University of Toronto Mississauga

Department of Mathematical and Computational Sciences

Course Notes

MAT102H5

Introduction to Mathematical Proofs

Shay Fuchs

December 21, 2021

Table of Contents

Preface

These notes were written with the intention of serving as the main source for the course *MAT102H5 - Introduction to Mathematical Proofs* – a first year course at the University of Toronto Mississauga, required in most mathematics, computer-science and statistics programs.

The primary goal of this course is to help students transition from high-school mathematics, where theory, proofs and precise use of language are rarely emphasized, to university-level mathematics, that requires creative thinking, problem-solving skills, and ability to communicate ideas in a coherent and precise fashion. In other words, the main goal of this course is to develop thinking, reasoning, and communication skills within the framework of mathematics. A great effort was made to include many non-technical exercises, and stay away from problems, in which a prescribed algorithm can be executed. It is my belief that one develops mathematical-thinking skills by constantly working on and being exposed to new problems, that require creative thinking, and the discovery and use of new ideas.

Nevertheless, developing mathematical-thinking skills requires some mathematical content, and so the secondary goal of the course is to present fundamental ideas, notions and results in mathematics, and use them as a vehicle to the development of problem-solving, writing, and communication skills. These fundamental topics (sets, functions, induction, equivalence relations, etc.) are important prerequisites for learning more advanced mathematics. My hope is, therefore, that these notes (and the course) will prepare the student, both in terms of skills and content knowledge, to higher-level mathematics, computer-science, and statistics courses.

Some of the topics (such as functions, inequalities and the quadratic formula) are not necessarily new to the student, but the treatment is deeper and more theoretical. Many other topics will be new to most students.

Every chapter ends with an Exercises section. Working on the exercises, and devoting a substantial amount of time and effort in attempting to solve them, is an integral part of the learning process. Throughout the course, exercises are assigned (and some are graded) on a weekly basis.

I would like to thank my colleagues Tyler Holden, Ali Mousavidehshikh, Alex Rennet and Marina

Tvalavadze, for reading through the notes, spotting several errors and misprints, and for making useful suggestions and recommendations.

Also, many thanks to Sahid Velji, a student in the Fall 2017 Semester of this course, who used to frequently send me lists of errors he found in the notes and suggestions for better wording. The notes are now much cleaner from misprints and other errors due to Sahid's helpful comments.

The notes, however, are still under review and development. Students, instructors, and mathematicians are encouraged to contact me with comments, suggestions, and any errors found in the text.

Sincerely,

Shay Fuchs (s.fuchs@utoronto.ca)

Chapter 1

Numbers, Quadratics and Inequalities

1.1 The Quadratic Formula

We begin by discussing a familiar topic - the quadratic formula. The general formula for solving an equation of the form $ax^2 + bx + c = 0$ is

$$x = \frac{-b \pm \sqrt{b^2 - 4ac}}{2a}.$$

This formula was most likely presented to you by your high school teacher, and you have learned how to use it for solving quadratic equations in various settings. However, if you have not seen a proof (or at least, some sort of explanation) of this formula, it would be hard to see where this formula comes from, and why it works. In fact, the proof of this formula is not too complicated, and only requires some algebraic manipulations. We therefore start by properly stating a theorem on quadratic equations, and then present a proof using the "completing the square" method.

Theorem 1.1.1 (The Quadratic Formula)**.** *Let a, b, c be three real numbers, with $a \neq 0$.*
Then the equation $ax^2 + bx + c = 0$ has...

(a) ***No*** *real solutions if $b^2 - 4ac < 0$.*

(b) A ***unique*** *solution if $b^2 - 4ac = 0$, given by $x = -\frac{b}{2a}$.*

(c) ***Two distinct*** *solutions if $b^2 - 4ac > 0$, given by $x = \frac{-b + \sqrt{b^2 - 4ac}}{2a}$ and $x = \frac{-b - \sqrt{b^2 - 4ac}}{2a}$.*

Note: We will explain later the term 'real numbers'. For now, a real number is any number on the number line (including whole numbers, negative numbers, fractions, etc.).

Proof. We start by rewriting the given equation, and complete the square, as follows:
Divide both sides by a (which is possible since $a \neq 0$)

$$x^2 + \frac{b}{a}x + \frac{c}{a} = 0,$$

rewrite the middle term as $2x \cdot \frac{b}{2a}$, and add and subtract the term $\frac{b^2}{4a^2}$

$$x^2 + 2x \cdot \frac{b}{2a} + \frac{b^2}{4a^2} - \frac{b^2}{4a^2} + \frac{c}{a} = 0.$$

Note that the first three terms on the left hand side form the square $\left(x + \frac{b}{2a}\right)^2$. The other two terms we move to the right hand side to get

$$\left(x + \frac{b}{2a}\right)^2 = \frac{b^2}{4a^2} - \frac{c}{a},$$

which is equivalent to

$$\left(x + \frac{b}{2a}\right)^2 = \frac{b^2 - 4ac}{4a^2}.$$

The last equation is a simpler form of the original equation, as the unknown x appears only once. It will be easier now to 'solve for x' and obtain the quadratic formula. However, we must be careful. Solving for x will involve the square root operation, and square roots cannot be applied to negative numbers. We therefore consider three separate cases:

(a) If $b^2 - 4ac < 0$, then $\frac{b^2 - 4ac}{4a^2} < 0$, and since $\left(x + \frac{b}{2a}\right)^2 \geq 0$ (for any real number x), we conclude that there are **no real solutions** in this case.

(b) If $b^2 - 4ac = 0$, then we get

$$\left(x + \frac{b}{2a}\right)^2 = 0 \qquad \Rightarrow \qquad x + \frac{b}{2a} = 0 \qquad \Rightarrow \qquad x = -\frac{b}{2a},$$

which gives, in this case, the **unique solution** to the equation.

(c) If $b^2 - 4ac > 0$, we use square roots and get

$$x + \frac{b}{2a} = \pm\sqrt{\frac{b^2 - 4ac}{4a^2}} \qquad \Rightarrow \qquad x = -\frac{b}{2a} \pm \frac{\sqrt{b^2 - 4ac}}{2a} = \frac{-b \pm \sqrt{b^2 - 4ac}}{2a}.$$

We see that there are **two distinct solutions** in this case, and we have obtained the famous quadratic formula, as needed.

□

This was our first proof in the course. Take a closer look at it! What features can you identify in a mathematical proof? Here are a few important observations.

- Our proof included quite a few words and sentences (in natural language), and not only mathematical symbols (such as equations, numbers and formulas). This will happen in most mathematical proofs. A mathematical argument should be made of **complete sentences**, containing words, symbols, or a combination of both. The words are meant to help the reader follow the logical flow of the argument, and to connect the various statements that appear in the proof. Words such as "if... then...", "and" and "or" (also called **connectives**) often appear in mathematical arguments, and should be used properly. In Chapter 3 we will discuss in detail the meaning of these connectives in mathematics, and how to use them in mathematical proofs.

- At the end of the proof, we placed the symbol □. This is a common way to denote the end of a mathematical proof (or, more generally, the end of an argument). In other books you might see the symbol ■ or the acronym **Q.E.D.** used instead (which comes from the Latin phrase **Quod Erat Demonstrandum**, meaning "which is what had to be shown"). However, in this book, we will keep using our square □.

- A mathematical argument is normally based on facts that have been previously validated, or agreed upon. For example, in the proof of Theorem 1.1.1, we used the identity $(x+y)^2 = x^2 + 2xy + y^2$, which is valid for any two real numbers x and y. Should we have also proved this formula? Well, we could, but we assumed that this formula was well established prior to proving the theorem, and so there was no need to re-explain (or prove) it again.

  This sort of judgment needs to be done each time a mathematical argument is presented to an audience, and you will need to ask yourself: Which facts should be well known to the reader? What other theorems or claims may I refer to in my proof? What are the main steps (or ideas) in the argument? What is the main tool (or tools) used in my proof? Should I mention them explicitly?

  With time and practice, you will develop your own style of writing mathematical proofs. The feedback you will get from the course staff (and even from your classmates), will help you improve your writing and polish your arguments.

There are several reasons why proofs are important. First, a proof validates the truth of a general statement. Once a theorem is proved, it remains true forever (unless an error is found). For instance, Theorem 1.1.1 implies that a quadratic equation can never have three distinct solutions, no matter how hard you try to find one, or how much time you spend searching. This is the strength of a proof. We can say now, without a doubt, that any given quadratic equation must have zero, one or two real solutions, and there are no other options or exceptions.

Secondly, a proof often gives us an insight as to **why** the theorem is valid, and may suggest strategies for proving similar statements. For instance, can we prove a similar theorem about cubic equations?

The Exercises at the end of the chapter cover a few applications of the quadratic formula, and some additional properties of quadratics (for which a proof will be requested).

1.2 Inequalities, and Arithmetic/Geometric Means

In your high school years, you have spent a substantial amount of time on equations: You had to re-arrange, simplify, and solve equations regularly. However, working with inequalities can be more challenging, and one has to be much more careful when an argument involves inequalities.

Consider, as an example, the equation $\frac{1}{x} = x$. Solving this equation is quite simple. All we have to do is multiply both sides by x, and then the equation becomes $1 = x^2$. Its solutions are, of course, $x = 1$ and $x = -1$.

But now how about solving the inequality $\frac{1}{x} > x$? We cannot just multiply by x as before, since if x is negative, we need to flip the inequality sign. We need to consider two possible cases: If $x > 0$, then multiplying by x gives $1 > x^2$, and the **positive** x's satisfying this condition are those between 0 and 1, namely, $0 < x < 1$. However, if $x < 0$, multiplication by x yields $1 < x^2$, and the **negative** x's satisfying the latter inequality are those which are smaller than -1, namely, $x < -1$.

To summarize, the solutions to $\frac{1}{x} > x$ are the numbers between 0 and 1, and those that are smaller than -1. We may write

$$\frac{1}{x} > x \qquad \text{if and only if} \qquad x < -1 \quad \text{or} \quad 0 < x < 1 \ .$$

(The words "if and only if" indicate, in mathematics, a two-sided implication: If x solves the inequality, then it must satisfy the condition "$x < -1$ or $0 < x < 1$", and if this condition is satisfied, then x solves the inequality.)

This example shows some of the complications that may arise while working with inequalities, and how important it is to be able to manipulate them properly. Here are a few familiar facts about numbers and inequalities, that we will (temporarily) call "Basic Facts":

Basic Facts. Let a, b and c be three real numbers.

(1) If $a < b$ (or $a \leq b$) and $c > 0$, then $ca < cb$ (or $ca \leq cb$).

(2) $a^2 \geq 0$.

(3) If $a \geq 0$, then there is a unique nonnegative number $\sqrt{a}$, whose square is a.
(A nonnegative number is a number that is positive or zero.)

