypt” an encrypted message, simply apply the inverse of ϕ_E : $\phi_E^{-1}(\phi_E(\mathcal{M})) = \mathcal{M}$. We call \mathcal{M} the **message**, ϕ_E the **encryption key**, and ϕ_E^{-1} the **decryption key**.

Suppose you want to communicate with a friend via written messages, in such a way that no third party who intercepts the messages will be able to read them. You and your friend get together in utmost secrecy and agree on an encryption key ϕ_E and a decryption key ϕ_D , which could be much more complicated than the bijection above (for instance, choose one bijection for the first word of the message, a different bijection for the second word, etc.). Now as long as you keep these keys secret, third parties who intercept encrypted messages $\phi_E(\mathcal{M})$ will have trouble decrypting them.

But what if you have no way to secretly share the encryption and decryption keys?

Public key cryptography is a way for strangers to communicate securely. The future message recipient, say Amazon.com, publishes an encryption key ϕ_E , which anyone can see, but keeps the decryption key ϕ_D private. When making a purchase, a shopper uses the encryption key to encrypt the details \mathcal{M} of the order (especially the credit card number!), and the encrypted message $\phi_E(\mathcal{M})$ is sent to Amazon. Amazon then decrypts the message $\phi_D(\phi_E(\mathcal{M})) = \mathcal{M}$ and draws money from the account. Since the encryption key is public, any snooping third party has access to it. So security hinges on whether decryption is a much harder process than encryption: the encryption key should not reveal the decryption key! Note that this is not the case in the cryptogram example above, where $\phi_D = \phi_E^{-1}$, namely the decryption key was easy to compute from the encryption key.

A famous public key cryptosystem used extensively online is called the **RSA cryptosystem**. RSA is based on number theory, so the first step is to convert letters into numbers. We can do this by the easy substitution $A \mapsto 01$, $B \mapsto 02$, $C \mapsto 03$, ..., $Z \mapsto 26$. We also want to keep track of spaces, so we replace each space by “00”.

We'll illustrate how RSA encryption and decryption work through a simple example.

4.5.1 Example (RSA encryption). Amazon publishes two carefully chosen numbers, say $n = 28907$ and $e = 11$ (in reality these will be *much* larger), which will be used for

encryption. For convenience, set ℓ to be the number of digits of n , in this case $\ell = 5$. You, the shopper, can see all of this public information, and you encode your order as follows. Suppose part of the order is

DISCRETE MATH.

First, you convert this text into numbers, by the above substitution, to get

04091903180520050013012008.

Next, you chop this number into blocks of length $\ell - 1 = 4$, namely

$$m_1 = 0409, m_2 = 1903, m_3 = 1805, m_4 = 2005, m_5 = 0013, m_6 = 0120, m_7 = 0800$$

(pad m_7 with extra zeros to ensure all blocks are the same length). Now you compute the remainder r_i of each m_i^e modulo n , write each r_i as a 5-digit number (padding the front with 0's if necessary), and concatenate everything into a long string, which gets sent to Amazon. The computations are

$$\begin{aligned} 0409^{11} &\equiv 20557 \pmod{28907}, \\ 1903^{11} &\equiv 21779 \pmod{28907}, \\ 1805^{11} &\equiv 06299 \pmod{28907}, \\ 2005^{11} &\equiv 04448 \pmod{28907}, \\ 0013^{11} &\equiv 20166 \pmod{28907}, \\ 0120^{11} &\equiv 11232 \pmod{28907}, \\ 0800^{11} &\equiv 11558 \pmod{28907}, \end{aligned}$$

so the remainders are

$$r_1 = 20557, r_2 = 21779, r_3 = 06299, r_4 = 04448, r_5 = 20166, r_6 = 11232, r_7 = 11558.$$

The string sent to Amazon is

20557217790629904448201661123211558.

4.5.2 Algorithm (RSA encryption). Summary of the encryption process ϕ_E of a string \mathcal{M} , using n and e , where ℓ is the number of digits of n :

- **Step 1:** Convert \mathcal{M} into a numbers in the usual way.
- **Step 2:** Break the result into blocks m_i of length $\ell - 1$ (pad the end of the last block with 0's).
- **Step 3:** Compute the remainder r_i of each m_i^e modulo n .
- **Step 4:** Consider each r_i as having length ℓ (pad the front of each r_i with 0's).

- **Step 5:** Concatenate the r_i to obtain $\phi_E(\mathcal{M})$.

Any third party who knows n and e and intercepts $\phi_E(\mathcal{M})$ will trouble decrypting the message because modular exponentiation (Step 3) is not easy to invert in general. But the situation is not general, and only Amazon knows why!

4.5.3 Example (RSA decryption). Only Amazon knows the prime factorization of n , which is $28907 = 137 \cdot 211$. Amazon chose $e = 11$ to be relatively prime to $(137 - 1) \cdot (211 - 1) = 28560$. Thus Amazon can use the Euclidean algorithm with substitution to find a linear combination of 11 and 28560 that equals 1:

$$28560 = 2596 \cdot 11 + 4$$

$$11 = 2 \cdot 4 + 3$$

$$4 = 1 \cdot 3 + \boxed{1}$$

$$3 = 3 \cdot 1 + 0,$$

so that

$$1 = 4 - 3 = 4 - (11 - 2 \cdot 4) = -11 + 3 \cdot 4 = -11 + 3 \cdot (28560 - 2596 \cdot 11) = \boxed{3} \cdot 28560 - \boxed{7789} \cdot 11.$$

Thus -7789 is a multiplicative inverse of 11 modulo 28560, and so is $-7789 + 28560 = 20771$. Set $d = 20771$; this is the magic number that will allow Amazon to decrypt the message!

Now Amazon breaks $\phi_E(\mathcal{M})$ into blocks r_i of length ℓ , and simply computes the remainders s_i of each m_i^d modulo n , considering each as an $(\ell - 1)$ -digit number. Thus to decrypt the string Amazon received in the previous example, Amazon computes

$$20557^{20771} \equiv 0409 \pmod{28907},$$

$$21779^{20771} \equiv 1903 \pmod{28907},$$

$$06299^{20771} \equiv 1805 \pmod{28907},$$

$$04448^{20771} \equiv 2005 \pmod{28907},$$

$$20166^{20771} \equiv 0013 \pmod{28907},$$

$$11232^{20771} \equiv 0120 \pmod{28907},$$

$$11558^{20771} \equiv 0800 \pmod{28907},$$

namely

$$s_1 = 0409, s_2 = 1903, s_3 = 1805, s_4 = 2005, s_5 = 0013, s_6 = 0120, s_7 = 0800.$$

Finally, Amazon concatenates all the s_i to obtain

$$0409190318052005001301200800,$$

and then converts this back into letters by the usual substitution, ending with the original message:

DISCRETE MATH .

4.5.4 Algorithm (RSA decryption). Summary of the decryption process ϕ_D of an encrypted string $\phi_E(\mathcal{M})$ using d , a multiplicative inverse of e modulo $(p_1 - 1)(p_2 - 1)$, where $n = p_1 p_2$ is the prime factorization of n and ℓ is the number of digits of n :

- **Step 1:** Break $\phi_E(\mathcal{M})$ into blocks r_i of length ℓ .
- **Step 2:** Compute the remainder s_i of each r_i^d modulo n .
- **Step 3:** Think of s_i as an $(\ell - 1)$ -digit number (in fact, $s_i = m_i!$)
- **Step 4:** Convert the m_i into letters in the usual way.
- **Step 5:** Concatenate the results to obtain $\phi_D(\phi_E(\mathcal{M}))$.

As in the example, $\phi_D(\phi_E(\mathcal{M})) = \mathcal{M}$, as we will prove shortly!

4.5.5 Note (RSA is secure). In practice, Amazon chooses two extremely large distinct prime numbers p_1 and p_2 (at least 200 digits each) and picks e to be some positive integer relatively prime to $(p_1 - 1)(p_2 - 1)$. Amazon publishes $n = p_1 p_2$ and e , and keeps the two prime factors of n secret. The number n is so large that there is no hope of finding its prime factors! Decryption relies on computing d , which cannot be done without knowing the prime factors, so a snooping third party that knows only n , e , and an encrypted message can't decrypt the message.

Why does RSA decryption recover the original message? The key step in encryption is to encrypt a block m by finding the remainder r of m^e modulo n , and the key step in decryption is to compute the remainder s of r^d modulo n . Thus

$$s \equiv r^d \equiv (m^e)^d \equiv m^{ed} \pmod{n},$$

so RSA will recover the original block if $s \equiv m \pmod{n}$. Thus we want to prove the theorem:

4.5.6 Theorem (RSA decryption works). *Let p_1, p_2 be distinct primes, let $e \in \mathbb{Z}_{\geq 1}$ satisfy $\gcd(e, (p_1 - 1)(p_2 - 1)) = 1$, and let $d \in \mathbb{Z}_{\geq 1}$ be a multiplicative inverse of e modulo $(p_1 - 1)(p_2 - 1)$. Then*

$$m^{ed} \equiv m \pmod{p_1 p_2}$$

for any $m \in \mathbb{Z}$.

The challenge in proving the theorem is that the modulus $p_1 p_2$ is a product of primes, whereas we are best at working modulo a prime, when we can use powerful tools like FLT. We want to prove that m^{ed} is a solution of the congruence $x \equiv m \pmod{p_1 p_2}$, and we do this as follows.

Proof. We break the proof into three steps.

Step 1: Split the congruence $x \equiv m \pmod{p_1 p_2}$ into the system of two congruences

$$\begin{aligned} x &\equiv m \pmod{p_1} \\ x &\equiv m \pmod{p_2}. \end{aligned}$$

We say an integer is a solution of the system if it satisfies both congruences.

Step 2: Show that any solution of the system must also be a solution of the original congruence.

Step 3: Show that m^{ed} is a solution of the system.

Step 1 does not require any work, so we begin by showing Step 2. Suppose a is a solution of the system. Then any $a + kp_1 p_2$ is also a solution of the system, since the term $k p_1 p_2$ is congruent to 0 modulo p_1 and modulo p_2 . Thus the remainder r of a modulo $p_1 p_2$ is a solution of the system. Note that $0 \leq r < p_1 p_2$. The system also has the obvious solution m , and by the same argument the remainder s of m modulo $p_1 p_2$ is also a solution of the system, and $0 \leq s < p_1 p_2$. But by the Chinese Remainder Theorem (abbreviated CRT), which I will state in a moment, the system has a unique solution in the interval $0 \leq x < p_1 p_2$, thus we must have $r = s$. But this implies

$$a \equiv r = s \equiv m \pmod{p_1 p_2},$$

so a satisfies the original congruence, as claimed.

Here is the statement of the mighty CRT, which I will prove after completing Step 3.

4.5.7 Theorem (Chinese Remainder Theorem, abbreviated CRT). *Let p_1, p_2 be distinct primes, and let $a_1, a_2 \in \mathbb{Z}$. Then the system of congruences*

$$\begin{aligned} x &\equiv a_1 \pmod{p_1}, \\ x &\equiv a_2 \pmod{p_2} \end{aligned}$$

has a unique solution in the interval $0 \leq x < p_1 p_2$.