(4) If $a < b$ and $b < c$, then $a < c$.

Most likely, you have seen these facts before and used them multiple times, and so we will take them for granted (even though they can be proved). However, we can now use these facts to prove other results, that may be less known, or not as intuitive as the basic facts.

Proposition 1.2.1. Let a and b be two real numbers.

(a) If $0 < a < b$, then $a^2 < b^2$ and $\sqrt{a} < \sqrt{b}$.

(b) Similarly, if $0 < a \leq b$, then $a^2 \leq b^2$ and $\sqrt{a} \leq \sqrt{b}$.

Proof. To prove (a), we assume that $0 < a < b$. Since $a < b$ and $a > 0$, we can use basic fact (1) (with $c = a$) to get $a^2 < ab$. Similarly, as $b > 0$, we can use basic fact (1) again to get $ab < b^2$. Now, from $a^2 < ab$ and $ab < b^2$, we get $a^2 < b^2$ (here, basic fact (4) is used).

To prove the second inequality, re-arrange $a < b$ as follows:

$$a < b \qquad \Rightarrow \qquad b - a > 0 \qquad \Rightarrow \qquad (\sqrt{b} + \sqrt{a})(\sqrt{b} - \sqrt{a}) > 0.$$

We used the well known difference of squares formula $x^2 - y^2 = (x + y)(x - y)$.

We now multiply both sides, using basic fact (1), by $c = \dfrac{1}{\sqrt{a} + \sqrt{b}}$ (which is a positive number), to get $\sqrt{b} - \sqrt{a} > 0$, or $\sqrt{a} < \sqrt{b}$, as needed. This concludes the proof of part (a). The proof of part (b) is very similar, and is left as an exercise. □

Note again how the proof contained **words**, and that **complete sentences** were used. A phrase like

$$a < b \qquad a^2 < ab \qquad ab < b^2 \qquad a^2 < b^2$$

cannot be considered as a proof of $a^2 < b^2$ (even though it does contain its key steps). Without words and complete sentences, it would be very hard for the reader to follow the argument, and the logic that was used. The reader may conclude that the argument is incomplete, unclear, or even flawed.

We now proceed to discuss two important and fundamental inequalities in mathematics: The Arithmetic-Geometric Mean Inequality, and the Triangle Inequality. These inequalities will be useful later on, but

even more importantly, the techniques and approaches we use to prove them will turn out to be valuable tools for proving other inequalities (and for constructing proofs, in general).

We begin with the following definition.

Definition 1.2.2. The **arithmetic mean** of two real numbers, x and y, is $\dfrac{x+y}{2}$.
If $x, y \geq 0$, then their **geometric mean** is $\sqrt{x \cdot y}$.

You have probably encountered the arithmetic mean before (also referred to as **the average** of x and y). The geometric mean is another type of 'average', that often shows up naturally in calculations and applications.

Examples. (a) The arithmetic mean of 2 and 8 is $\dfrac{2+8}{2} = 5$, and their geometric mean is $\sqrt{2 \cdot 8} = 4$. The arithmetic mean of -10 and 7 is -1.5, and their geometric mean is undefined.

(b) Imagine a bank, offering a savings account, that pays interest once a year as follows. One year after the opening date, 10% of the current amount is deposited into your account. Then, at the end of every subsequent year, an interest of 20% is applied.

For instance, if the initial investment is \$250, then after one year, this amount grows to $\$250 \cdot 1.1 = \$275$, and after two years, to $\$275 \cdot 1.2 = \$330$.

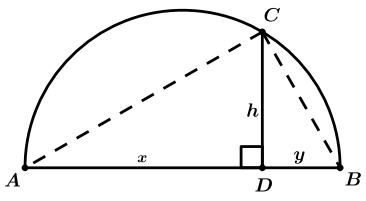
In general, if the initial investment is x, then after two years, the investment grows to $x \cdot 1.1 \cdot 1.2 = 1.32x$.

What would be a sensible way to define an **"average" growth rate** for the first two-year period? We might want to look for a hypothetical fixed rate r that would lead to the same final amount. Thus, we want r to satisfy the condition $x \cdot r \cdot r = x \cdot 1.1 \cdot 1.2$ (for every value of x). We get

$$r^2 = 1.1 \cdot 1.2 \qquad \Rightarrow \qquad r = \sqrt{1.1 \cdot 1.2} \qquad \Rightarrow \qquad r = \sqrt{1.32} \approx 1.1489.$$

We conclude that the average interest rate is about 14.89% (and **not** 15%). Note that the number 1.1489 is the **geometric mean** of 1.1 and 1.2.

(c) In the diagram, AB is a diameter of a circle, and CD is perpendicular to AB.
If x, y, h are the lengths of AD, BD, and CD, respectively,
how can we express h in terms of x and y?



Solution:

One way to proceed, is to observe that the triangles ADC, CDB and ACB are all right triangles (remember that inscribed angles subtended by a diameter are right). Therefore, we can apply the Pythagorean Theorem to get

$$AC^2 = AD^2 + DC^2 \quad , \quad CB^2 = CD^2 + DB^2 \quad \text{and} \quad AB^2 = AC^2 + CB^2.$$

We now use the first two equalities to replace AC^2 and CB^2 by $AD^2 + DC^2$ and $CD^2 + DB^2$ in the third equality:

$$AB^2 = (AD^2 + DC^2) + (CD^2 + DB^2) = AD^2 + 2CD^2 + DB^2.$$

Expressing all quantities in terms of x, y and h leads to

$$(x+y)^2 = x^2 + 2h^2 + y^2 \quad \Rightarrow \quad x^2 + 2xy + y^2 = x^2 + 2h^2 + y^2 \quad \Rightarrow \quad h^2 = xy$$

which gives $h = \sqrt{xy}$. In other words, h is the geometric mean of x and y.

(**Note:** Another way to obtain the same formula, is to use similar triangles. Can you figure out the details?)

In Examples (a) and (b) above, we see that the geometric mean is **smaller** than the arithmetic mean. It turns out that this will always be the case. This is, essentially, the content of the famous **Arithmetic-Geometric Mean (or AGM) Inequality**, stated below.

Proposition 1.2.3 (The Arithmetic-Geometric Mean Inequality)**.**
For any two real numbers x and y, $x \cdot y \leq \left(\frac{x+y}{2}\right)^2$, and equality holds iff (if and only if) $x = y$.
If, in addition, $x, y \geq 0$, then $\sqrt{xy} \leq \frac{x+y}{2}$.

Before presenting the proof, let us make sure that we fully understand the statement, and what needs to be proved. Keeping in mind that iff (if and only if) means a double-sided implication, we see that there are three statements included in the first sentence:

(1) $x \cdot y \leq \left(\frac{x+y}{2}\right)^2$ for any x and y.

(2) If $x = y$, then $x \cdot y = \left(\frac{x+y}{2}\right)^2$.

(3) If $x \cdot y = \left(\frac{x+y}{2}\right)^2$, then $x = y$.

The second sentence in Proposition 1.2.3 also needs to be proved, but this will easily follow by applying square roots to the inequality $x \cdot y \leq \left(\frac{x+y}{2}\right)^2$.

Let us start by focusing on (1). How would one prove such an inequality (for all x and y)? Clearly, we cannot substitute numbers for x and y, as our argument must be completely general. But as a start, we can try to re-write the given inequality, with the hope of simplifying it to an inequality that would be easier to prove. We call this kind of work "rough work", since we are not writing an actual proof (yet), but only doing preliminary experimentation to try and **discover** a proof.

Rough Work:

$$x \cdot y \leq \left(\frac{x+y}{2}\right)^2 \quad \Rightarrow \quad x \cdot y \leq \frac{x^2+2xy+y^2}{4} \quad \Rightarrow \quad 4xy \leq x^2+2xy+y^2$$

$$\Rightarrow \quad 0 \leq x^2-2xy+y^2 \quad \Rightarrow \quad 0 \leq (x-y)^2.$$

Can this rough work considered a proof? No. First, there are no words and full sentences explaining the argument. More importantly, a proof cannot begin with the statement that needs to be proved. Remember - we cannot assume the validity of the inequality $x \cdot y \leq \left(\frac{x+y}{2}\right)^2$. Our task is to provide an argument (i.e., a proof) **validating** this inequality. We may only use facts that are known to be true (such as simple high school algebra).

However, we did achieve something. Using algebraic manipulations, we were able to obtain a simpler inequality, $0 \leq (x-y)^2$, which holds true for any x and y, by Basic Fact (2). We might be able to use it as our starting point, and work backwards in the rough work. If we manage to reverse all the steps, we will end up with the desired inequality, and that would be our proof.

We are now ready to prove the proposition.

Proof (of Proposition 1.2.3). (1) For any two real numbers, x and y, we have $0 \leq (x-y)^2$ (as $a^2 \geq 0$ for any real number a). We expand and re-arrange this inequality, as follows:

$$0 \leq (x-y)^2 \quad \Rightarrow \quad 0 \leq x^2-2xy+y^2 \quad \Rightarrow \quad 4xy \leq x^2+2xy+y^2$$

(in the second step, we added $4xy$ to both sides). We divide by 4, and get

$$xy \leq \frac{x^2+2xy+y^2}{4} \quad \Rightarrow \quad xy \leq \left(\frac{x+y}{2}\right)^2,$$

as needed.

(2) For proving (2), we do not need rough work, as it is quite clear what needs to be done. We assume that $x = y$, and show that the inequality (in (1)), becomes an equality. Well, if $x = y$, then the

left-hand side becomes $x \cdot y = x^2$, and the right-hand side –

$$\left(\frac{x+y}{2}\right)^2 = \left(\frac{2x}{2}\right)^2 = x^2.$$

This proves that when $x = y$, we have equality, as needed.

(3) For this part, we assume that $x \cdot y = \left(\frac{x+y}{2}\right)^2$, and prove that $x = y$. We do that by simplifying the equality as much as we can, until it can be easily seen that x and y must be equal. The steps we follow resemble those from the proof of (1).

$$xy = \left(\frac{x+y}{2}\right)^2 \quad \Rightarrow \quad xy = \frac{x^2+2xy+y^2}{4} \quad \Rightarrow \quad 4xy = x^2+2xy+y^2$$

$$\Rightarrow \quad 0 = x^2 - 2xy + y^2 \quad \Rightarrow \quad 0 = (x-y)^2 \quad \Rightarrow \quad x - y = 0.$$

We conclude that $x = y$, as needed.

To prove the second sentence of Proposition 1.2.3, we assume that $x, y \geq 0$, and apply square roots to $x \cdot y \leq \left(\frac{x+y}{2}\right)^2$ (Proposition 1.2.1 is used here). □

Inequalities are used all over mathematics, and have many important applications. Here is an example.

Example 1.2.4. If a and b are two real numbers, satisfying $a + 2b = 50$, what is the maximum (i.e., the largest possible) value of $a \cdot b$?

Solution:

This problem can be solved using Calculus, but we show an alternate solution, relying on Proposition 1.2.3. By the AGM inequality, with $x = a$ and $y = 2b$, we see that

$$a \cdot 2b \leq \left(\frac{a+2b}{2}\right)^2 \quad \Rightarrow \quad 2ab \leq \left(\frac{50}{2}\right)^2 \quad \Rightarrow \quad ab \leq \frac{25^2}{2} = 312.5,$$

which means that the product $a \cdot b$ cannot exceed 312.5. To show that 312.5 is, in fact, the maximal value of the product, we need to show that $a \cdot b = 312.5$ for **some** numbers a and b, satisfying $a + 2b = 50$.

We know, from Proposition 1.2.3, that the AGM inequality becomes an equality when $x = y$. This translates, in our case, to $a = 2b$. Together with the condition $a + 2b = 50$, we get $a + a = 50$, or $a = 25$, and hence $b = 12.5$. Indeed, when $a = 25$ and $b = 12.5$, we have $a + 2b = 50$, and $ab = 312.5$. This proves that 312.5 is the largest possible value of ab, as needed.

1.3 The Triangle Inequality

In this section we present, and prove, another fundamental inequality in mathematics - The Triangle Inequality. We start by reviewing absolute values and their properties.

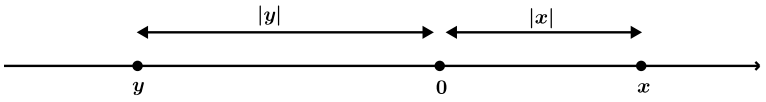
Definition 1.3.1. The absolute value of a real number x, denoted as $|x|$, is defined as

$$|x| = \begin{cases} x & \text{if } x \geq 0 \\ -x & \text{if } x < 0 \end{cases} .$$

For instance,

$$|-5| = 5 \quad , \quad |0| = 0 \quad , \quad \left|\frac{4}{7}\right| = \frac{4}{7} \quad , \quad |-\pi| = \pi.$$

Geometrically, one can interpret absolute values, as measuring distances from zero on the number line.



The following proposition lists a few basic properties of absolute values.

Proposition 1.3.2. For any two real numbers x, y, we have

$$\sqrt{x^2} = |x| \quad , \quad |x|^2 = x^2 \quad , \quad x \leq |x| \quad \text{and} \quad |x \cdot y| = |x| \cdot |y|.$$

The proof follows directly from Definition 1.3.1, and is left as an exercise.

The last equality in the proposition, $|x \cdot y| = |x| \cdot |y|$, shows that products are preserved by absolute values. Namely, the absolute value of a product of two numbers equals the product of their absolute values. This property does not hold for sums (can you find an example?). However, there is still a relation between the absolute value of a sum, and the sum of absolute values. This is the well-known triangle inequality, stated below.

Proposition 1.3.3 (The Triangle Inequality)**.** For any two real numbers x and y, we have

$$|x + y| \leq |x| + |y|.$$

In other words, the absolute value of a sum is always less than or equal to the sum of the absolute values.

You might wonder why this inequality is called "The Triangle Inequality", as there does not seem to be any triangles involved here. The version stated above is the **one-dimensional version** of the triangle inequality (concerning numbers on the number line), but the triangle inequality can be generalized to higher dimensions. In two dimensions, for instance, the triangle inequality states that the sum of the lengths of any two sides of a triangle must be greater than the length of the remaining side. However, we will restrict ourselves to the one-dimensional version.

How can we go about proving the triangle inequality? There are a few possibilities (such as looking at cases, according to the sign of x, y and $x+y$). We follow an approach similar to the one we used for the AGM inequality, and begin with some rough work.

Rough Work:

$$|x+y| \leq |x|+|y| \quad \Rightarrow \quad |x+y|^2 \leq (|x|+|y|)^2 \quad \Rightarrow \quad (x+y)^2 \leq |x|^2+2|x||y|+|y|^2$$

$$\Rightarrow \quad x^2+2xy+y^2 \leq x^2+2|xy|+y^2 \quad \Rightarrow \quad xy \leq |xy|.$$

Note that we relied on Proposition 1.3.2, and that the last inequality holds true. This will be the starting point of our proof.

Proof (of Proposition 1.3.3). For any real number, we have $a \leq |a|$, and so for any x, y we have $xy \leq |xy|$. Using basic algebra and Proposition 1.3.2, we have

$$xy \leq |xy| \quad \Rightarrow \quad x^2+2xy+y^2 \leq x^2+2|xy|+y^2 \quad \Rightarrow \quad x^2+2xy+y^2 \leq |x|^2+2|x||y|+|y|^2$$

$$\Rightarrow \quad (x+y)^2 \leq (|x|+|y|)^2 \quad \Rightarrow \quad |x+y| \leq |x|+|y|,$$

as needed. Note that in the last step, Proposition 1.2.1 was used. □

The following example shows how the triangle inequality can be used in "bounding arguments".

Example. Find a number M, such that $|x^5-2x-5| \leq M$ whenever $|x| \leq 2$.

Solution:

The key to finding such an M is to relate the absolute value of x^5-2x-5 to the absolute value of x. This can be done using the triangle inequality. For any number x, we have

$$|x^5-2x-5| = |x^5+(-2x)+(-5)| \leq |x^5|+|-2x|+|-5| \leq |x|^5+2|x|+5.$$

Using the fact that $|x| \leq 2$, we get

$$|x|^5 + 2|x| + 5 \leq 2^5 + 2 \cdot 2 + 5 = 32 + 4 + 5 = 41.$$

This shows that $M = 41$ satisfies the required condition.

(Note: Other numbers work too. For instance, any number greater than 41 can be taken as our M.)

1.4 Types of Numbers

In this section we review, informally, the basic types of numbers used in mathematics, and introduce relevant terminology. The first type of numbers one normally encounters as a child are the **Natural Numbers** $1, 2, 3, 4, 5, \ldots$. These are also often called the **Positive Integers**, or the **Counting Numbers** (as we use them to count objects or people on a regular basis). Note that in some books, 0 is considered a natural number as well. However, in this course, we employ the convention that 0 is **not** a natural number (and so the smallest natural number for us is 1).

What can we do with natural numbers? There are several **operations** we can use with natural numbers to perform various computations. We can add and multiply natural numbers to produce new natural numbers: $3 + 5 = 8$, $7 \cdot 3 = 21$, etc. We can take powers of natural numbers: $2^3 = 8$, $5^4 = 625$, etc. We can also subtract and divide natural numbers, for instance: $13 - 9 = 4$ and $42 : 7 = 6$, but not every such computation will lead to an answer which is a natural number. To be able to calculate $12 - 36$ or $18 : 4$, we need to extend our number system to include other type of numbers, such as negative numbers and fractions.

However, before doing so, there is one more thing worth mentioning about the natural numbers - we can **order** them. We all know that 3 is smaller than 7, and that 32 is greater than 19, and we even have symbols to denote these facts: $3 < 7$ and $32 > 19$. Whenever we are given two natural numbers, they can be either equal to each other, or one is greater than the other. We also use the symbols $\leq$ and $\geq$ for 'less than or equal to' and 'greater than or equal to'. For instance, $16 \leq 16$, $9 \geq 7$ and $11 \leq 111$ are all correct statements.

The **integer numbers** are obtained by joining the number 0 (zero), and the negative numbers $-1, -2, -3, \ldots$ to the natural numbers. The integers are thus the numbers $\ldots, -3, -2, -1, 0, 1, 2, \ldots$. They form an extension of the natural numbers, and are also ordered (for instance, $-15 < -9$, $4 \geq -4$, etc.). Every time we add, subtract or multiply two integers, the result will be also an integer (e.g., $(-2) \cdot (-6) = 12$, $(-8) - 3 = -11$, $5 + (-4) = 1$, etc.), but that is not the case for division. The result of a division problem (with integers) may or may not be an integer: The answer to $(-24) : 4$ is in an integer, while the answer to $32 : (-5)$ is not. This leads to the following definition.

Definition 1.4.1. Let a be an integer, and b a **nonzero** integer.
We say that a is **divisible** by b (or that b **divides** a), if there exists an integer m, for which $a = m \cdot b$.
(In other words, a is divisible by b, if it is an integer multiple of b.)

Note that the definition above does not use (nor refers to) the division operation, and that the notion of "fraction" is not mentioned. Divisibility is defined by referring to multiplication only. This means that a person, who has never learned about fractions, and has no idea what is 'a half' or 'a third', should be able to decide whether an integer is divisible (or not) by another integer.

Examples. (a) 15 is divisible by 3, since $15 = 5 \cdot 3$ (here $a = 15$, $b = 3$ and $m = 5$).

(b) -7 is divisible by -1, since $-7 = 7 \cdot (-1)$ (here $a = -7$, $b = -1$ and $m = 7$).

(c) 0 is divisible by 13, since $0 = 0 \cdot 13$.

(d) Any integer a is divisible by 1, since $a = a \cdot 1$.

(e) 15 is not divisible by 4, since 15 is not a multiple of 4.

Now, we can use divisibility to define a few related notions.

Definition 1.4.2. (a) An integer is **even** if it is divisible by 2. Otherwise, it is **odd**.

(b) A natural number $p > 1$ is called a **prime number**, if the only natural numbers that divide p are 1 and p.

As we will see later on, prime numbers are, in some sense, the building blocks of the integers. The first few prime numbers are $2, 3, 5, 7, 11, 13, 17, 19, 23, 29, \ldots$. We will prove later that there are infinitely many of them. The first prime number, 2, is the only even prime number (why?).

Next we introduce fractions, or, using more accurate terminology, the **Rational Numbers**. A rational number is a number (on the number line), that can be represented as a quotient of two integers. In other words, it is a number of the form $\frac{a}{b}$, where a and b are integers, and b is nonzero.

We have all learned about fractions in elementary school. Numbers such as $\frac{1}{2}$, $\frac{2}{3}$ and $\frac{17}{32}$ are examples of rational numbers. Improper fractions like $\frac{18}{10}$ and $\frac{19}{-5}$ are also rational numbers. Moreover, every number that can be written as a quotient of two integers is a rational number. For instance, 4, $5\frac{6}{7}$, 0.12 and $-\frac{3}{2}$ are rational numbers, as they can be written as quotients of integers: $\frac{4}{1}$, $\frac{41}{7}$, $\frac{12}{100}$ and $\frac{-3}{2}$.

We conclude that every integer is a rational number (why?), and so the rational numbers form an extension of the integers. The rational numbers are also ordered (e.g., $\frac{2}{3}$ is larger than $\frac{7}{12}$), and we can add, subtract, multiply, and divide pairs of rational numbers (except that we cannot divide by zero!).