To finish the proof that RSA decryption works, we now complete Step 3, namely we must show that $m^{ed} \equiv m \pmod{p_1}$ and $m^{ed} \equiv m \pmod{p_2}$. The argument is the same for each congruence up to interchanging the roles of p_1 and p_2 , so we will only prove that $m^{ed} \equiv m \pmod{p_1}$. An easy case is when $p_1 \mid m$, in which case $m \equiv 0 \pmod{p_1}$, so $m^{ed} \equiv 0 \equiv m \pmod{p_1}$ is what we are trying to show. So assume $p_1 \nmid m$, which will allow us to use FLT. We will now finally use the properties of e and d . Since d is a multiplicative inverse of e modulo $(p_1 - 1)(p_2 - 1)$, $ed \equiv 1 \pmod{(p_1 - 1)(p_2 - 1)}$, namely $de = 1 + k(p_1 - 1)(p_2 - 1)$ for some $k \in \mathbb{Z}$. Thus by FLT,

$$m^{ed} = m^{1+k(p_1-1)(p_2-1)} = m \cdot (m^{p_1-1})^{k(p_2-1)} \equiv m \cdot 1^{k(p_2-1)} = m \pmod{p_1},$$

so we are done. □

To prove the CRT, stated above, we need an easy lemma. Remember Exercise 4.1.16(e)? You probably proved (e) using prime factorizations. Here's a different proof.

4.5.8 Lemma. *Let $a, b, p \in \mathbb{Z}$ with p prime. If $p \mid ab$, then $p \mid a$ or $p \mid b$.*

Proof. Suppose $p \nmid a$. Then $\gcd(p, a) = 1$, so by Bézout's identity (4.2.15), there are $r, s \in \mathbb{Z}$ such that $rp + sa = 1$. Multiplying both sides by b , we get $rpb + sab = b$. Since $p \mid ab$ by assumption, $p \mid sab$, and also $p \mid rpb$, thus $p \mid (rpb + sab) = b$, which completes the proof. \square

Proof of the CRT. Since p_1 and p_2 are distinct primes, $\gcd(p_1, p_2) = 1$, so by Bézout's identity (4.2.15) there exist $r_1, r_2 \in \mathbb{Z}$ such that $r_1p_1 + r_2p_2 = 1$. Then $x = a_1r_2p_2 + a_2r_1p_1$ is a solution to the system of congruences, and finding the remainder modulo p_1p_2 gives a solution in the interval $0 \leq x < p_1p_2$. To see that the solution is unique, suppose that $0 \leq x, y < p_1p_2$ are two solutions. Then $x \equiv a_1 \equiv y \pmod{p_1}$ and $x \equiv a_2 \equiv y \pmod{p_2}$, so there are $k_1, k_2 \in \mathbb{Z}$ such that $x = y + k_1p_1$ and $x = y + k_2p_2$. Thus $k_1p_1 = k_2p_2$, so $p_1 \mid k_2p_2$. Since $p_1 \nmid p_2$, the lemma implies $p_1 \mid k_2$, namely there is some $l \in \mathbb{Z}$ such that $k_2 = lp_1$. Thus $x = y + lp_1p_2$, so $x \equiv y \pmod{p_1p_2}$, which implies $x = y$ since $0 \leq x, y < p_1p_2$. \square

Chapter 5

Induction

5.1 Proof by Induction

Have you ever set up a domino rally? You place dominoes upright in a line, then knock over the first domino, which knocks over the second, which knocks over the third, and so on. Keep this image in mind while learning about mathematical induction, because the basic idea of induction is that of the domino rally.

Mathematical induction is a powerful technique for proving a propositional function is true for all positive (or nonnegative) integers. Here is a typical proof by induction:

5.1.1 Claim. *Let $n \in \mathbb{Z}_{\geq 1}$. Then*

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

Proof. Induction on n . For the base case $n = 1$, note that $1 = \frac{1(1+1)}{2}$. Now suppose the claim is true for some $n = k \in \mathbb{Z}_{\geq 1}$, namely that

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

Then we compute

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

so the claim is true for $n = k + 1$. Thus we are done by induction. \square

Why did the previous argument prove the claim? Consider the propositional function “the sum of the first n positive integers is $\frac{n(n+1)}{2}$ ”, which we’ll denote $P(n)$. The claim is stating that $P(n)$ is true for all $n \in \mathbb{Z}_{\geq 1}$. To prove this, we argue as follows:

- (1) “Base case”: Prove the claim for the lowest value of n ; in our case, show $P(1)$.
(*Knock over the first domino.*)

- (2) “Inductive step”: Prove that $P(k)$ implies $P(k + 1)$ for arbitrary $k \in \mathbb{Z}_{\geq 1}$. (*Show that a falling domino knocks over the next domino.*)

These two steps prove the theorem because (1) establishes $P(1)$, while (2) establishes the chain of implications

$$P(1) \implies P(2) \implies P(3) \implies P(4) \implies \dots$$

Since $P(1)$ is true, so is $P(2)$. Since $P(2)$ is true, so is $P(3)$. And so on. Each falling domino knocks over the next domino in line, so all of the infinitely many dominoes get knocked over.

5.1.2 Note. The power of induction is that for the inductive step, we get to assume a case $P(k)$, and use that to prove $P(k + 1)$. For instance, our assumption that the claim is true for $n = k$ gave us a formula for adding up all the terms of $1 + 2 + \dots + k + (k + 1)$ except the last one. This is much easier than directly proving $P(k + 1)$ without any assumptions.

5.1.3 Note. Note that we do not use the $P(n)$ notation when writing a proof! Instead of writing $P(k)$, for instance, we refer to the theorem being true for $n = k$, as in the proof of Claim 5.1.1.

5.1.4 Note. Induction can be confusing at first, but it is incredibly important in mathematics. Make sure you understand the logic of induction and how to write a proof by induction.

Let’s try another proof by induction.

5.1.5 Claim. *Let $n \in \mathbb{Z}_{\geq 1}$. Then the sum of the first n odd positive integers is n^2 , namely*

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

Proof. Induction on n . For the base case $n = 1$, we simply note that $1 = 1^2$. Now suppose the theorem is true for some $n = k \in \mathbb{Z}_{\geq 1}$, namely that

$$1 + 3 + \dots + (2k - 3) + (2k - 1) = k^2.$$

We can use this to compute

$$1 + 3 + \dots + (2k - 3) + (2k - 1) + (2k + 1) = k^2 + (2k + 1) = (k + 1)^2,$$

which shows that the theorem is true for $n = k + 1$. Thus we are done by induction. \square

Here’s a proof by induction that does not involve a summation formula.

5.1.6 Claim. *Let $n \in \mathbb{Z}_{\geq 1}$. Then $n^3 - n \equiv 0 \pmod{3}$.*

Proof. Induction on n . The base case $n = 1$ holds since $1^3 - 1 = 0 \equiv 0 \pmod{3}$. Now suppose that the claim holds for some $n = k \in \mathbb{Z}_{\geq 1}$, namely that $k^3 - k \equiv 0 \pmod{3}$. Then

$$(k+1)^3 - (k+1) = k^3 + 3k^2 + 3k + 1 - (k+1) = (k^3 - k) + 3k(k+1) \equiv 0 + 0 \pmod{3},$$

which proves the claim when $n = k$. Thus we are done by induction. \square

Here's a less abstract proof by induction.

5.1.7 Claim. *In any non-empty set S of horses, all the horses are the same color.*

Proof. Proof by induction on n , the number of horses in S . The base case $n = 1$ holds since there is only one horse in the set. Now suppose that for some $k \in \mathbb{Z}_{\geq 1}$, any set of k horses has all horses the same color. Given a set S of $k+1$ horses, pick two different subsets A_1, A_2 of S with k horses each. By assumption, all the horses of A_1 are the same color and all the horses of A_2 are the same color. But the intersection $A \cap B$ contains $k-1$ horses, hence the color of the horses in A is the same as the color of the horses in B . By induction, we have proved the theorem. \square

Wait, what did we just prove?! Analyze the proof carefully and try to find an error. Hint: there is a gap in the proof that $P(k) \implies P(k+1)$ when $k = 1$. The above theorem is a *flawed* proof by induction: don't write proofs like this!

Thus the base case holds, but the rest of the proof falls apart! The first domino gets knocked over, but the falling first domino fails to knock over the second domino, so no other dominoes fall. Note that a small mistake in one case of this proof by induction is the difference between the claim being true for all $n \geq 2$ and being false for all $n \geq 2$. So we must be careful when trying to prove a claim by induction!

5.1.8 Exercise. Let $n \in \mathbb{Z}_{\geq 1}$. Prove each of the following claims by induction or disprove by giving a counterexample.

- (a) The sum of the first n positive integers is $\frac{n(n+1)}{2}$, namely $1 + 2 + \cdots + n = \frac{n(n+1)}{2}$.
- (b) The sum of the first n even positive integers is $n(n+1)$, namely $2 + 4 + \cdots + 2n = n(n+1)$.
- (c) The sum of the first n odd positive integers is n^2 , namely $1 + 3 + 5 + \cdots + (2n-1) = n^2$.
- (d) $1 \cdot 1! + 2 \cdot 2! + \cdots + n \cdot n! = (n+1)! - 1$.
- (e) $1^2 - 2^2 + 3^2 - \cdots + (-1)^{n-1}n^2 = (-1)^{n-1}\frac{n(n+1)}{2}$.
- (f) $1 \cdot 2 + 2 \cdot 3 + \cdots + n(n+1) = \frac{n(n+1)(n+2)}{3}$.
- (g) $1^3 + 2^3 + \cdots + n^3 = \frac{n^2(n+1)^2}{4}$.
- (h) $1^2 + 2^2 + \cdots + n^2 = \frac{n(n+1)(2n+1)}{6}$.

5.1.9 Exercise. Let $n \in \mathbb{Z}_{\geq 1}$. Prove each of the following claims by induction or disprove by giving a counterexample.

- (a) $n^2 + n \equiv 0 \pmod{2}$.
- (b) $n^3 + 2n \equiv 0 \pmod{3}$.
- (c) $n^2 - n \equiv 0 \pmod{3}$.
- (d) $n^3 - n \equiv 0 \pmod{6}$.
- (e) Every n -element set A has exactly $\frac{n(n-1)}{2}$ 2-element subsets.

5.1.10 Exercise. Let $n \in \mathbb{Z}_{\geq 1}$. Prove that every n -element set A has exactly $\frac{n(n-1)}{2}$ 2-element subsets.

5.1.11 Exercise. Prove the “power rule” for derivatives: Let $n \in \mathbb{Z}_{\geq 1}$ and set $f_n(x) = x^n$. Then $f'_n(x) = nx^{n-1}$.

(Hint: Induction on n . For the base case, use the definition of the derivative. For the inductive step, use the product rule for derivatives.)

5.1.12 Exercise. Prove the following claim: Let $n \in \mathbb{Z}_{\geq 2}$, let $a_1, \dots, a_n \in \mathbb{Z}$, and let p be a prime. If $p \mid a_1 a_2 \cdots a_n$ where each $a_i \in \mathbb{Z}$, then $p \mid a_i$ for some i .

(Hint: Induction on n . Use Lemma 4.5.8 for the base case and for the inductive step.)