Now that we have the rational numbers, or fractions, at our disposal, can we use them to describe any number (on the number line)? In other words, can we represent any (real) number as a quotient of two integers? The answer, which was a surprising revelation to the ancient Greek mathematicians (around the 5th century BC), is **no**. There are numbers which cannot be represented as quotients of integers. In fact, as we will see in Chapter 5, there are many such numbers, and we call them the **irrational numbers**. Examples of irrational numbers are $\sqrt{7}$, π and $\log_2 3$. These numbers cannot be expressed in the form $\frac{a}{b}$, with a and b being integers (and b nonzero). How do we know that? Well, this requires a proof. We will see how to prove irrationality of some numbers in Chapters 3 and 6. In some cases, proving irrationality can be difficult. For instance, to show that π is irrational, we need some tools from calculus (and a precise definition of π), and so we leave this proof for another course.

By joining the irrational numbers to the rational numbers, we get a larger set of numbers, called the **real numbers**. We will discuss a more precise way of defining the set of real numbers later, but for now, let us just say that the real numbers are all the numbers on the number line - rationals and irrationals. Real numbers are ordered (e.g., $-2 < \pi$ and $\sqrt{2} < 1.5$), and various algebraic operations can be performed with them.

1.5 Exercises for Chapter 1

1.5.1. (a) Show, that if x_1 and x_2 are two solutions of a quadratic equation $ax^2 + bx + c = 0$ (with $a \neq 0$), then $x_1 + x_2 = -\frac{b}{a}$ and $x_1 \cdot x_2 = \frac{c}{a}$ (these are often called **Vieta's Formulas**). Also, find a formula for $x_1^2 + x_2^2$ in terms of a, b and c.

(b) Use Part (a) to find a quadratic equation with two distinct real solutions, given that the sum of the solutions is 47 and their product -59 .

1.5.2. What are the dimensions of a rectangle with perimeter 12 and diagonal $\sqrt{20}$?

1.5.3. (a) Find all the real solutions of the equation $x^2+x+1=0$, if there are any.

(b) Alex, a MAT102 student, presented the following argument in his solution to part (a):

Clearly, $x\neq 0$ (as $0^2+0+1\neq 0$), so we can divide by x to get $x+1+\frac{1}{x}=0 \quad\Rightarrow\quad x=-1-\frac{1}{x}$.

Now, replace the term 'x' in the original equation with $-1-\frac{1}{x}$, to get:

$$x^2+\left(-1-\frac{1}{x}\right)+1=0 \qquad \Rightarrow \qquad x^2-\frac{1}{x}=0 \qquad \Rightarrow \qquad x^3=1 \qquad \Rightarrow \qquad x=1$$

and so the solution of the equation is $x=1$.

Is Alex right? If not, where does the mistake occur? Explain.

1.5.4. For which positive numbers b does there exist a rectangle with perimeter $2b$ and area $\frac{b}{2}$?

1.5.5. True or False? Explain your answer briefly.

(a) For any real number c, the quadratic equation $x^2+x-c^2=0$ has two distinct (real) solutions.

(b) If $a>4$, then the equation $ax^2+4x+1=0$ has no (real) solutions.

(c) If $b^2-4ac\geq 0$, then the quadratic equation $ax^2+bx+c=0$ has at most one solution.

1.5.6. (a) Solve the inequality $\frac{2}{x}>3x$.

(b) Solve the equation $x^3=x$ and the inequality $x^3>x$.

1.5.7. Prove part (b) of Proposition 1.2.1.

1.5.8. Show that the statement "If $a<b$, then $a^2<b^2$" is not valid in general (this is one of the reasons for requiring that a and b are positive in Proposition 1.2.1).

1.5.9. Let a,b be two **positive** numbers.
Decide whether the given statement is true or false. Give a proof or a counterexample.

(a) If $a+b\leq\frac{1}{2}$, then $\frac{1-a}{a}\cdot\frac{1-b}{b}\geq 1$.

(b) If $\frac{1-a}{a}\cdot\frac{1-b}{b}\geq 1$, then $a+b\leq\frac{1}{2}$.

1.5.10. Let $a,b>0$. Prove that $\frac{2}{\frac{1}{a}+\frac{1}{b}}\leq\sqrt{ab}$.

1.5.11. Use the AGM inequality (Proposition 1.2.3) to find the maximum of $\left(5+\sqrt{x^4+1}\right)\cdot\left(9-\sqrt{x^4+1}\right)$.
Do not use calculus.

1.5.12. Let x,y,u,v be real numbers.

(a) Prove that $4xyuv\leq 2x^2y^2+2u^2v^2$.

(b) Prove that $(xu+yv)^2 \leq (x^2+y^2)(u^2+v^2)$.

1.5.13. Prove that for any two real numbers x, y, with $x \neq 0$, we have $2y \leq \frac{y^2}{x^2} + x^2$.

1.5.14. Show that for any two real numbers x and y, we have $2xy \leq \frac{2}{3} \cdot x^2 + \frac{3}{2} \cdot y^2$.

1.5.15. Prove that if $x > y > z$, then $xy + yz > \frac{(x+y)(y+z)}{2}$.

1.5.16. Let x, y, z be nonnegative real numbers such that $x + z \leq 2$. Prove that $(x - 2y + z)^2 \geq 4xz - 8y$. Determine when equality holds.

1.5.17. Show that for any two real numbers a, b and $\varepsilon > 0$, we have $ab \leq \frac{a^2}{2\varepsilon} + \frac{\varepsilon b^2}{2}$.

1.5.18. Show that for any two **positive** real numbers a, b: $\frac{a}{a+2b} + \frac{b}{b+2a} \geq \frac{1}{2}$.

1.5.19. (a) Prove that $\sqrt[4]{xyzw} \leq \frac{x+y+z+w}{4}$ for any $x, y, z, w \geq 0$ (this is the AGM inequality for four numbers).

Hint: Write $\sqrt[4]{xyzw}$ as $\sqrt{\sqrt{xy} \cdot \sqrt{zw}}$ and use Proposition 1.2.3 multiple times.

(b) Prove the AGM inequality for three numbers, $\sqrt[3]{xyz} \leq \frac{x+y+z}{3}$ (where $x, y, z \geq 0$), by using part (a) with $w = (xyz)^{1/3}$.

1.5.20. Prove Proposition 1.3.2.

1.5.21. When does equality hold in the triangle inequality?

State your answer using the words "if and only if", and justify it.

1.5.22. Prove the following inequality for any x and y: $\quad ||x| - |y|| \leq |x - y|$.

This is an alternate version of the triangle inequality.

1.5.23. Prove that for any $x, y \geq 0$, we have $\left|\sqrt{x} - \sqrt{y}\right| \leq \sqrt{|x-y|}$.

(Hint: Treat the cases $x \geq y$ and $x \leq y$ separately.)

1.5.24. Find a number M, such that $|x^3 - 4x^2 + x + 1| < M$ for all $1 < x < 3$.

Do not use calculus!

1.5.25. If $1 < x < 2$ find a bound for $\left|\frac{x^3 + x^2 - 1}{x - 6}\right|$.

(i.e., find a number M, such that $\left|\frac{x^3 + x^2 - 1}{x - 6}\right| < M$ for all $1 < x < 2$.)

1.5.26. Let a, b, c be three real numbers. Prove that $|a - c| \leq |a - b| + |b - c|$.

1.5.27. Answer the following questions. Explain your answer briefly.

(a) Is -117 divisible by -13 ?

(b) Is 3 divisible by 9 ?

(c) Find all the integers that divide -20 .

(d) If k is a nonzero integer, does it divide $(k+1)^2 - 1$?

1.5.28. In Definition 1.4.1, what might be the reason for requiring that b is a **nonzero** integer?

1.5.29. True or False? Explain your decision briefly.

(a) Every nonzero integer divides zero.

(b) The only integers that divide 17 are 1 and 17.

(c) Let a and b be two nonzero integers. If a divides b and b divides a, then $a = b$.

(d) For any integer k, the quantity $k^2 + k$ is divisible by 2.

1.5.30. (a) What are the rational solutions to the equation $(x^2 - 1)(x^2 - 7) = 0$?

(b) Why is the number $1.111\ldots$ a rational number?

(c) Show that $\dfrac{\sqrt{3}+\sqrt{2}}{\sqrt{3}-\sqrt{2}} - 2\sqrt{6}$ is a rational number.

1.5.31. The following statements are **False**. Show that by finding a counterexample for each statement.

(a) The square root of any rational number is irrational.

(b) The sum of two irrational numbers is irrational.

(c) The product of an integer and an irrational number is irrational.

(d) The quotient of two irrational numbers cannot be a natural number.

(e) For any two real numbers a and b, with b nonzero, the quotient $\dfrac{a}{b}$ is a rational number.

Chapter 2

Sets, Functions and the Field Axioms

2.1 Sets

Sets are often considered as one of the most fundamental objects in mathematics. With some effort, nearly all of mathematics can be formulated in terms of sets. Precise definitions for functions, natural numbers, pairs of numbers, and other mathematical objects can be entirely formulated in terms of sets, and so it is essential that we understand sets, and know how to work with them.

A formal definition of a set in mathematics is well beyond the scope of these notes. Instead, we employ a naive approach, and define a set, informally, as a **collection of distinct objects**. The objects may be referred to as the **elements**, or **members** of the set.

Notation and Terminology.

A set will be normally labeled by an **upper-case Roman letter** (such as A, B, C, S, T, etc.). We use **braces** to explicitly list the elements of a set. For instance, if A is a set whose elements are the numbers 1, 7, -1 and 5, and B is a set whose elements are the word 'apple', the letter n, and the number $\frac{1}{2}$, we can write

$$A = \{1, 7, -1, 5\} \qquad \text{and} \qquad B = \left\{\text{apple}, n, \frac{1}{2}\right\}.$$

The symbol $\in$ is used to indicate that a particular object belongs to a set. The notation $y \in C$ reads "y is an element (or a member) of the set C". For example, we can write $-1 \in A$ and $n \in B$ to indicate that the number -1 belongs to the set A, and that n is a member of the set B (defined above). To indicate that a certain object is **not** an element of a set, we use the symbol $\notin$. For instance, $3 \notin A$ means that 3 is **not** an element of the set A.

When all the elements of a set C are also members of another set D, we say that C is a **subset** of D,

and denote this fact as $C \subseteq D$ (or as $D \supseteq C$). Similarly, $C \not\subseteq D$ means that C is **not** a subset of D.

Naturally, two sets A and B are said to be **equal** if they have the same elements. That is, if every element of A is also in B, and vice versa. In that case, we write $A = B$.

There is one set that has no elements whatsoever - **the empty set**. It is commonly denoted by ϕ (the Greek letter Phi), or by a pair of braces (with nothing between them): $\{\ \}$.