5.2 Strong Induction and the FTA

Our goal in this section is to prove the FTA (4.1.9):

Theorem (Fundamental theorem of arithmetic). *Let $n \in \mathbb{Z}_{\geq 2}$. Then*

- (a) *n can be written as a product of primes $n = p_1 p_2 \cdots p_r$;*
- (b) *this prime factorization is unique if we insist that $p_1 \leq p_2 \leq \cdots \leq p_r$.*

For the proof of (a), we will use a slightly different version of induction, called **strong induction**, which is logically equivalent to induction, but sometimes makes a proof easier. The two steps for using strong induction are:

- (1) Prove the “base case” $P(1)$. (*Knock over the first domino.*)
- (2) For arbitrary $k \in \mathbb{Z}_{\geq 1}$, prove that $P(i)$ for all $i \leq k$ implies $P(k+1)$. (*Show that if all the dominoes up to a certain point have fallen over, then the next one gets knocked over.*)

Here is how we can use strong induction to prove (a):

Proof of (a). Strong induction on n . For the base case $n = 2$, note that 2 is prime, hence factors uniquely into primes as $2 = 2$. Now assume that for some $k \in \mathbb{Z}_{\geq 2}$, any integer i in the range $2 \leq i \leq k$ can be written as a product of primes. Then either $k + 1$ is prime, in which case $k + 1 = k + 1$ is a prime factorization, or $k + 1$ is composite, in which case it has a divisor d in the interval $2 \leq d \leq k$. Since $d \mid k + 1$, there is an $e \in \mathbb{Z}$ such that $k + 1 = de$, and the bounds on d imply that $2 \leq e \leq k$. Thus by assumption, $d = p_1 \cdots p_s$ and $e = p_{s+1} \cdots p_r$, so combining these factorizations yields $k + 1 = de = p_1 \cdots p_s p_{s+1} \cdots p_r$. By strong induction, I have proved that every $n \in \mathbb{Z}_{\geq 2}$ has a prime factorization. \square

Proof of (b). To see that a prime factorization $n = p_1 p_2 \cdots p_r$ is unique when we insist that $p_1 \leq p_2 \leq \cdots \leq p_r$, suppose for contradiction that there is a different prime factorization $n = q_1 q_2 \cdots q_s$ with $q_1 \leq q_2 \leq \cdots \leq q_s$. Removing prime factors that appear in both factorizations, the leftover primes give an equation

$$p_{i_1} p_{i_2} \cdots p_{i_u} = q_{j_1} q_{j_2} \cdots q_{j_v},$$

where no prime appears on both sides. But p_{i_1} divides the left side, hence also the right side, so by Exercise 5.1.12, $p_{i_1} \mid q_{j_l}$ for some l , hence $p_{i_1} = q_{j_l}$ since q_{j_l} is prime, which is a contradiction. \square

Chapter 6

Combinatorics

6.1 Counting

Let's start with a very easy counting problem.

6.1.1 Example. Suppose you have 7 candy bars and 11 lollipops. In how many ways can you choose one piece of candy?

Solution: You can choose either a candy bar or a lollipop. Since there are 7 candy bars and 11 lollipops, you have $7 + 11 = 18$ different choices of a piece of candy.

The simple principle at work here is the following:

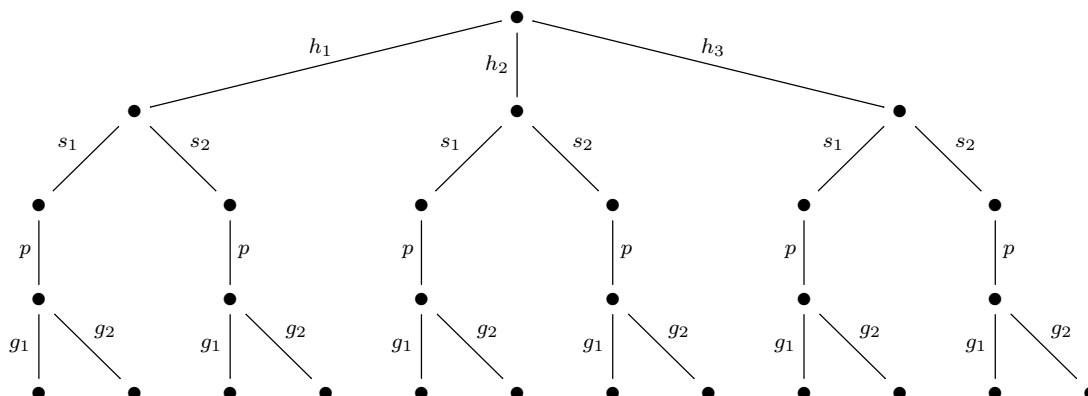
6.1.2 Proposition (Sum rule). *If a task can be done either in one of n_1 ways or in one of n_2 ways, where none of the set of n_1 ways is the same as any of the set of n_2 ways, then there are $n_1 + n_2$ ways to do the task.*

Here's a trickier kind of problem.

6.1.3 Example. Now suppose you have 3 pointy hats, 2 bloody shirts, 1 coal-black pair of pants, and 2 ghoulish pairs of socks. How many different costumes of 1 hat, 1 shirt, 1 pair of pants, and 1 pair of socks can you make?

Solution: Draw a tree diagram. Starting with a point (the root of the tree), draw 3 branches, one for each choice of hat. From the end of each hat branch, draw 2 branches, one for each type of shirt. From the end of each shirt branch, draw 1 branch, since there is only one choice of pants. Now for each pants branch, draw 2 branches, one for each pair of socks. Each path from the root of the tree to the end of a socks branch corresponds to a costume. Counting the branches, there are a total of $3 \cdot 2 \cdot 1 \cdot 2 = 12$ different costumes. Your diagram might look like this, where $\{h_1, h_2, h_3\}$ is the set of hats, $\{s_1, s_2\}$ is the set

of shirts, $\{p\}$ is the set of pants, and $\{g_1, g_2\}$ is the set of pairs of socks:



We can formalize this method of drawing a tree diagram to obtain:

6.1.4 Proposition (Product rule). *Suppose a procedure consists of independent tasks T_1, \dots, T_r . Suppose there are n_i ways of doing the task T_i for each $1 \leq i \leq r$. Then there are $n_1 n_2 \cdots n_r$ ways to do the procedure.*

6.1.5 Note. The word “independent” means that the number of ways of doing task T_i does not depend on which way the other tasks are being done. For instance, in the example, the number of ways to pick a shirt did not depend on the hat that was chosen.

6.1.6 Example. How many different 6-letter “words” are possible? (By “word” I mean any string of 6 letters.)

Solution: Think of choosing a word as a procedure with 6 tasks T_1, T_2, \dots, T_6 , where each T_i is the task of choosing the i th letter of the word. Since there are 26 letters, there are 26 ways to do each T_i , thus by the product rule there are 26^6 6-letter words. That’s over 300 million possibilities! Of course most of these, like “orltkq”, are not (yet!) words in the English language.

6.1.7 Example. How many functions are there from a set $A = \{a_1, a_2, a_3\}$ with three elements to a set $B = \{b_1, b_2, b_3, b_4\}$ with four elements?

Solution: To define a function $A \xrightarrow{f} B$, we need to specify an image for each a_i . Think of this as a procedure with three tasks T_1, T_2, T_3 , where each T_i is the task of specifying an image for a_i . Then there are 4 ways of doing each T_i since there are 4 elements in B . Thus by the product rule, the number of functions $A \xrightarrow{f} B$ is $4 \cdot 4 \cdot 4 = 64$.

6.1.8 Example. How many *injective* (one-to-one) functions are there from $A = \{a_1, a_2, a_3\}$ to $B = \{b_1, b_2, b_3, b_4\}$?

Solution: To define an injective function $A \xrightarrow{f} B$, we need to specify an image for each $a_i \in A$, and none of these images can be the same. As before, let T_i be the task of specifying an image for a_i , and think of doing T_1 , then T_2 , and then T_3 . Then there

are 4 ways to do T_1 , but only 3 ways to do T_2 , because a_2 can't map to the image you chose for T_1 . Similarly, there are only 2 ways to do T_3 , because you cannot map a_3 to the images of a_1 or a_2 . Thus by the product rule, the number of injective functions is $4 \cdot 3 \cdot 2 = 24$.

6.1.9 Note. In Example 6.1.8, the choice of image for a_1 did affect the possible images for a_2 . For instance, if $a_1 \mapsto b_1$, then a_2 can only map to b_2 , b_3 , or b_4 , whereas if $a_1 \mapsto b_2$, then a_2 can only map to b_1 , b_3 , or b_4 . But the crucial point is that the *number of ways* of choosing an image of a_2 is 3, regardless of which image for a_1 you chose. Thus the tasks of choosing the images for the a_i are independent, allowing us to use the product rule.

6.1.10 Exercise. Count some things! Prove $2^n \dots$

6.2 Permutations

Let's increase the complexity of the objects we are counting.

6.2.1 Definition. An r -**permutation** of a set S is a sequence in S of length r that contains no repeated elements.

6.2.2 Example. The 1-permutations of $\{a, b, c\}$ are (a) , (b) , and (c) . The 2-permutations are (a, b) , (b, a) , (a, c) , (c, a) , (b, c) , (c, b) . The 3-permutations are (a, b, c) , (a, c, b) , (b, a, c) , (b, c, a) , (c, a, b) , (c, b, a) . There is exactly one 0-permutation of $\{a, b, c\}$, namely the empty sequence $()$, and there are no 4-permutations.

6.2.3 Note. Note that we write permutations in parentheses since they are *sequences*, not sets (see 3.1.20). In particular, the order in which the elements of a *sequence* are listed is important, whereas the order in which the elements of a *set* are written is irrelevant.

How many r -permutations of a set S of n elements are there? Equivalently, we want to count the number of ways to do the procedure

★ Procedure: choose a sequence of r elements of S . (How many ways?)

To count the ways, we break up the procedure into r tasks:

- T_1 : choose the first element of the sequence. (n ways)
- T_2 : choose the second element of the sequence. ($n - 1$ ways)
- ⋮
- T_r : choose the r th element of the sequence. ($n - r + 1$ ways)

Thus by the product rule, we get

6.2.4 Proposition. Let $n, r \in \mathbb{Z}_{\geq 0}$ with $r \leq n$. Then the number of r -permutations of a set with n elements is

$$P(n, r) = n(n-1)(n-2) \cdots (n-r+1).$$

A useful way to express this product is using factorial notation.

6.2.5 Definition. Let $n \in \mathbb{Z}_{\geq 1}$. Then define n **factorial**, written $n!$, to be the product $n! = n(n-1)(n-2) \cdots 1$. It is also convenient to define $0! = 1$.

6.2.6 Example. $0! = 1$, $1! = 1$, $2! = 2$, $3! = 6$, $4! = 24$, $5! = 120$, $6! = 720$, $7! = 5040$, $10! = 3628800$.

With this notation we get:

6.2.7 Corollary. Let $n, r \in \mathbb{Z}_{\geq 0}$ with $r \leq n$. Then $P(n, r) = \frac{n!}{(n-r)!}$.

6.2.8 Note. Note that $P(n, n) = n!$. That is, the number of ways to order all n elements of a set is $n!$.

6.2.9 Exercise. Exercises about permutations!

6.3 Combinations

Suppose now that we don't care about the order in which choices are made.

6.3.1 Example. What are the possible 3-person teams in a group of 5 people? Equivalently, what are the 3-element subsets of a set with 5 elements, say $\{a, b, c, d, e\}$? The 10 possibilities are

$$\{a, b, c\}, \{a, b, d\}, \{a, b, e\}, \{a, c, d\}, \{a, c, e\}, \{a, d, e\}, \{b, c, d\}, \{b, c, e\}, \{b, d, e\}, \{c, d, e\}.$$

6.3.2 Definition. An r -**combination** of a set S is an r -element subset of S .

How many r -combinations of a set S of n elements are there? We can deduce this from our study of permutations. The only difference between an r -permutation and an r -combination is that in an r -permutation, the chosen elements are given an order. This observation inspires the following analysis:

★ Procedure: choose an r -permutation of S . ($P(n, r)$ ways)

Now we cleverly break up the procedure into 2 tasks:

- $T1$: choose an r -combination of S . (How many ways?)
- $T2$: choose an order for the r -combination, namely an r -permutation of the r -combination. ($P(r, r)$ ways)

Thus, by the product rule, the number $C(n, r)$ (which we read as “ n choose r ”) of r -combinations satisfies

$$P(n, r) = C(n, r) \cdot P(r, r).$$

Solving for $C(n, r)$, we get

$$C(n, r) = \frac{P(n, r)}{P(r, r)} = \frac{n!/(n-r)!}{r!} = \frac{n!}{(n-r)!r!}.$$

Thus we have proved:

6.3.3 Proposition. *Let $n, r \in \mathbb{Z}_{\geq 0}$ and $r \leq n$. Then the number of r -combinations of a set with n elements is*

$$C(n, r) = \frac{n!}{(n-r)!r!}.$$

6.3.4 Example. Returning to Example 6.3.1, the number of 3-person teams in a group of 5 people is

$$C(5, 3) = \frac{5!}{(5-3)!3!} = 10,$$

which agrees with the list we compiled.