We also have special symbols to denote various sets of numbers (and in particular, those mentioned in Section 1.4):

$$\mathbb{N} = \{1, 2, 3, 4, \dots\} \qquad \text{(The Natural Numbers)}$$

$$\mathbb{Z} = \{\dots, -3, -2, -1, 0, 1, 2, 3, 4, \dots\} \qquad \text{(The Integers)}$$

$$\mathbb{Q} = \left\{ \frac{a}{b} \ : \ a, b \in \mathbb{Z} \ \text{ and } \ b \neq 0 \right\} \qquad \text{(The Rational Numbers)}$$

$$\mathbb{R} = \text{The set of all real numbers (rational and irrational).}$$

Note that each of the above sets is a **subset** of the subsequent set: $\mathbb{N} \subseteq \mathbb{Z} \subseteq \mathbb{Q} \subseteq \mathbb{R}$.

Examples.

- In a set, the elements are not ordered, and a repeated element just counts as one element. For instance, the set $\{1, 2, 1, 2, 3\}$ has only three elements, and since order does not matter, we can write

$$\{1, 2, 1, 2, 3\} = \{1, 2, 3\} = \{3, 2, 1\}.$$

- All the following statements are correct. Can you explain why?

$$1 \in \mathbb{N} \quad , \quad 0 \notin \mathbb{N} \quad , \quad \frac{2}{3} \in \mathbb{Q} \quad , \quad \sqrt{2} \in \mathbb{R} \quad , \quad \sqrt{2} \notin \mathbb{Q},$$

$$\{2, 4, 6\} \subseteq \mathbb{N} \quad , \quad \mathbb{Z} \subseteq \mathbb{R} \quad , \quad \mathbb{R} \not\subseteq \mathbb{Z}.$$

- Set notation and the symbols introduced must be used carefully. For instance, what is the difference between $\mathbb{N}$ and $\{\mathbb{N}\}$?

  The set $\mathbb{N}$ is just the set of natural numbers, while $\{\mathbb{N}\}$ represents a set with a single element (that happens to be itself a set - the set of natural numbers).

  Similarly, ϕ denotes the empty set, while the set $\{\phi\}$ is not empty at all (it has an element, and that element happens to be the empty set).

  The set $\mathbb{Z}$ is the set of integers, while $\{\mathbb{Z}\}$ is a set containing one element - the set of all integers. Note that $\mathbb{Z} \subseteq \mathbb{R}$, while $\{\mathbb{Z}\} \not\subseteq \mathbb{R}$.

Before moving on to proofs involving sets, here are a few important observations.

Remarks.

- For any set B, we have $\phi \subseteq B$ and $B \subseteq B$.

  In other words, the empty set is a subset of any set, and every set is a subset of itself.

  Agreeing with the second statement, $B \subseteq B$, is probably easier. Every element of B is (obviously) also an element of B, which implies that $B \subseteq B$. But why is it true that $\phi \subseteq B$ (for any set B)? Well, let us think carefully about the meaning of being a subset. $A \subseteq B$ means that every element of A is also an element of B. Consequently, $A \not\subseteq B$ if there is at least one element in A, which does not belong to B. Since there are no elements in the empty set, we cannot argue that it is not a subset of B, and hence $\phi \subseteq B$. In other words, it is correct to say that any element in ϕ is also in B (as there are no elements at all in ϕ).

  This might be confusing, but it is a crucial point in understanding mathematical reasoning. A statement of the form "For any $x \in A$, we have..." is considered true when the set A is empty.

  In Chapter 3, we will take a closer look at the language of mathematics, and discuss similar statements in detail.

- A set can be defined (or described) using the **set-builder notation**. I.e., using a rule or a condition its elements must satisfy. This is particularly useful when a set has many elements, or is infinite, which makes it difficult (or impossible) to list all of its elements. In fact, the description of the set $\mathbb{Q}$, the rational numbers, was done using the set-builder notation:

$$\mathbb{Q} = \left\{ \frac{a}{b} \ : \ a, b \in \mathbb{Z} \ \text{ and } \ b \neq 0 \right\}.$$

  We could not possibly list all the rational numbers between the braces. Instead, we indicated that the rationals are numbers of the form $\frac{a}{b}$, where a and b are integers (and $b \neq 0$). Here are three more examples of using the set-builder notation:

$$C = \{t \in \mathbb{R} \ : \ t^2 - 4 = 0\} \qquad D = \{x \in \mathbb{R} \ : \ |x| > 1\} \qquad E = \{2k \colon k \in \mathbb{Z}\}.$$

  The set C is the solution set of the equation $t^2 - 4 = 0$. Clearly, as there are only two solutions, we could have written $C = \{-2, 2\}$. The set D, however, is infinite, so we have to use a rule or a pattern to describe its elements. The set E is the set of all numbers of the form $2k$, where k is an integer. This is, of course, the set of **even numbers**, and as the even numbers follow a simple pattern, we can also write $E = \{0, \pm 2, \pm 4, \pm 6, \dots\}$ or $E = \{\dots, -4, -2, 0, 2, 4, 6, \dots\}$.

- When two sets A and B are equal, each element in one set is also a member of the other set. In other words,

$$A = B \qquad \text{if and only if} \qquad A \subseteq B \ \text{ and } \ B \subseteq A.$$

Although straightforward, this observation is quite important, as it provides a **strategy** for proving equality of sets. Namely, if we are able to prove that $A \subseteq B$ and $B \subseteq A$, we can conclude that the sets A and B are equal.

Example 2.1.1 (Proving equality of sets)**.**
Consider the following two sets

$$A = \left\{ x \in \mathbb{R} : \ x \neq -1 \ \text{ and } \ \left(\frac{x}{x+1}\right)^2 \leq x \right\} \qquad , \qquad B = \{x \in \mathbb{R} : \ x \geq 0\}.$$

Prove that $A = B$.

Proof. Following the remark above, we show that $A = B$ by proving the two inclusions $A \subseteq B$ and $B \subseteq A$.

- If $x \in A$, then we have $x \geq \left(\frac{x}{x+1}\right)^2$, and since the square of a real number is always nonnegative, we conclude that
$$x \geq \left(\frac{x}{x+1}\right)^2 \geq 0.$$
However, $x \geq 0$ implies $x \in B$, which proves that $A \subseteq B$.
(Read this short argument carefully, and several times, to make sure the logic is clear: To prove $A \subseteq B$, we consider an arbitrary element x from A, and explain why this x must be a member of the set B.)

- To prove the other inclusion, we pick an arbitrary element $x \in B$ (which means that $x \geq 0$). Our task is to show that $x \in A$. To do that, there are two things that need to be verified. The first one is $x \neq -1$, which clearly holds, as $x \geq 0$, and the second is $\left(\frac{x}{x+1}\right)^2 \leq x$.
However, this inequality can be rearranged as follows:
$$\left(\frac{x}{x+1}\right)^2 \leq x \qquad \Leftrightarrow \qquad x^2 \leq (x+1)^2 \cdot x \qquad \Leftrightarrow \qquad x^3 + x^2 + x \geq 0.$$
The last inequality $x^3 + x^2 + x \geq 0$ is valid as $x \geq 0$, and hence the inequality $\left(\frac{x}{x+1}\right)^2 \leq x$ holds true as well. This shows that $x \in A$, and hence $B \subseteq A$.

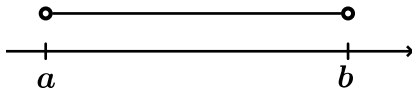
We see that $A \subseteq B$, and $B \subseteq A$, which imply that $A = B$, as needed. □

The Interval Notation and Set Operations.

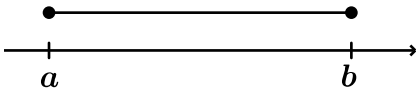
Open and closed intervals are commonly used sets of numbers. Using set-notation, we can defined them precisely, as follows.

Definition 2.1.2. If $a, b \in \mathbb{R}$, and $a \leq b$, we define:

(a) The **open interval** with endpoints a, b: $(a, b) = \{x \in \mathbb{R} \colon a < x < b\}$.



(b) The **closed interval** with endpoints a, b: $[a, b] = \{x \in \mathbb{R} \colon a \leq x \leq b\}$.



Note how we use **open and solid dots** in the diagrams, to indicate whether an endpoint is or is not a member of the interval.

The definitions of half-open (or half-closed) and infinite intervals are defined in a natural way. For instance, $[a, b)$ is the set of all real numbers between a and b, including a but not including b. The infinite interval $(-\infty, b)$ consists of all the real numbers which are smaller than b.

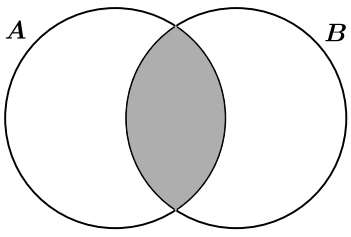
As with numbers, we can define **operations** on sets, that can be used to generate new sets from existing ones. Each definition below is accompanied by a **Venn diagram**, a diagram in which the shaded region represents the result of the corresponding set operation.

Definition 2.1.3. Let A, B be any two sets.

(a) The **intersection** of A and B is the set

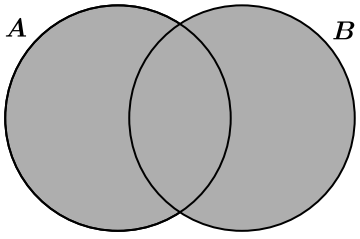
$$A \cap B = \{x \colon x \in A \text{ and } x \in B\}.$$

If $A \cap B = \phi$, we say the A and B are **disjoint** sets.



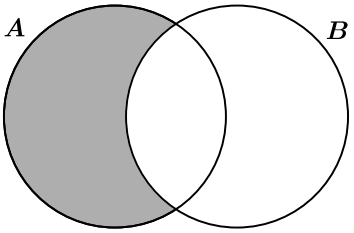
(b) The **union** of A and B is the set

$$A \cup B = \{x \colon x \in A \text{ or } x \in B\}.$$



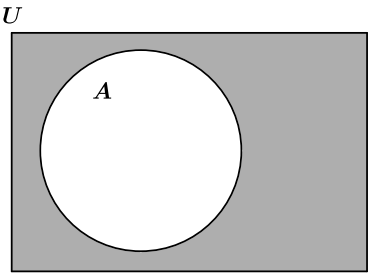
(c) The **difference** between A and B is the set

$$A \setminus B = \{x \colon x \in A \text{ and } x \notin B\}.$$[1]



(d) If A is a subset of some **universal set** U, we define the **complement** of A (with respect to U) as

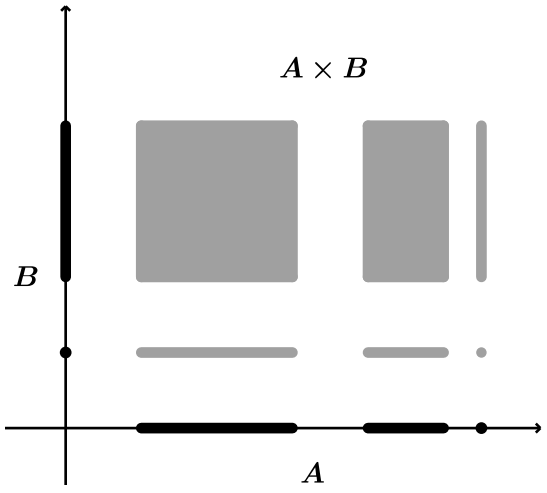
$$A^c = \{x \in U \colon x \notin A\}.$$



(e) The **Cartesian Product** of A and B is the set of all pairs (x, y), in which x is an element of A and y is an element of B:

$$A \times B = \{(x, y) \colon x \in A \text{ and } y \in B\} \ .$$

Note that members of $A \times B$ are **not** elements of A nor B, and therefore it is not possible to draw a Venn diagram for this operation. However, we use the following diagram to demonstrate the Cartesian product operation, in the case where A and B are sets of numbers.



In this diagram, A and B are sets of real numbers. A is drawn on the x-axis, and B on the y-axis. Each element (x, y) of $A \times B$ is a pair of numbers, which can be thought of as **a point in the plane**

[1] $A - B$ is another common notation for the difference of two sets.

(with coordinates x and y). The gray region in the first quadrant of the coordinate system represents the Cartesian product $A \times B$.

Example. Consider the following sets of numbers:

$$A = [0, \infty) \qquad , \qquad B = [-2.5, 3.5] \qquad , \qquad C = (1, 2).$$

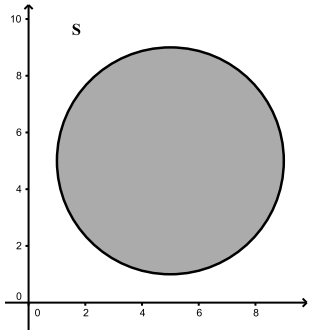
(The interval notation is used. Note that $[0, \infty)$ is the infinite interval $\{x \in \mathbb{R} \colon x \geq 0\}$.)

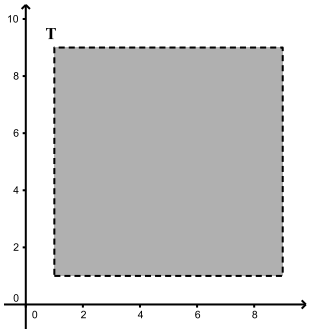
We can use set operations to create new sets, as follows:

- $A \cup B = [0, \infty) \cup [-2.5, 3.5] = [-2.5, \infty)$.
- $B \setminus C = [-2.5, 3.5] \setminus (1, 2) = [-2.5, 1] \cup [2, 3.5]$.
- $A \cap B = [0, \infty) \cap [-2.5, 3.5] = [0, 3.5]$.
- $B \cap \mathbb{Z} = [-2.5, 3.5] \cap \mathbb{Z} = \{-2, -1, 0, 1, 2, 3\}$.
- $C \cap \mathbb{N} = \phi$ (there are no natural numbers in the open interval $(1, 2)$).
- $(B \cap \mathbb{Z}) \times \{x, y\} = \{-2, -1, 0, 1, 2, 3\} \times \{x, y\} =$
  $= \{(-2, x), (-1, x), (0, x), (1, x), (2, x), (3, x), (-2, y), (-1, y), (0, y), (1, y), (2, y), (3, y)\}$
  (note that there are twelve elements in $(B \cap \mathbb{Z}) \times \{x, y\}$, and that each element is a pair).

Remark. The set $\mathbb{R} \times \mathbb{R}$, also denoted as $\mathbb{R}^2$, is the set of all **pairs** of real numbers, and hence can be thought of as the set of **all points in an (infinite) two-dimensional plane**. Any subset of $\mathbb{R}^2$ (and in particular, any Cartesian product of two sets of numbers) can be thought of as a region in the plane.

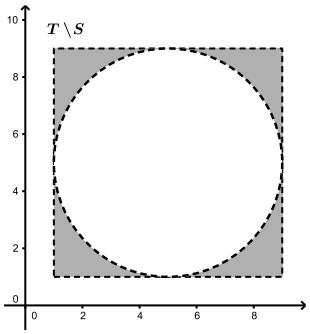
For instance, the sets $S = \{(x, y) \colon (x - 5)^2 + (y - 5)^2 \leq 16\}$ and $T = (1, 9) \times (1, 9)$ are both subsets of $\mathbb{R}^2$. S is a full circle (or a disk) of radius 4, and T is a square.

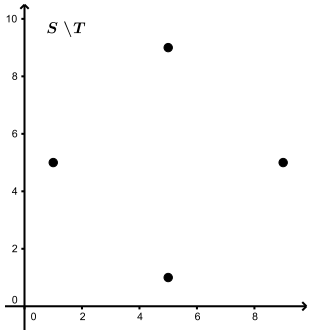




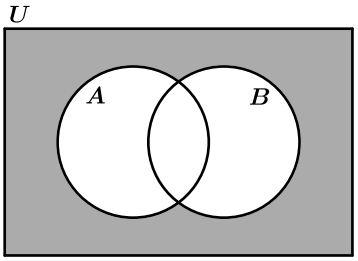
Note how we used solid and dotted lines, to indicate whether the boundary is or is not part of the set.

We can perform various operations on S and T, and obtain more subsets of the plane. For example, here are the sets $T \setminus S$ and $S \setminus T$. Can you draw diagrams for the sets $S \cap T$ and $S \cup T$?





Claim 2.1.4 (A set identity)**.** Let A and B be two subset of some universal set U.
Then $(A \cup B)^c = A^c \cap B^c$.



This is an example of a set identity. Much like algebraic identities (such as $(a+b)(a-b) = a^2-b^2$), set identities highlight relations between various set operations, and can be used to simplify expressions involving sets. But first – we need to prove them. In the Venn diagram on the right, the region $(A \cup B)^c$ is shaded, and with some imagination, it is not hard to see that the same region represents $A^c \cap B^c$. However, we must be cautious. An informal argument, that relies too heavily on diagrams, may be incomplete or flawed (and the risk is higher when there are many sets, or other objects, involved). Although such arguments are often acceptable by mathematicians, we prefer to be more careful, and validate the claim with a proof, that relies more on the definitions of set operations, rather than on diagrams.

Proof (of Claim 2.1.4).
We prove the identity by proving two inclusions: $(A \cup B)^c \subseteq A^c \cap B^c$ and $(A \cup B)^c \supseteq A^c \cap B^c$.

- Let $x \in (A \cup B)^c$. From the definitions of complements and unions, we have

$$x \notin A \cup B \qquad \Rightarrow \qquad x \notin A \ \text{ and } \ x \notin B,$$

  which implies that x is in both A^c and B^c. This means that $x \in A^c \cap B^c$ (definition of intersections), and so the inclusion $(A \cup B)^c \subseteq A^c \cap B^c$ is confirmed.

- To prove the other inclusion, we start with an $x \in A^c \cap B^c$. Again, referring to the definitions of complements and intersections, we get

$$x \in A^c \ \text{ and } \ x \in B^c \qquad \Rightarrow \qquad x \notin A \ \text{ and } \ x \notin B.$$

But if x is in none of the sets A and B, it cannot be in their union:

$$x \notin A \cup B \qquad \Rightarrow \qquad x \in (A \cup B)^c.$$

The inclusion $(A \cup B)^c \supseteq A^c \cap B^c$ is now also confirmed.

Both inclusions imply that $(A \cup B)^c = A^c \cap B^c$, as required. □

2.2 Functions

A function is a name we use for a "mathematical machine", creating a unique output for a given input. For instance, the formula $f(x) = x^2$ defines a function (on the set of real numbers), that assigns to each number, x, its square, x^2: $f(3) = 9$, $f(-0.5) = 0.25$, etc. Most likely you have already encountered the notion of a function in your high school mathematics classes, and have seen many examples of functions (linear, trigonometric, exponential, logarithmic functions, and more). Functions appear everywhere in mathematics and in the sciences, and are used to describe relations between two quantities, procedures and processes, geometric transformations, and more.

Nevertheless, as surprising as it may seem, the inputs and outputs of a function need not be numbers. Moreover, functions do not have to be necessarily defined (or described) by a formula. For instance, imagine a function g that assigns to any word, its first letter:

$$g(\text{hello}) = \text{h} \qquad , \qquad g(\text{chair}) = \text{c} \qquad , \qquad g(\text{love}) = \text{l} \qquad , \qquad \ldots$$

This is a perfectly legitimate function. In advanced mathematics, inputs and outputs of a function can be sets, vectors, matrices, or even other functions. This broader approach allows more flexibility in using and applying functions, while keeping common language, terminology, and symbols.

With these remarks in mind, we present the definition of a function.

Definition 2.2.1. A function f from a set A to a set B is a rule, that assigns to **each** element $a \in A$, a **unique** element $f(a) \in B$, called **the image** of a under f.

(To be honest, Definition 2.2.1 is still somewhat informal, as we do not explicitly state what a 'rule' is. Nevertheless, this definition is good enough for our needs in this course.)

Notation and Terminology.

- The set A and B (in Definition 2.2.1) are called **the domain** and **the codomain** (or **target space**) of f, respectively.

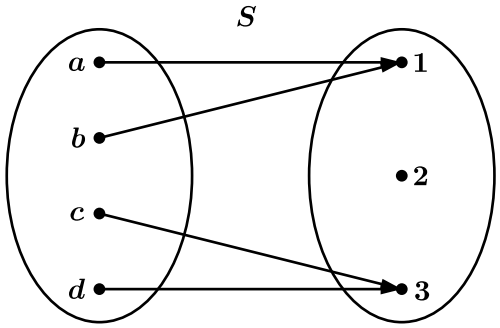
- The notation $f\colon A \to B$ means that f is a function with domain A and codomain B.
- The set $\{f(a)\colon a \in A\}$, denoted as $f(A)$, is called **the range** (or **the image**) of the function f. Note that $f(A)$ is always a subset of the codomain B.

Examples.

(a) We define a function $f\colon [-3,4] \to (-2,2)$ by $f(x) = \sin x$ (where x is in radians).
The image of $\frac{\pi}{2}$ is $f\left(\frac{\pi}{2}\right) = \sin\frac{\pi}{2} = 1$, and the image of $-\frac{\pi}{6}$ is $f\left(-\frac{\pi}{6}\right) = \sin\left(-\frac{\pi}{6}\right) = -\frac{1}{2}$. Note that as the domain of the function is the closed interval $[-3,4]$, quantities such as $f(2\pi)$ and $f(-\pi)$ are undefined.