6.3.5 Example. How many distinct 5-card hands are possible in a deck of 52 cards? Simply compute:

$$C(52, 5) = \frac{52!}{47!5!} = \frac{52 \cdot 51 \cdot 50 \cdot 49 \cdot 48}{5 \cdot 4 \cdot 3 \cdot 2 \cdot 1} = 52 \cdot 51 \cdot 10 \cdot 49 \cdot 2 = 2598960.$$

6.3.6 Example. How many 5-card hands are a flush (have all 5 cards of the same suit)? Think of choosing a flush as a procedure consisting of two tasks: first choose a suit, then choose 5 cards from that suit. There are $\binom{4}{1}$ ways to choose a suit, and $C(13, 5)$ ways to choose 5 cards from that suit, so by the product rule the number of flush hands is

$$\binom{4}{1} \cdot \binom{13}{5} = 4 \cdot \frac{13!}{8!5!} = 4 \cdot \frac{13 \cdot 12 \cdot 11 \cdot 10 \cdot 9}{5 \cdot 4 \cdot 3 \cdot 2 \cdot 1} = 13 \cdot 12 \cdot 11 \cdot 3 = 5148.$$

An easy corollary of the theorem is

6.3.7 Corollary. $C(n, r) = C(n-r, r)$.

This is intuitive because choosing a subset of r elements of a set n is equivalent to choosing the $n-r$ elements to *leave out* of the subset.

6.3.8 Exercise. Exercises about combinations!

6.4 Binomial Theorem

The binomial theorem is a formula for expanding powers of the form $(x + y)^n$. As we will see, the formula is intimately related to combinations.

6.4.1 Definition. A **binomial** is the sum of two terms, for instance $x + y$.

We want to study powers of the binomial $x + y$. The first few of these are

$$\begin{aligned}(x + y)^1 &= x + y, & (x + y)^2 &= x^2 + 2xy + y^2, & (x + y)^3 &= x^3 + 3x^2y + 3xy^2 + y^3, \\ (x + y)^4 &= x^4 + 4x^3y + 6x^2y^2 + 4xy^3 + y^4, & (x + y)^5 &= x^5 + 5x^4y + 10x^3y^2 + 10x^2y^3 + 5xy^4 + y^5.\end{aligned}$$

Let's try to find a general formula for $(x + y)^n$. Expand the power $(x + y)^n$ as a product of n factors

$$(x + y)(x + y) \cdots (x + y).$$

If we multiply this out without grouping like terms, then every term arises from a choice of either x or y in each factor of $x + y$. For example, for $(x + y)(x + y)$, choosing x in both factors yields x^2 , choosing x in the first and y in the second yields xy , choosing y in the first and x in the second yields yx (which is the same as xy), and choosing y in both yields y^2 . Note that we got xy twice, because there were two ways of choosing one x and one y .

Let's return to the product of n factors of $(x + y)$. The possible terms are $x^n, x^{n-1}y, x^{n-2}y^2, \dots, y^n$. We want to count how many times each one occurs. Pick a term, say $x^{n-j}y^j$, and focus on the power of y of that term. In order to get exactly the j th power of y , you have to choose y in exactly j of the n factors (all the rest will be x), and there are $C(n, j)$ ways to do this. For instance, when $j = 0$, the term is x^n , and there is only $C(n, 0) = 1$ way to choose y in 0 of the factors. Likewise, when $j = 1$, there are $C(n, 1) = n$ ways to choose y in just one factor.

Thus for each $0 \leq j \leq n$, the coefficient of $x^{n-j}y^j$ is $C(n, j)$. These $C(n, j)$ are so important that we use a special notation for them: $\binom{n}{j}$. We call them **binomial coefficients** because in this notation, we can write our result as:

6.4.2 Theorem (Binomial theorem). *Let x and y be variables, and $n \in \mathbb{Z}_{\geq 1}$. Then*

$$(x + y)^n = \sum_{j=0}^n \binom{n}{j} x^{n-j} y^j = \binom{n}{0} x^n + \binom{n}{1} x^{n-1} y + \binom{n}{2} x^{n-2} y^2 + \cdots + \binom{n}{n} y^n.$$

The binomial theorem is a fantastic way to compute powers of binomials. Check that the binomial theorem gives the same results as the powers of $x + y$ above! We can also compute powers of other binomials by substituting whatever we like for x and y :

6.4.3 Example. Compute $(x + 1)^4$. We just replace y by 1 in the binomial theorem, to get

$$\begin{aligned}(x + 1)^4 &= \binom{4}{0} x^4 1^0 + \binom{4}{1} x^3 1^1 + \binom{4}{2} x^2 1^2 + \binom{4}{3} x^1 1^3 + \binom{4}{4} x^0 1^4 \\ &= x^4 + 4x^3 + 6x^2 + 4x + 1.\end{aligned}$$

6.4.4 Example. Compute $(2x - 3)^4$. For this, we replace x by $2x$ and y by -3 in the binomial theorem. This yields

$$\begin{aligned}(2x - 3)^4 &= \binom{4}{0}(2x)^4(-3)^0 + \binom{4}{1}(2x)^3(-3)^1 + \binom{4}{2}(2x)^2(-3)^2 + \binom{4}{3}(2x)^1(-3)^3 + \binom{4}{4}(2x)^0(-3)^4 \\ &= 16x^4 - 96x^3 + 216x^2 - 216x + 81.\end{aligned}$$

Setting $x = y = 1$ in the binomial theorem, we get an identity relating 2^n to a sum of binomial coefficients:

6.4.5 Corollary. *Let $n \in \mathbb{Z}_{\geq 1}$. Then $2^n = \sum_{j=0}^n \binom{n}{j}$.*

6.4.6 Example. How many subsets does a set $S = \{a_1, \dots, a_n\}$ with n elements have? The total number of subsets is the number of subsets with 0 elements, namely $\binom{n}{0}$, plus the number of subsets with 1 element, which is $\binom{n}{1}$, plus the number of subsets with 2 elements, and so on. By Corollary 4, adding up all these binomial coefficients gives 2^n .

The number of subsets of S can also be computed using just the product rule. Think of choosing a subset as a procedure, consisting of tasks T_1, \dots, T_n , where the task T_i is choosing whether or not to include a_i in the subset. There are two ways to do each T_i (either include a_i or don't), so the total number of ways to do the procedure is $2 \cdot 2 \cdots 2 = 2^n$.

We can obtain another identity by setting $x = 1$ and $y = -1$ in the binomial theorem:

6.4.7 Corollary. *Let $n \in \mathbb{Z}_{\geq 1}$. Then*

$$0 = \sum_{j=0}^n (-1)^j \binom{n}{j}.$$

This is not surprising when n is odd because the symmetry $\binom{n}{j} = \binom{n}{n-j}$ and the alternating signs will cause all the terms to cancel in pairs. But it is an interesting identity when n is even!

6.5 Pascal's Triangle

There is a beautiful way to generate the binomial coefficients, called **Pascal's triangle**. The first row is an infinite row of 0s with a single 1 in the middle. The second row is obtained as follows: for every pair of adjacent entries of the first row, add the two numbers, and write the sum below the midpoint of the two numbers, in the second row. The third row is obtained by taking sums of pairs of entries in the second row, and so on. Here are the first 9 rows of the triangle, with the 0s omitted:

Combinatorial proof. We can also prove the identity by finding a different way to count the $\binom{n+1}{k+1}$ subsets of $k+1$ elements of a set $S = \{a_1, a_2, \dots, a_{n+1}\}$ of $n+1$ elements. First, we count the subsets with $k+1$ elements that contain a_1 . Since a_1 is in the subset, the subset must contain k of the remaining n elements $\{a_2, \dots, a_{n+1}\}$, so there are $\binom{n}{k}$ such subsets. Second, we count the subsets that do not contain a_1 . Such a subset must contain $k+1$ of the remaining n elements, thus the number of such subsets is $\binom{n}{k+1}$. Since every subset of $k+1$ elements either contains a_1 or doesn't contain a_1 , the total number of subsets of $k+1$ elements is $\binom{n}{k} + \binom{n}{k+1}$, which proves the identity. \square

6.5.2 Note. If you don't feel like messing with factorials, Pascal's triangle can be a painless way to find the binomial coefficients!

We can also use Pascal's identity to give another proof of the binomial theorem, using induction:

Theorem (Binomial theorem). *Let x and y be variables, and $n \in \mathbb{Z}_{\geq 1}$. Then*

$$(x+y)^n = \sum_{j=0}^n \binom{n}{j} x^{n-j} y^j = \binom{n}{0} x^n + \binom{n}{1} x^{n-1} y + \binom{n}{2} x^{n-2} y^2 + \dots + \binom{n}{n} y^n.$$

Inductive proof of the binomial theorem. Induction on n . For the base case $n = 1$, note that $(x+y)^1 = x+y = \binom{1}{0}x + \binom{1}{1}y$. Now suppose the binomial theorem holds for some $k \in \mathbb{Z}^{>0}$, namely that

$$(x+y)^k = \binom{k}{0} x^k + \binom{k}{1} x^{k-1} y + \binom{k}{2} x^{k-2} y^2 + \dots + \binom{k}{k} y^k.$$

Multiplying both sides by $(x+y)$, we get

$$\begin{aligned} (x+y)^{k+1} &= \binom{k}{0} x^{k+1} + \binom{k}{1} x^k y + \binom{k}{2} x^{k-1} y^2 + \dots + \binom{k}{k} x y^k \\ &\quad + \binom{k}{0} x^k y + \binom{k}{1} x^{k-1} y^2 + \dots + \binom{k}{k-1} x y^k + \binom{k}{k} y^{k+1}. \end{aligned}$$

We can replace $\binom{k}{0}$ by $\binom{k+1}{0}$ and $\binom{k}{k}$ by $\binom{k+1}{k+1}$ since these binomial coefficients are all equal to 1. Moreover, we can group each pair of like terms and apply Pascal's identity to the sum of their coefficients, to obtain

$$(x+y)^{k+1} = \binom{k+1}{0} x^{k+1} + \binom{k+1}{1} x^k y + \binom{k+1}{2} x^{k-1} y^2 + \dots + \binom{k+1}{k} x y^k + \binom{k+1}{k+1} y^{k+1}.$$

This is the statement of the binomial theorem for $n = k+1$, so we are done by induction. \square

6.6 Application of Combinations: 5-Card Poker Hands

What is the probability (chance, likelihood) of getting a pair (two cards of the same rank), a triple, a quadruple, 2-pair, a full house (a pair and a triple), a flush (five cards of the same suit), a straight (five consecutive ranks), or a straight flush in a 5-card poker hand from a standard 52-card deck? We assume the 5-card hand is completely random. Then we can calculate the probability $P(H)$ of getting a particular kind of hand H by counting the number $N(H)$ of such hands and then dividing by the total number N of 5-card hands. Namely,

$$P(H) = \frac{N(H)}{N}.$$

Note that the probability of getting a particular kind of hand will always be a real number between 0 and 1 since $0 \leq N(H) \leq N$.