The range of f is the set of all possible images of f. Although the codomain of f is the interval $(-2,2)$, numbers such as -1.5 or 1.9 will be never obtained as the sine of a number. From our knowledge of the sine function, we see that the image of f is $f([-3,4]) = [-1,1]$.

(b) A function can be defined using an **arrow-diagram**. For instance, the diagram below defines a function $S : \{a,b,c,d\} \to \{1,2,3\}$.



From the diagram, it follows that $S(a) = S(b) = 1$ and $S(c) = S(d) = 3$. The range of S is the set $\{1,3\}$.

(c) The function $g\colon \mathbb{Z} \times \mathbb{Z} \to \mathbb{Z}$, $g(m,n) = m^2 + n^2$, takes, as an input, a pair of integers, and returns, as the output, the sum of their squares. For instance, $g(2,5) = 2^2 + 5^2 = 29$, $g(-1,3) = 1 + 9 = 10$, etc. Describing the range of g (in an explicit way) requires additional work, which we omit for now.

(d) We can define a function T on the set of all calendar months, that returns the number of days in a given month:

$$T\colon \{\text{Jan.}, \text{Feb.}, \ldots, \text{Dec.}\} \to \{25, 26, 27, \ldots, 35\}$$

$$T(y) = (\text{number of days in month } y, \text{ in a non-leap year}).$$

For example, $T(\text{Jul.}) = 31$, $T(\text{Apr.}) = 30$ and $T(\text{Feb.}) = 28$. The image of T is the set of possible outcomes, $\{28, 30, 31\}$.

Remark. Note that if A is a set, and $f, g\colon A \to \mathbb{R}$ are two functions, we can use basic arithmetic to create new functions, such as $f+g$, $f-g$, $f \cdot g$, $5f^2$, and so on.

For instance, if $f, g\colon \mathbb{R} \to \mathbb{R}$ are given by $f(x) = x^2$ and $g(x) = \sin x$, then $(f+g)(x) = x^2 + \sin x$, $(f \cdot g)(x) = x^2 \sin x$, and $f^3(x) = (x^2)^3 = x^6$.

Next, we turn to the definition of the graph of a function.

Definition 2.2.2. The **graph** of a function $f\colon A \to B$ is the set $\{(a, f(a))\colon a \in A\}$.
This is a subset of $A \times B$.

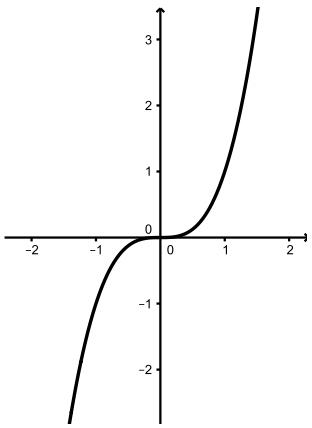
In other words, the graph of a function f is the set of all pairs (x, y), where x is an element of the domain A, and y is the image of that x. Interestingly, this definition applies to all functions, even those involving non-numerical elements. For instance, the graph of the function T from the example above (giving the number of days in a calendar month) is the following set:

$$\{(\text{Jan.}, 31), (\text{Feb.}, 28), (\text{Mar.}, 31), (\text{Apr.}, 30), (\text{May.}, 31), (\text{Jun.}, 30),$$
$$(\text{Jul.}, 31), (\text{Aug.}, 31), (\text{Sep.}, 30), (\text{Oct.}, 31), (\text{Nov.}, 30), (\text{Dec.}, 31)\}.$$

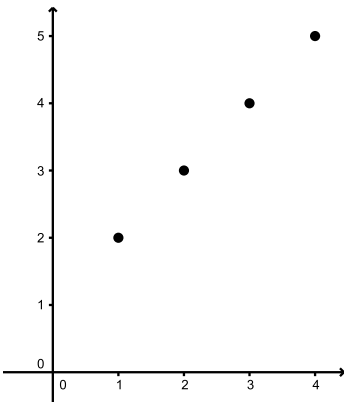
Note that we use the term 'graph', even when it is not clear at all whether we can actually draw a graph in the usual sense.

If the domain and the codomain of a function are **sets of numbers**, then every element in the graph is a pair of numbers, which can be drawn as a point in the two-dimensional plane $\mathbb{R}^2$.

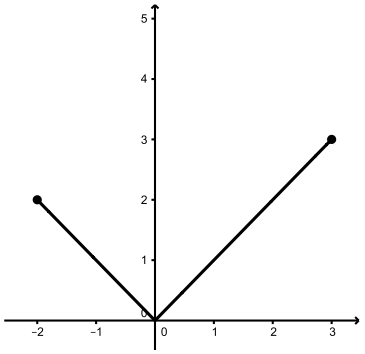
Examples.
Here are three functions, and their graphs.



$\boldsymbol{f : \mathbb{R} \to \mathbb{R}\ ,\quad f(x) = x^3}$



$\boldsymbol{p : \mathbb{N} \to \mathbb{N}\ ,\quad p(n) = n+1}$



$\boldsymbol{R : [-2, 3] \to \mathbb{R}\ ,\quad R(x) = |x|}$

Note how each point on the graph of a function corresponds to an element in its domain. For instance, the domain of p is the set of all natural numbers, and so every point on its graph must correspond to a pair (x, y) where $x = n \in \mathbb{N}$ and $y = n + 1$. Connecting the points with a straight line would be incorrect!

Our next example requires a careful proof, and involves several ideas from previous sections.

Example. Consider the function $f\colon \mathbb{R} \to \mathbb{R}$, $f(x) = \dfrac{x}{1+x^2}$.
Prove that the range of f, $f(\mathbb{R})$, is the closed interval $\left[-\frac{1}{2}, \frac{1}{2}\right]$.

Proof. Our task is to prove that $f(\mathbb{R}) = \left[-\frac{1}{2}, \frac{1}{2}\right]$, which is an equality between two sets (the range, or image of f, and the closed interval). As we did previously, we prove it by showing mutual subset inclusion.

- Let $y \in f(\mathbb{R})$. Then, by the definition of range, we get that $y = f(x) = \dfrac{x}{1+x^2}$ for some $x \in \mathbb{R}$. Note that the statement $y \in \left[-\frac{1}{2}, \frac{1}{2}\right]$ is equivalent to

$$-\frac{1}{2} \le y \le \frac{1}{2} \qquad \Leftrightarrow \qquad -\frac{1}{2} \le \frac{x}{1+x^2} \le \frac{1}{2} \qquad \Leftrightarrow \qquad -(1+x^2) \le 2x \le 1+x^2$$

$$\Leftrightarrow \qquad -x^2 - 2x - 1 \le 0 \le x^2 - 2x + 1 \qquad \Leftrightarrow \qquad -(x+1)^2 \le 0 \le (x-1)^2,$$

  and since the latter inequalities are clearly valid, we conclude that $y \in \left[-\frac{1}{2}, \frac{1}{2}\right]$. This proves the inclusion $f(\mathbb{R}) \subseteq \left[-\frac{1}{2}, \frac{1}{2}\right]$.

- Conversely, we now start with a $y \in \left[-\frac{1}{2}, \frac{1}{2}\right]$, and we need to show that $y \in f(\mathbb{R})$. More explicitly, we need to show that $y = f(x) = \dfrac{x}{1+x^2}$ for some real number x.

  To do so, let us think of the equality $y = \dfrac{x}{1+x^2}$ as an equation in x (where y is a fixed number). This equation can be rearranged as

$$y(1+x^2) = x \qquad \Leftrightarrow \qquad y \cdot x^2 - x + y = 0,$$

  which is a quadratic equation in x (as long as $y \neq 0$). To show that $y \in f(\mathbb{R})$, we argue that this quadratic has (at least) one solution.

  Recall (from Theorem 1.1.1) that a quadratic equation $ax^2 + bx + c = 0$ has real solutions if (and only if) $b^2 - 4ac \ge 0$. In our case, $b^2 - 4ac = 1 - 4y^2$, and since $-\frac{1}{2} \le y \le \frac{1}{2}$, the condition $1 - 4y^2 \ge 0$ is satisfied, from which it follows that $y \in f(\mathbb{R})$ (the case $y = 0$ is easy to handle: $f(0) = 0$ and so $0 \in f(\mathbb{R})$).

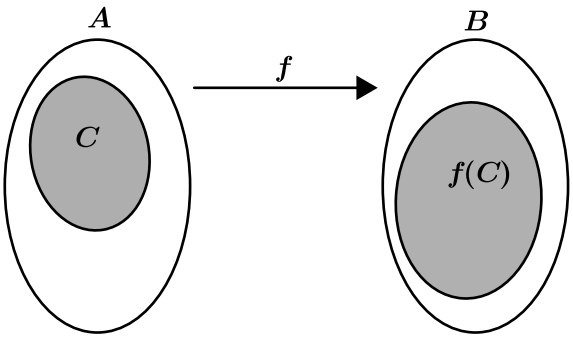
  We conclude that $y \in f(\mathbb{R})$, which proves the other inclusion $\left[-\frac{1}{2}, \frac{1}{2}\right] \subseteq f(\mathbb{R})$, as needed.

□

We end this section with another proof. This time, however, we prove a **general statement** about functions and sets (rather than dealing with a specific function). Here is a preliminary definition.

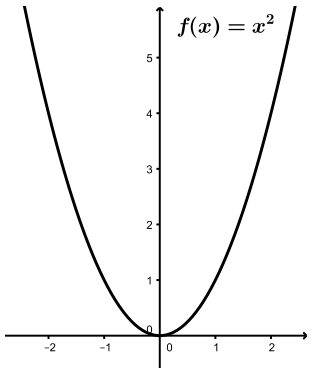
Definition 2.2.3. If $f\colon A \to B$ is a function, and $C \subseteq A$, then **the image of C under f** is the set $f(C) = \{f(a) : a \in C\}$. Note that $f(C)$ is a subset of the codomain B.

In words, the image of C under f is the set of all images of elements in C. Note that when C is the whole domain of f (namely, $C = A$), then the image of C under f is the same as the image of f (see remark on page 33). The following diagram visualizes the notion of the image of a set.



Example.

Consider the function $f\colon \mathbb{R} \to \mathbb{R}$, $f(x) = x^2$, whose graph is given on the right.



Then, we have the following equalities (make sure you can justify each of these equalities):

$$f(\{-1,1,2,3\}) = \{1,4,9\} \quad , \quad f([2,4]) = [4,16] \quad ,$$
$$f((-1,1)) = [0,1) \quad , \quad f(\mathbb{R}) = [0,\infty) \quad ,$$
$$f(\mathbb{Z}) = \{0,1,4,9,16,25,\dots\} \quad .$$

We are now ready to state and prove the following proposition.