Since we are considering 5-card hands coming from a 52-card deck, the total number of hands is

$$N = \binom{52}{5} = 2598960.$$

Using combinatorics, we will compute the values in the following table:

Type of Hand H	Number of Hands $N(H)$	Combinatorial Expression for $N(H)$	Probability $P(H)$
Pair	1098240	$\binom{13}{1} \binom{4}{2} \binom{12}{3} \binom{4}{1}^3$	0.42257
Triple	54912	$\binom{13}{1} \binom{4}{3} \binom{12}{2} \binom{4}{1}^2$	0.02113
Quadruple	624	$\binom{13}{1} \binom{4}{4} \binom{12}{1} \binom{4}{1}$	0.00024
2-pair	123552	$\binom{13}{2} \binom{4}{2}^2 \binom{11}{1} \binom{4}{1}$	0.04754
Full house	3744	$\binom{13}{1} \binom{4}{3} \binom{12}{1} \binom{4}{2}$	0.00144
Straight flush	40	$\binom{10}{1} \binom{4}{1}^5$	0.00002
Flush	5108	$\binom{4}{1} \binom{13}{5} - \binom{10}{1} \binom{4}{1}$	0.00197
Straight	10200	$\binom{10}{1} \binom{4}{1}^5 - \binom{10}{1} \binom{4}{1}$	0.00392
Total	1296420		0.49882

Now we'll show how some of the values $N(H)$ can be computed.

6.6.1 Example (Pair).

★ Procedure: Choose a hand with a pair (exactly two cards of the same rank).

Choose the pair.

- $T1$: Choose a rank for the pair: $\binom{13}{1}$ ways.
- $T2$: Choose two cards of that rank: $\binom{4}{2}$ ways.

Choose the remaining three cards.

- $T3$: Choose three other ranks: $\binom{12}{3}$ ways.
- $T4$: Choose one card for each of those three ranks: $\binom{4}{1}^3$ ways.

Thus by the product rule, $N(\text{pair}) = \binom{13}{1} \binom{4}{2} \binom{12}{3} \binom{4}{1}^3$.

Note that by choosing the remaining three cards to be of different ranks, we ensure that the hand does not have a triple or two-pair. Also, a flush is impossible since the two cards of the same rank we chose for the pair must have different suits. One important detail is that we chose the ranks for the three remaining cards all at once (using $\binom{12}{3}$), instead of choosing those ranks one-by-one (using $\binom{12}{1} \binom{11}{1} \binom{10}{1}$). This is crucial, because choosing one-by-one imposes an order on our choices that results in overcounting.

6.6.2 Example (2-pair).

★ Procedure: Choose a hand with two pairs.

First, choose the two pairs:

- $T1$: Choose two ranks for the two pairs: $\binom{13}{2}$ ways.
- $T2$: Choose two cards of each of those ranks: $\binom{4}{2}^2$ ways.

Then choose the last card:

- $T3$: Choose one other rank: $\binom{11}{1}$ ways.
- $T4$: Choose one card of that rank: $\binom{4}{1}$ ways.

So by the product rule, $N(2\text{-pair}) = \binom{13}{2} \binom{4}{2}^2 \binom{11}{1} \binom{4}{1}$.

Once again, it is crucial that we choose the ranks of the pairs at once, otherwise we would be imposing an unwanted order on the two pairs that would result in overcounting.

6.6.3 Example (Straight flush).

★ Procedure: Choose a hand with a straight flush.

The trick is that the ranks occurring in the straight are determined by the highest rank, which can be 5, 6, 7, 8, 9, 10, J, Q, K, A. (We allow the straight A 2 3 4 5.)

- $T1$: Choose the highest card of the straight: $\binom{10}{1}$ ways.
- $T2$: Choose the suit: $\binom{4}{1}$ ways.

So $N(\text{straight flush}) = \binom{10}{1} \binom{4}{1}$.

6.6.4 Example (Flush).

★ Procedure: Choose a hand with a flush.

First, we count how many hands have a flush.

- $T1$: Choose the suit for the flush: $\binom{4}{1}$ ways.
- $T2$: Choose 5 cards of that suit: $\binom{13}{5}$ ways.

We also have to exclude the number $\binom{10}{1} \binom{4}{1}$ of straight flushes, since those count as a straight flush rather than just a flush. Thus

$$N(\text{flush}) = \binom{4}{1} \binom{13}{5} - \binom{10}{1} \binom{4}{1}.$$

6.6.5 Exercise. Compute combinatorial expressions for the remaining special hands. Write down tasks and use the product rule! You can check your answers in the table above.

6.6.6 Example (Getting nothing). How many hands are not one of the above special hands? We can simply take the total number of hands and subtract the number of special hands:

$$N(\text{nothing}) = 2598960 - 1296420 = 1302540.$$

Another way to compute $N(\text{nothing})$ is as follows. The main idea is that to get nothing, the ranks of the five cards have to be different. The only special hands with all ranks different are the flushes, straights, and straight flushes, so we just have to subtract the numbers of those from the number of hands with five cards of different ranks. Thus we get

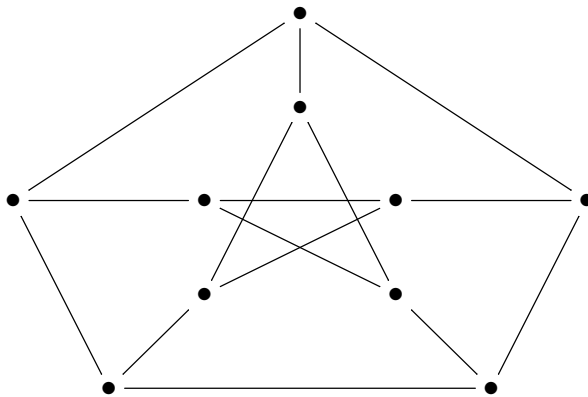
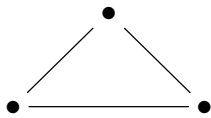
$$N(\text{nothing}) = \binom{13}{5} \binom{4}{1}^5 - N(\text{flush}) - N(\text{straight}) - N(\text{straight flush}) = 1302540,$$

as before.

Chapter 7

Graph Theory

Let's start a brand new topic, called **graph theory**! The kinds of graphs studied in graph theory are completely different from graphs of functions; instead, a **graph** is a collection of points in the plane, called **vertices**, with some lines drawn between them, called **edges**. Here are two examples:

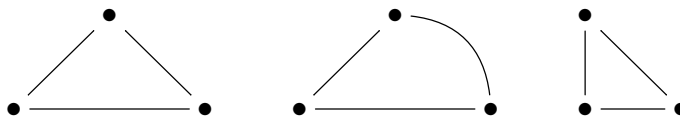


In a graph:

We care about	We don't care about
The number of vertices.	How the vertices are placed in the plane.
The number of edges.	How the edges are drawn.
Which vertices are connected by which edges.	Edge crossings at points that are not vertices.

Thus we consider the following three graphs to be the same:

Same:



Graphs may have **multiple edges** between two vertices, as well as **loops**, which are edges that are connected to only one vertex. For example:



We want to study the properties of graphs rigorously and prove things about them. For this we need to formulate a more careful definition of what a graph is. The goal of our new definition is to capture the information we care about, while ignoring the data we don't care about.

7.1 Graphs

A common theme in pure mathematics is the following: in order to study a geometric object, it is useful to define its structure abstractly. For instance, above we couldn't precisely define what it should mean for graphs to be the "same", but we will be able to do this with our new definition.

7.1.1 Definition. A (finite, undirected) **graph** $G = (V_G, E_G, \phi_G)$ consists of a set V_G , whose elements are called **vertices**, a set E_G , whose elements are called **edges**, and a function $E_G \xrightarrow{\phi} \Sigma_G$, where Σ_G is the set of all one or two element subsets of V_G , which assigns one or two vertices, called **endpoints**, to each edge. We say an edge is **incident** to its endpoints. If $\phi_G(e) = \{v, w\}$, we say e **connects** the vertices v and w . If $\phi_G(e)$ is a single vertex, then we call e a **loop**. The **degree** $\deg(v)$ of a vertex v is the number of edges incident to it, with loops counted twice. The number of vertices of G is $|V_G|$ and the number of edges of G is $|E_G|$.

Note that this definition captures the data in the "We care about" column of the above table, while throwing out what "We don't care about".

7.1.2 Note. We sometimes write V, E, ϕ, Σ instead of $V_G, E_G, \phi_G, \Sigma_G$ when it is clear which graph G is being discussed.

7.1.3 Example. Let G be the graph with $V_G = \{u, v, w\}$, $E_G = \{e, f\}$, and the map $E_G \xrightarrow{\phi_G} \Sigma_G$ defined by $e \mapsto \{u, v\}$, $f \mapsto \{v, w\}$. We have $\deg(u) = 1$, $\deg(v) = 2$, and $\deg(w) = 1$. Also, $|V_G| = 3$ and $|E_G| = 2$.

7.1.4 Example. Let H be the graph with $V_H = \{x\}$, $E_H = \{c, d\}$, and the map $E_H \xrightarrow{\phi_H} \Sigma_H$ defined by $c \mapsto \{x\}$, $d \mapsto \{x\}$. For this graph, $\deg(x) = 4$, since loops are counted twice. Also, $|V_H| = 1$ and $|E_H| = 2$.

This definition of graph is terribly abstract, but there is an easy procedure to recover the geometric picture:

7.1.5 Definition. A **drawing** of a graph is obtained by:

- Drawing each vertex as a point in the plane.
- Drawing each edge as a line connecting its endpoints.

7.1.6 Example. One way to draw the graph G from the example above is



One way to draw H is



There are many ways to draw a graph, but fortunately:

7.1.7 Note. Any two drawings of a graph are the “same”, in terms of the criteria in the table above. In other words, any two drawings of a given graph differ only in where the vertices are drawn in the plane and how the edges are drawn.

7.1.8 Exercise. Prove or disprove the following claim: Let $G = (V, E, \phi)$ be a graph. Then the sum of the degrees of all vertices in G is equal to twice the number of edges of G , namely

$$\sum_{v \in V} \deg(v) = 2 \cdot |E|.$$

7.2 Isomorphism

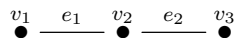
Intuitively, we want to think about graphs in terms of their drawings, without explicitly thinking about the sets V_G and E_G and the function ϕ_G . To justify this, we need to think hard about the connection between the abstract definition of a graph and the data inherent in a drawing of a graph. Because any two drawings of a graph are the “same”, we get a function

$$\{\text{graphs}\} \xrightarrow{\text{draw the graph}} \{\text{drawings up to “sameness”}\}.$$

(Recall that a function maps each element of the domain to exactly one element of the codomain. Thus to get a function, we need a “unique” way of drawing the graph, which we have because any two drawings of a graph are the “same”.)

Our goal is to get a function going the other direction, ideally an inverse of the “draw the graph” function. How can we get a graph from a drawing?

7.2.1 Note. Given a drawing of a graph, we can define sets V and E and a function ϕ . For instance, we can choose a labeling of vertices and edges



and then write $V = \{v_1, v_2, v_3\}$ and $E = \{e_1, e_2\}$ and define $E \xrightarrow{\phi} \Sigma$ by $\phi(e_1) = \{v_1, v_2\}$ and $\phi(e_2) = \{v_2, v_3\}$.

So we have a way of going from a drawing to a graph. Unfortunately, this will not give us the function we want, because there are many, many ways to choose labels for a given drawing. The problem isn't in our way of labeling a drawing, but rather in our definition of graph, which has too much information. In order to define a graph, you have to choose a set of vertices and a set of edges, which forces you to choose names for your vertices and edges.

Should the exact labeling be so important? Mathematically, no! When studying graphs, we care mostly about the relative arrangement of vertices and edges, and much less about the exact labels chosen. Unfortunately, the natural definition of “sameness” for graphs is:

7.2.2 Definition. Two graphs $G = (V_G, E_G, \phi_G)$ and $H = (V_H, E_H, \phi_H)$ are **equal** if $V_G = V_H$, $E_G = E_H$, and $\phi_G = \phi_H$.

Equality of graphs is too strict, because if the names of the elements in V_G and V_H are different, then $V_G \neq V_H$. So we need a different notion of when graphs are the “same”.

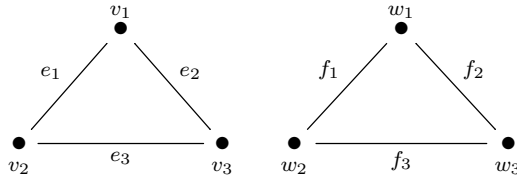
Here is the notion of “sameness” for graphs that we want. It is constructed exactly so that different labelings of the same drawing can be considered the “same”.

7.2.3 Definition. An **isomorphism** of two graphs $G = (V_G, E_G, \phi_G)$ and $H = (V_H, E_H, \phi_H)$ is a pair of bijections $V_G \xrightarrow{\nu} V_H$ and $E_G \xrightarrow{\epsilon} E_H$ such for that all $e \in E_G$, $\phi_H(\epsilon(e)) = \{\nu(v_1), \nu(v_2)\}$, where $\phi_G(e) = \{v_1, v_2\}$.

What the last condition means is that for each edge e of G , with endpoints $\{v_1, v_2\}$, the endpoints of the corresponding edge $\epsilon(e)$ in H are $\{\nu(v_1), \nu(v_2)\}$. In other words, the bijection of the edges is “compatible” with the bijection of the vertices with regard to how edges connect vertices.

7.2.4 Definition. Two graphs are **isomorphic** if there is an isomorphism between them.

7.2.5 Example. Let's show that the graphs corresponding to any two labelings



are isomorphic. Let ϕ_1 denote the map from the edges of the left graph to their endpoints, and let ϕ_2 be the corresponding map for the right graph. The most obvious isomorphism is given by

$$\begin{aligned}\nu(v_1) &= w_1, \nu(v_2) = w_2, \nu(v_3) = w_3 \\ \epsilon(e_1) &= f_1, \epsilon(e_2) = f_2, \epsilon(e_3) = f_3.\end{aligned}$$

Since $\phi_1(e_1) = \{v_1, v_2\}$ and $\phi_2(f_1) = \{w_1, w_2\}$, we see that $\phi_2(\epsilon(e_1)) = \{w_1, w_2\} = \{\nu(v_1), \nu(v_2)\}$, so the isomorphism condition holds for e_1 . Checking the condition for e_2 and e_3 is similar.

This is only one possible isomorphism for these two graphs. In fact, there are six such isomorphisms. Can you see why? (Hint: the isomorphism is determined by the map ν on the vertices.)

As the example shows, it is very tedious to explicitly write down isomorphisms and check all the conditions. Fortunately, the following proposition (which is obvious if you understand the definitions!) guarantees that we need not worry about which labeling we choose for a given drawing. Moreover, isomorphic graphs have the same drawings.

7.2.6 Proposition. *Any two graphs obtained by labeling a given drawing are isomorphic. Any two drawings of two isomorphic graphs are the “same”.*

Thus we finally get our map in the opposite direction:

$$\{\text{graphs up to isomorphism}\} \begin{array}{c} \xrightarrow{\text{draw the graph}} \\ \xleftarrow{\text{label the drawing}} \end{array} \{\text{drawings up to “sameness”}\}.$$

(We get a function “label the drawing” because any drawing gives us a unique graph when we consider isomorphic graphs to be the same.)

In fact,

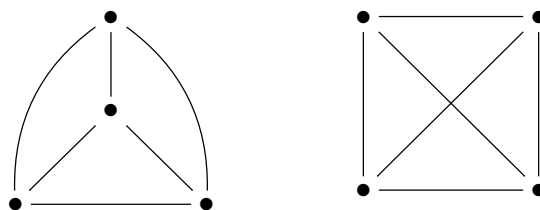
7.2.7 Proposition. *The maps “draw the graph” and “label the drawing” are inverses, and thus give a bijection between the set of graphs up to isomorphism and the set of drawings up to “sameness”.*

7.2.8 Note. From now on, we will always consider graphs to be the “same” if they are isomorphic. Thus the previous theorem allows us to refer to drawings of graphs as graphs, since each drawing of a graph corresponds naturally to a unique graph. Conversely, we can always think about a graph in terms of a drawing. The bijection also allows us to use the rigorous language of graph isomorphism, which is defined very precisely above, instead of referring to our fuzzy notion of “sameness” of drawings.

7.2.9 Note. If we think of graphs as drawings, then two graphs are isomorphic if and only if you can move around the vertices in the first graph and change how the edges are drawn (without changing their endpoints) to get the second graph.

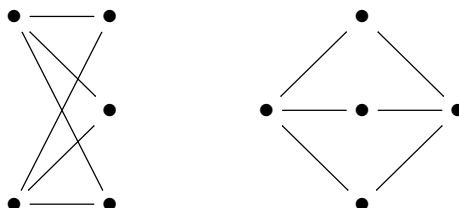
We’ll end the section with few more examples of isomorphic graphs.

7.2.10 Example. The following two graphs are isomorphic:



Think about which bijections of the sets of vertices could give an isomorphism. Note that the bijection of the vertices in a graph isomorphism determines the bijection of the edges (to see where to map an edge, look at where its endpoints are being mapped!). Alternatively, think about how you can move the vertices and redraw the edges in the first graph to get the second graph.

Another pair of isomorphic graphs is



Which bijections of the sets of vertices can give the isomorphism? How can you move the vertices and redraw the edges?

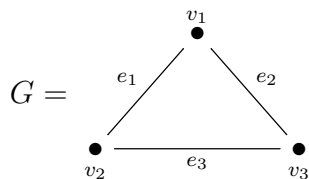
7.2.11 Note. As in the example, we like to think of graphs visually in terms of their drawings. But the abstract definition is still crucially important because it is the foundation for rigorous proofs. We will prove plenty of things about graphs in the coming weeks, and you will see how important it is to have set descriptions of the vertices and edges of graphs.

7.3 Some Types of Graphs

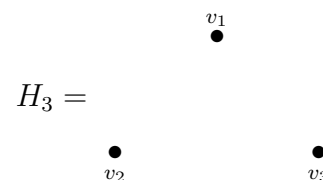
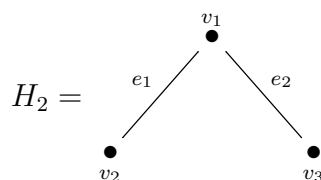
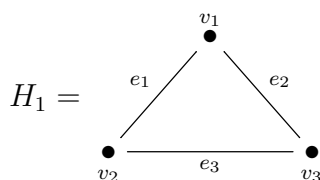
Whenever you see a new definition in mathematics, try to write down examples! This both forces you to think about each part of the definition and also gives you something to visualize when the definitions start to pile up (which they always do!).

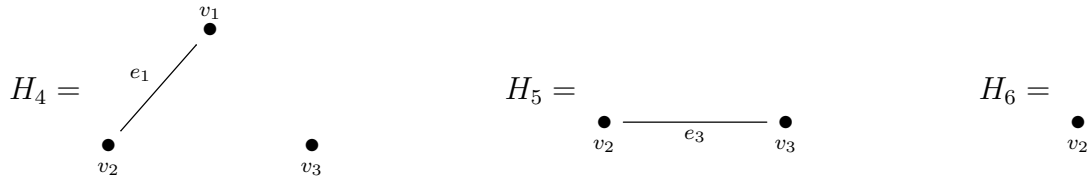
7.3.1 Definition. A **subgraph** of a graph $G = (V_G, E_G, \phi_G)$ is a graph $H = (V_H, E_H, \phi_H)$ such that $V_H \subseteq V_G$, $E_H \subseteq E_G$, and ϕ_H is the restriction of the function $E_G \xrightarrow{\phi_G} \Sigma_G$ to the domain E_H and codomain Σ_H .

7.3.2 Example. Six of the subgraphs of

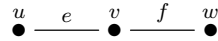


are





Since we usually consider isomorphic graphs to be the same, we may also think of a graph like



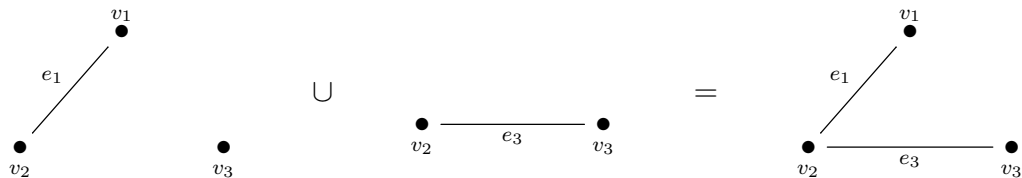
as a subgraph of G , as long as we specify an isomorphism onto a subgraph, for instance onto H_2 .

7.3.3 Note. When describing a subgraph H of a graph G , we often don't explicitly define the function ϕ_H since it is fully determined by ϕ_G .

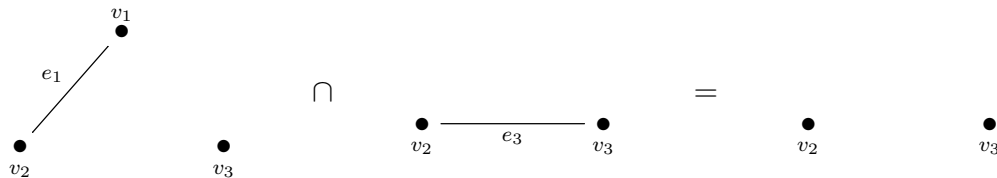
7.3.4 Exercise. Prove the following claim: Let G and H be graphs. Then $G = H$ if and only if G is a subgraph of H and H is a subgraph of G .

7.3.5 Definition. Let H_1, H_2 be subgraphs of a graph G . The **union** of H_1 and H_2 , denoted $H_1 \cup H_2$, is the subgraph of G with vertex set $V_{H_1} \cup V_{H_2}$ and edge set $E_{H_1} \cup E_{H_2}$. The **intersection** of H_1 and H_2 , denoted $H_1 \cap H_2$, is the subgraph of G with vertex set $V_{H_1} \cap V_{H_2}$ and edge set $E_{H_1} \cap E_{H_2}$.

7.3.6 Example. Returning to the previous example,



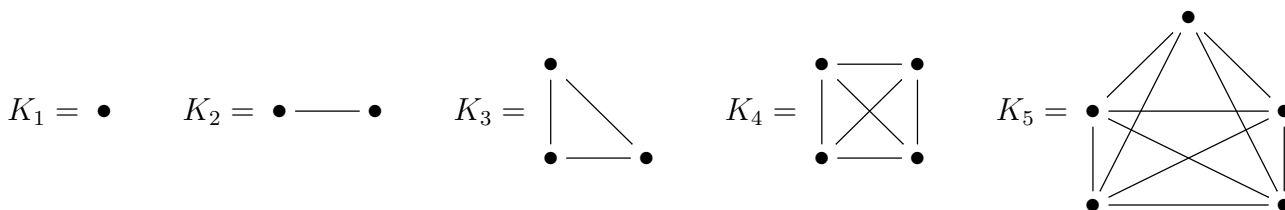
and



7.3.7 Definition. A graph is said to have **multiple edges** if there are two distinct edges connecting the same two vertices, as in the graph $\bullet \text{---} \text{---} \bullet$. A graph is said to be **simple** if it has no loops and no multiple edges.

7.3.8 Definition. Let $n \in \mathbb{Z}^{\geq 1}$. The **complete graph** K_n is the simple graph with n vertices and exactly one edge between each pair of distinct vertices.

7.3.9 Example. The first five complete graphs are

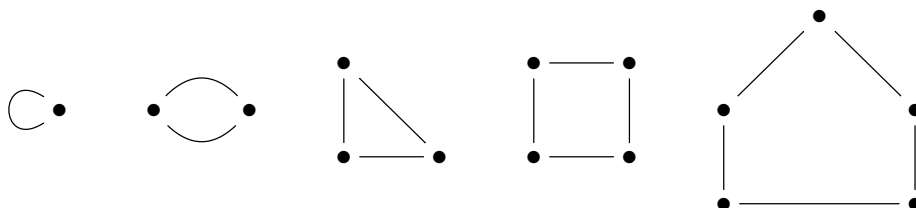


7.3.10 Exercise. Let $G = (V, E, \phi)$ be a graph. Prove or disprove the following claim: If G is simple, then

$$|E| \leq \binom{|V|}{2}.$$

7.3.11 Definition. Let $n \in \mathbb{Z}_{\geq 1}$. A **cycle of length n** is a graph isomorphic to the graph with vertex set $\{v_1, \dots, v_n\}$, edge set $\{e_1, \dots, e_n\}$, and function $\phi(e_1) = \{e_n, e_2\}$ and $\phi(e_i) = \{v_{i-1}, v_{i+1}\}$ for all $2 \leq i \leq n$. We say a graph **has a cycle of length n** if it has a subgraph isomorphic to a cycle of length n .

7.3.12 Example. Cycles of length 1 through 5 are



7.3.13 Example. The graph



has two cycles of length 1 (the loops), 2 cycles of length 2 (the sets of multiple edges), and 4 cycles of length 3 (can you see why?). It has no cycles of length ≥ 4 since it has only 3 vertices.

7.3.14 Example. The complete graph K_n has cycles of every length between 3 and n .

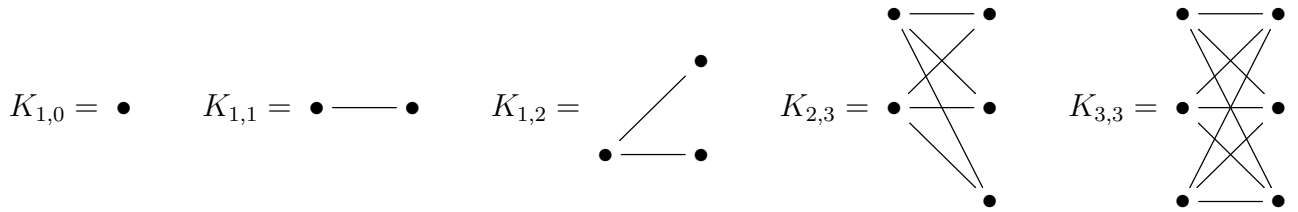
7.3.15 Exercise. Let G be a graph. Prove or disprove the following claim: G is simple if and only if it has no cycles of length 1 or 2.

7.3.16 Definition. A graph $G = (V, E, \phi)$ is **bipartite** if there is a partition of the vertices $V = V_1 \sqcup V_2$ (disjoint union) such that every edge $e \in E$ has one endpoint in V_1 and the other endpoint in V_2 .

7.3.17 Example. K_1 and K_2 are bipartite, but all other complete graphs are not bipartite.

7.3.18 Definition. Let $n, m \in \mathbb{Z}^{\geq 0}$. The **complete bipartite graph $K_{n,m}$** is the simple graph consisting of a set V_1 of n vertices, a set V_2 of m vertices, and exactly one edge from each vertex in V_1 to each vertex in V_2 .

7.3.19 Example. Some complete bipartite graphs are



7.3.20 Exercise. Let G be a graph and H be a subgraph of G . Prove or disprove the following claims:

- (a) If G is simple, then H is simple.
- (b) If G is a cycle, then H is a cycle.
- (c) If G is bipartite, then H is bipartite.
- (d) If G is complete bipartite, then H is complete bipartite.
- (e) If G is bipartite and H is a cycle, then H has even length.

7.3.21 Exercise. Let $G = (V, E, \phi)$ be a graph. Prove or disprove the following claims:

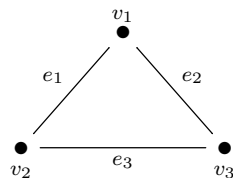
- (a) If G is simple, then $|E| \leq \binom{|V|}{2}$.
- (b) If G is bipartite, then
- (c)
- (d)
- (e)

7.4 Connected Graphs

7.4.1 Definition. A **walk** W in a graph $G = (V, E, \phi)$ is a sequence $(v_0, e_1, v_1, e_2, v_2, \dots, e_n, v_n)$, where $n \in \mathbb{Z}^{\geq 0}$, each $v_i \in V$, each $e_i \in E$, and $\phi(e_i) = \{v_{i-1}, v_i\}$. We say that W **connects** v_0 and v_n and that W is a walk **from** v_0 **to** v_n .

Think of a walk as an actual walk: imagine standing at a vertex, then walking along an edge emanating from that vertex, arriving at another vertex, walking along another edge, etc.

7.4.2 Example. Some walks in the graph



are

$$(v_1), \quad (v_1, e_2, v_3), \quad (v_2, e_1, v_1, e_1, v_2), \quad (v_3, e_2, v_1, e_1, v_2, e_3, v_3).$$

7.4.3 Example. Some walks in the graph

$$e \bigcirc \overset{v}{\bullet} \bigcirc f$$

are

$$(v), \quad (v, e, v, e, v), \quad (v, f, v, e, v, f, v).$$

7.4.4 Definition. A graph $G = (V, E, \phi)$ is **connected** if for every pair of vertices $v, w \in G$, there is a walk in G from v to w .

This is just the formal way of saying that every vertex is “connected” to every other vertex by some sequence of edges.

7.4.5 Example. The empty graph with no vertices or edges is connected! This is because the “if” condition in the definition of connected is trivially satisfied.

7.4.6 Example. The graphs H_1, H_2, H_5, H_6 in the example above are connected. The graphs H_3, H_4 are not connected.

7.4.7 Exercise. Let G be a graph. Prove or disprove each of the following claims:

- (a) If G is simple, then G is connected.
- (b) If G is a complete graph, then G is connected.
- (c) If G is a cycle, then G is connected.
- (d) If G is bipartite, then G is connected.
- (e) If G is a complete bipartite graph, then G is connected.
- (f) If G is a complete bipartite graph of the form $K_{m,n}$ with $m, n \in \mathbb{Z}_{\geq 1}$, then G is connected.
- (g) G is a cycle if and only if it is connected and every vertex has degree 2.

7.4.8 Exercise. Let H_1 and H_2 be connected subgraphs of a graph G . Prove or disprove each of the following:

- (a) $H_1 \cup H_2$ is connected.
- (b) $H_1 \cap H_2$ is connected.
- (c)
- (d)

Here's a useful proposition. Think up some examples before embarking on the proof!

7.4.9 Proposition. *Let H_1, H_2 be connected subgraphs of a graph G and suppose H_1 and H_2 have a common vertex. Then $H_1 \cup H_2$ is connected.*

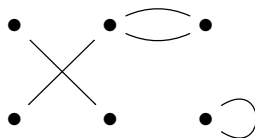
Proof. By assumption, there is a vertex $v \in V_{H_1} \cap V_{H_2}$. Let u, w be any two vertices in $H_1 \cup H_2$. If u, w are both in H_1 , then since H_1 is connected, there is a walk in H_1 from u to w , and this walk is also in $H_1 \cup H_2$ since H_1 is a subgraph of $H_1 \cup H_2$. A similar argument works when u, w are both in H_2 . The last case to consider is when u is in H_1 , while w is in H_2 . Since H_1 is connected, there is a walk $(u, e_1, v_1, \dots, e_n, v)$ in H_1 . Since H_2 is connected, there is a walk $(v, f_1, w_1, \dots, f_m, w)$ in H_2 . Combining these two walks, we get a walk

$$(u, e_1, v_1, \dots, e_n, v, f_1, w_1, \dots, f_m, w)$$

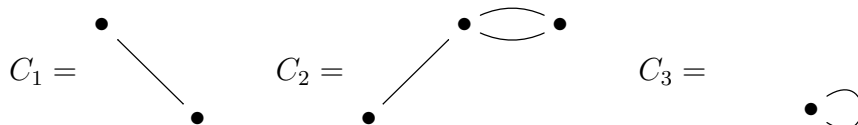
from u to w , and this walk is in $H_1 \cup H_2$ since the first part is in H_1 and the second part is in H_2 . Thus we have shown that for any two vertices u, w in $H_1 \cup H_2$, there is a walk from u to w in $H_1 \cup H_2$. So $H_1 \cup H_2$ is connected. \square

7.4.10 Definition. A **connected component** of a graph G is a maximal connected subgraph of G . (“Maximal” means not contained in any strictly larger connected subgraph.)

7.4.11 Example. The connected components of the graph



are the three subgraphs



It is easy to pick out the connected components in a drawing of a graph. Here is the algebraic description of the connected components:

7.4.12 Proposition. *Let $G = (V, E, \phi)$ be a graph. For any vertex v , the subgraph C of G defined by*

$$V_C = \{w \in V \mid \text{there is a walk in } G \text{ from } v \text{ to } w\} \quad \text{and} \quad E_C = \{e \in E \mid \phi(e) \subseteq V_C\}$$

is a connected component of G containing v .

Proof. We have to prove that (i) $v \in V_C$, (ii) C is connected, and (iii) C is maximal.

(i) Since (v) is a walk in G from v to v , $v \in V_C$.

- (ii) Suppose $u, w \in V_C$. Then there is a walk $(v, e_1, v_1, e_2, v_2, \dots, e_r, u)$ in G from v to u and a walk $(v, f_1, w_1, f_2, w_2, \dots, f_s, w)$ in G from v to w . In fact, each of these walks is also a walk in C . To see this, note that each v_i is in C because we can stop the first walk at v_i , hence obtaining a walk in G from v to v_i . Similarly, each w_i is in C . But then all the edges in the walks are also in C since their endpoints are vertices in the walk, which are in V_C . Thus both walks are in C .

Now we can construct a walk from u to w by walking from u to v (the reverse of the walk from v to u), then from v to w :

$$(u, e_r, \dots, v_1, e_1, v, f_1, w_1, \dots, f_s, w).$$

Since each of the two walks was in C , this new walk is in C . Thus C is connected.

- (iii) Let C' be a connected subgraph of G containing C . Suppose for contradiction that C' is strictly larger than C . Then C' must have a vertex not contained in C (C' cannot just have an extra edge since E_C already contains all possible edges whose endpoints are in V_C); call it w . Since C' contains C , C' contains v . Since C' is connected, there must be a walk W in C' from v to w . But W is also a walk in G , so $w \in V_C$, contradicting how we chose w .

□

Theorem 39 shows us that each vertex v of G is contained in a connected component. Even better, it gives us a recipe for describing a connected component containing v . As we will see in a moment, if we let v vary then this recipe gives us all the connected components.

7.4.13 Lemma. *Let C_1, C_2 be two connected components of a graph $G = (V, E, \phi)$. If C_1, C_2 have a common vertex, then $C_1 = C_2$.*

Proof. This is an easy application of Theorem 38. Since C_1 and C_2 are connected and share a common vertex, $C_1 \cup C_2$ is connected. Thus $C_1 \cup C_2$ is a connected subgraph of G that contains both C_1 and C_2 . By the maximality of C_1 and C_2 , the only way this can happen is if $C_1 = C_1 \cup C_2 = C_2$. □

Thus the distinct connected components of a graph are disjoint, namely they have no vertices in common. The previous theorem gives us a connected component for each vertex v , the disjointness lemma implies that there are no other connected components. Thus we have proved the important theorem:

7.4.14 Theorem. *Let $G = (V, E, \phi)$ be a graph and let C_1, \dots, C_r be the connected components of G . Then $G = C_1 \sqcup \dots \sqcup C_r$ (disjoint union). Moreover, choosing a vertex v_i in C_i , we can describe C_i by*

$$V_{C_i} = \{w \in V \mid \text{there is a walk from } v_i \text{ to } w\} \quad \text{and} \quad E_{C_i} = \{e \in E \mid \phi(e) \subseteq V_{C_i}\}.$$

Thus any graph G is the disjoint union of its connected components. Since the connected components are connected, we think of connected graphs as the “building blocks” of all graphs. Think of this theorem as some sort of analogue to the Fundamental Theorem of Arithmetic (unique factorization of any integer ≥ 2 into primes). Number theorists study primes in order to understand all integers; similarly, graph theorists study connected graphs in order to understand all graphs.

7.5 Induction on Connected Graphs

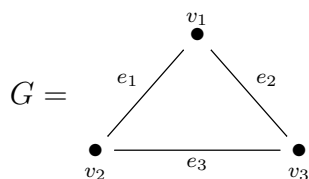
Strong induction on the size of a graph, which for us will be the number of vertices and/or edges, is a useful tool for proving facts about graphs. In order to apply it, we need a way of relating a connected graph to a smaller connected subgraph.

Every connected graph can be constructed as a sequence of subgraphs, starting with a single vertex and repeatedly using the following two operations:

- **(Op. 1)** Add a new vertex and one edge connecting the new vertex to the existing subgraph.
- **(Op. 2)** Add an edge connecting two vertices in the existing subgraph.

To convince yourself that this works, simply draw any connected graph, choose one vertex as your starting point, and start performing the two operations until you get the entire graph. In fact, we can begin by using only (Op. 1) until we have all the vertices, and then add all the leftover edges by (Op. 2).

7.5.1 Example. One way to construct the graph



in this way would be:

- Start with v_1 .
- (Op. 1) Add v_2 and e_1 .
- (Op. 1) Add v_3 and e_2 .
- (Op. 2) Add e_3 .

Here is a rigorous argument for why the construction works.

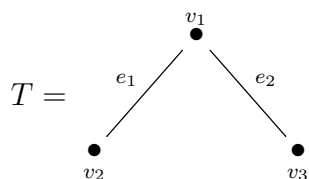
7.5.2 Lemma. *Let H be a non-empty connected subgraph of a connected graph G . Suppose $V_H \neq V_G$. Then there is a vertex $v \in V_G - V_H$ and an edge $e \in E_G - E_H$ such that the endpoints of e are v and some vertex in H . Moreover, the subgraph of G obtained by adding e and v to H is connected.*

Proof. Since H is non-empty, it has some vertex $u \in V_H$. Since $V_H \neq V_G$, there is some vertex $w \in V_G - V_H$. Since G is connected, there is a walk $(w, e_1, v_1, \dots, v_{n-1}, e_n, u)$ in G from w to u . Since the walk starts at a vertex not in V_H and ends with a vertex in V_H , there must be some edge $e = e_i$ with one endpoint $u' \in V_H$ and the other endpoint $v \in V_G - V_H$. The pair e and v have the desired properties. Let H' denote the subgraph of G with vertex set $\{v, u'\}$ and edge set $\{e\}$. Then H' is connected, and H and H' share the vertex u' , so by Theorem 38, $H \cup H'$ is connected. \square

The lemma justifies the following definition:

7.5.3 Definition. Let G be a connected graph. Starting with any one vertex of G , we can repeatedly apply (Op. 1) to get a connected subgraph T containing all the vertices of G . We call T a **spanning tree** of G .

7.5.4 Example. The spanning tree constructed for the graph G in the previous example is



The graph G actually has two other spanning trees. Can you find them?

Once we have constructed a spanning tree of G , we can construct the rest of G by repeating (Op. 2), namely by adding all the leftover edges one-by-one.

7.5.5 Note. If G is a connected graph with n vertices, then any spanning tree of G will have n vertices and $n - 1$ edges. This is because (Op. 1) must be used exactly $n - 1$ times in the construction of the spanning tree.

The key for strong induction is the following:

7.5.6 Theorem. *Let G be a connected graph with at least one edge. Then there is a connected subgraph H of G such that G is obtained from H by adding either one edge and one vertex or just one edge.*

Proof. By the above, we can construct G starting with a single vertex and repeatedly applying (Op. 1) and (Op. 2). If we stop the construction one operation before obtaining G , we get the desired subgraph H . \square

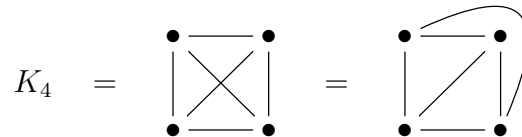
We will soon use strong induction to prove a nice result called Euler's formula.

7.6 Planar Graphs and Euler's Formula

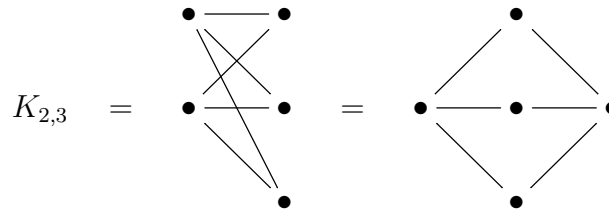
Now, for the first time, we will define a property of graphs that cares about edge crossings.

7.6.1 Definition. A graph is **planar** if it can be drawn in the plane without any edge crossings.

7.6.2 Example. The complete graph K_4 is planar (!) because we can redraw one of the diagonal edges to eliminate the edge crossing:



The bipartite graph $K_{2,3}$ is also planar, since we can move the top left vertex to the far right to eliminate all three edge crossings:



7.6.3 Note. Any subgraph of a planar graph is planar.

7.6.4 Note. Not all graphs are planar! We will eventually prove that the complete graph K_5 and the complete bipartite graph $K_{3,3}$ are not planar.

First we have to develop a bit of theory about planar graphs.

7.6.5 Definition. Let G be a planar graph, drawn in the plane without edge crossings. The edges of the graph divide the plane into **regions**, which we call the regions of G .

7.6.6 Example. The above planar drawing of K_4 divides the plane into 4 regions (counting the infinite region outside the graph). The planar drawing of $K_{2,3}$ divides the plane into 3 regions.

For us, the big theorem in the theory of planar graphs is:

7.6.7 Theorem (Euler's formula). *Let $G = (V, E, \phi)$ be a non-empty connected planar graph drawn in the plane. Then r , the number of regions of G , is*

$$r = |E| - |V| + 2.$$

Check the theorem for planar graphs like K_4 and $K_{2,3}$ before embarking on the proof!

Proof. Strong induction on $|V| + |E|$. The base case is $|V| + |E| = 1$, when G is the graph consisting of just one vertex with no edges. There is one region, so Euler's formula holds: $1 = 0 - 1 + 2$.

Now suppose that for some $k \in \mathbb{Z}^{\geq 1}$, Euler's formula holds for all connected planar graphs with $|V| + |E| \leq k$. Then let G be a connected planar graph with $|V| + |E| = k + 1$. Since G is connected, by Theorem 41 there is a connected subgraph H of G such that G is obtained from H by adding either (Case 1) one vertex and one edge or (Case 2) just one edge. Since G is planar, so is its subgraph H , so we will be able to use induction. We treat the two cases separately:

Case 1: G is obtained from H by adding one edge, by (Op. 2). Then H has v vertices and $e - 1$ edges, and $|V| + |E| - 1 = k$. Thus by the inductive assumption, we know that Euler's formula holds for H , namely

$$r_H = (|E| - 1) - |V| + 2,$$

where r_H is the number of regions of H . Now, adding one edge will split one region into two regions, so G has one more region than H . Thus

$$r = r_H + 1 = ((|E| - 1) - |V| + 2) + 1 = |E| - |V| + 2,$$

which proves Euler's formula for G .

Case 2: G is obtained from H by adding one vertex and one edge, by (Op. 1). Then H has $|V| - 1$ vertices and $|E| - 1$ edges. Since $(|V| - 1) + (|E| - 1) = k - 1 \leq k$, the inductive assumption guarantees that

$$r_H = (|E| - 1) - (|V| - 1) + 2 = |E| - |V| + 2.$$

Now, adding adding one vertex and one edge connecting that vertex to H will not create a new region, thus

$$r = r_H = |E| - |V| + 2,$$

which again proves Euler's formula.

In either case, we see that Euler's formula for H implies Euler's formula for G , so we are done by induction. \square

7.6.8 Corollary. *Let G be a connected planar graph. Then every planar drawing of G in the plane has the same number of regions.*

If we consider connected planar graphs that are also simple, then we can use Euler's formula to deduce some bounds on the maximum number of edges in terms of the number of vertices. The idea is that if a graph has “too many edges”, then you can't avoid edge crossings when you try to draw it in the plane.

7.6.9 Proposition. *Let G be a connected simple planar graph. Suppose $|V| \geq 3$. Then*

$$|E| \leq 3 \cdot |V| - 6.$$

Proof. Since there are no loops or multiple edges, every region has at least 3 edges on its boundary (with appropriate double counting). Since every edge is on the boundary of two regions, we see that

$$3r \leq 2 \cdot |E|.$$

Substituting this into Euler's formula, we see that

$$3(|E| - |V| + 2) = 3r \leq 2 \cdot |E|,$$

which simplifies to give the desired inequality

$$|E| \leq 3 \cdot |V| - 6.$$

□

This inequality is mainly useful because it allows us to deduce that some simple connected graphs are *not* planar. For instance, K_5 has 10 edges but only 5 vertices, so the inequality fails! Thus we see that

7.6.10 Corollary. *The complete graph K_5 is not planar.*

For simple graphs with no cycles of length 3 (such as bipartite graphs!) we can prove a stronger inequality.

7.6.11 Proposition. *Let G be a connected simple planar graph with no cycles of length 3. Suppose $v \geq 3$. Then*

$$e \leq 2 \cdot |V| - 4.$$

Proof. Since there are no loops, multiple edges, or cycles of length 3, every region has at least 4 edges on its boundary. Thus

$$4r \leq 2 \cdot |E|.$$

Substituting into Euler's formula, we obtain

$$2(|E| - |V| + 2) = 2r \leq |E|,$$

which simplifies to

$$|E| \leq 2 \cdot |V| - 4.$$

□

Applying this inequality to the complete bipartite graph $K_{3,3}$, which has 9 edges but only 6 vertices, we see that

7.6.12 Corollary. *The complete bipartite graph $K_{3,3}$ is not planar.*