Proposition 2.2.4. If $f\colon A \to B$ is a function, and C, D are subsets of A, then $f(C \cup D) = f(C) \cup f(D)$.

In words, the proposition says that the image of a union of two sets, equals the union of their images under the function f. Again, as we need to prove equality between two sets, we proceed by showing mutual subset inclusion.

Proof. Let $y \in f(C \cup D)$, then $y = f(x)$ for some x in $C \cup D$. Since $x \in C \cup D$, then either $x \in C$ or $x \in D$ (or both[2]). This implies that either $y = f(x) \in f(C)$ or $y = f(x) \in f(D)$, from which we get that $y \in f(C) \cup f(D)$. This proves the inclusion $f(C \cup D) \subseteq f(C) \cup f(D)$.

Conversely, let $y \in f(C) \cup f(D)$. Then either $y \in f(C)$ or $y \in f(D)$. If $y \in f(C)$, then $y = f(a)$ for some $a \in C$. If $y \in f(D)$, then $y = f(b)$ for some $b \in D$. In either case, we see that y is equal to $f(x)$ for some element x in $C \cup D$, and hence $y \in f(C \cup D)$. This proves the inclusion $f(C) \cup f(D) \subseteq f(C \cup D)$.

Therefore, $f(C \cup D) = f(C) \cup f(D)$, as needed. □

2.3 The Field Axioms

In mathematics, proofs are used to validate statements. In most cases, these proofs rely on earlier established facts - things we have already proved, or that we are convinced are true. For example, to prove the quadratic formula (Theorem 1.1.1), we relied on the known formula $(x+y)^2 = x^2 + 2xy + y^2$, and on the fact that squares of real numbers cannot be negative. Looking more carefully at our work so far, we notice that many other "known" algebraic and arithmetic rules, such as $(a \cdot b)^2 = a^2 \cdot b^2$ and $2 \cdot \frac{1}{2} = 1$, were used. In Section 1.2, we used, without proof, several Basic Facts (see page 10) to prove the Arithmetic-Geometric Mean Inequality.

This raises an important fundamental question: How can we rely on "facts" or "rules" that were not properly proved? Basing our arguments on rules that were not fully justified can be risky, and jeopardize the credibility of our conclusions. After all, if we are relying on invalid rules to prove a theorem, the proof may be flawed.

How can we resolve this issue? One possible approach would be to go back, and try to construct proofs for all these rules and facts we relied on (for instance, we might want to try and supply proofs for the Basic Facts from page 10). This, however, raises another question: Do we expect to be able to prove **every single algebraic identity and arithmetic rule from scratch**? For instance, can we prove that $x + y = y + x$ for any two real numbers x, y? Or that $(-1) \cdot (-1) = 1$? Or $0 \cdot a = 0$ for any real number a? After all, mathematical proofs depend on results that have been previously established, and so proving a fundamental result without having 'earlier statements' to rely on seems impossible.

Indeed, there must be a starting point. We have no choice but to accept some mathematical facts as true (or assume that they are true), and then build our theory on this list of assumptions, from which other conclusions follow. Rules such as $x + y = y + x$ or $x + 0 = x$ will be called **axioms**, and assumed to be true (for any x, y). Other, more advanced rules, will be proved from these assumptions.

[2] In mathematics, when we use the word 'or', we allow both possibilities to occur.

You may ask: How can we assume something without proving it? Or, in other words, how do we know our assumptions are correct? The answer to this question will probably surprise you: **It does not matter whether our assumptions (or axioms) are correct or not**. In fact, the question of whether an axiom is correct or not is problematic. What do we mean by "correct"? An axiom may be a reasonable assumption in some contexts, and inappropriate in others.

The main point here is that mathematics is not concerned with absolute truth (whether there is such a thing or not). Mathematics is all about conclusions one can make from a given set of assumptions (or axioms).

To dive in, and make the discussion more explicit, we present the definition of a field and the field axioms.

Definition 2.3.1. A set F, with two operations, $+$ (addition) and $\cdot$ (multiplication), and distinguished elements 0 and 1 (with $0 \neq 1$)[3], is called a **field**, if the following list of axioms hold.

(0) $x + y \in F$ and $x \cdot y \in F$ for any $x, y \in F$ (**closure** under addition and multiplication).

(1) $x + (y + z) = (x + y) + z$ and $x \cdot (y \cdot z) = (x \cdot y) \cdot z$ for any $x, y, z \in F$ (**associativity** of addition and multiplication).

(2) $x + y = y + x$ and $x \cdot y = y \cdot x$ for any $x, y \in F$ (**commutativity** of addition and multiplication).

(3) $x + 0 = x$ and $x \cdot 1 = x$ for all $x \in F$
(0 and 1 are called the **additive identity** and the **multiplicative identity**, respectively).

(4) For any $x \in F$, there is a $w \in F$ such that $x + w = 0$ (existence of **negatives**).
Moreover, if $x \neq 0$, then there is also an $r \in F$ such that $x \cdot r = 1$ (existence of **reciprocals**).
We denote $w = -x$ and $r = x^{-1}$.[4]

(5) $x \cdot (y + z) = x \cdot y + x \cdot z$ for any $x, y, z \in F$ (**distributivity** of addition over multiplication).

A few remarks are in place.

- None of the field axioms are new to you (in the context of real numbers). You have seen them all before and used them since elementary school. However, when we are given a field, we assume that axioms (0)-(5) hold true, and may use them to prove other statements. It turns out that many other

[3] 0 and 1 need not be the well known numbers zero and one from the number line, and some prefer to write 0_F and 1_F to emphasize that fact. That is, 0_F and 1_F are the zero and one **of the field** F.

[4] $-x$ and x^{-1} are often called the **additive inverse** and **multiplicative inverse** of x, respectively.

basic and more advanced properties of numbers can be derived from the field axioms, as we will see soon.

- In the definition of a field, the operations of **subtraction** and **division** are **not** mentioned. However, axiom (4) requires the existence of negatives and reciprocals. This allows us to **define** subtraction and division in terms of addition and multiplication:

$$x - y = x + (-y) \qquad \text{and} \qquad \frac{x}{y} = x \cdot y^{-1} \quad (\text{if } y \neq 0).$$

  We also use the usual notation for powers, and interpret them, of course, as repeated multiplication: $x^2 = x \cdot x,\ x^3 = x \cdot x \cdot x$, etc.

- We know well that every number x has a **unique** negative $-x$ (for instance, there is only one number w such that $2+w = 0$). The uniqueness, however, was not included in axiom (4), as it can be actually **derived** from the field axioms, as follows: If $x + w_1 = 0$ and $x + w_2 = 0$, then

$$w_1 \overset{(3)}{=} w_1 + 0 = w_1 + (x + w_2) \overset{(1)}{=} (w_1 + x) + w_2 \overset{(2)}{=} (x + w_1) + w_2 = 0 + w_2 \overset{(2)}{=} w_2 + 0 \overset{(3)}{=} w_2$$

  (the numbers indicate the axiom that was used in each step).
  Similarly, one can prove that reciprocals are unique.

- There are many other basic properties of numbers that were not included in the definition of a field. We know, for instance, that if a is a real number, then $a \cdot 0 = 0$ and $-(-a) = a$. Why aren't these on the above list of axioms? Mathematicians often try to keep the list of axioms as short as possible, and not include facts that can be derived from other axioms. Here is how the property $a \cdot 0 = 0$ can be **proved**.

Claim 2.3.2. Let F be a field, and $a \in F$. Then $a \cdot 0 = 0$.

Note that in the proof, we must be extremely cautious not to use any arithmetic or algebraic rule that is not part of the field axioms.

Proof.

$$a \cdot 0 \overset{(3)}{=} (a \cdot 0) + 0 \overset{(4)}{=} (a \cdot 0) + [(a \cdot 0) + (-a \cdot 0)] \overset{(1)}{=} [(a \cdot 0) + (a \cdot 0)] + (-a \cdot 0) \overset{(5)}{=} a \cdot (0 + 0) + (-a \cdot 0) \overset{(3)}{=}$$

$$\overset{(3)}{=} (a \cdot 0) + (-a \cdot 0) \overset{(4)}{=} 0\ .$$

Again, the numbers in brackets correspond to the field axiom used in each step. □

Examples. The set of all real numbers form a field,[5] as the field axioms hold true. What about other sets of numbers (with the usual addition and multiplication)?

- The set of rational numbers $\mathbb{Q}$ is also a field, as all the axioms are satisfied (check!).
- The set of integers $\mathbb{Z}$ is **not** a field, as numbers such as 2 and -5 have no integer reciprocals, and so axiom (4) fails.
- The closed interval $[-1,1]$ is **not** a field, as it is **not** closed under addition (axiom (0)). For instance, $\frac{1}{2}$ and $\frac{2}{3}$ are both in $[-1,1]$, but their sum is not.

Let us take another look at Claim 2.3.2. What do we learn from this claim and its proof? The fact that $a \cdot 0 = 0$ is well known to us, and we have been using it for years, so why bother proving it? Here are two reasons:

(1) Even though the content of the claim is not new to us, we now see how it can be derived from the field axioms only. Therefore, there is no need to present it as an additional axiom (or assumption).

(2) Furthermore, note that this claim is, in fact, quite general, and need not be applied to 'numbers' only! In its proof, we relied solely on the field axioms, and so whenever we happen to notice a 'world' in which the field axioms hold, the claim will be valid.
But what others 'worlds' can be called fields? Here are a few examples.

Examples.

- **Rational Functions.**
Recall that a polynomial is a sum of expressions of the form $a \cdot x^k$, where k is a nonnegative integer ($k = 0, 1, 2, \dots$), and a is any real number. For instance, $3x^2 + 2x + 1$, $x^7 - 9$ and $x^3 - 6x^{10}$ are polynomials. Clearly, polynomials can be added and multiplied, and we even have a 'zero polynomial' and a 'one polynomial' (the constant functions $f(x) = 0$ and $g(x) = 1$). Does the set of polynomials form a field?
No, it does not, as axiom (4) fails (why?). In fact, this is the only axiom that fails for polynomials (check!). However, the set of rational functions, i.e., the set
$$\left\{ \frac{f(x)}{g(x)} \ : \ f \text{ and } g \text{ are polynomials, and } g \neq 0 \right\}$$
is a field, as all the axioms are satisfied (verify this!).

[5] This matches our past experience with real numbers, and we take this fact for granted. There are ways to formally construct the real numbers and then verify the field axioms, but this is beyond the scope of the notes.

- **A Field with Two Elements.**

  What is the smallest possible field? Every field, according to the definition, must include at least two elements: 0 and 1. Is it possible to define addition and multiplication on the set $F = \{0, 1\}$ in such a way that all the field axioms hold true?

  We must have set $0 + 0 = 0$, $0 + 1 = 1 + 0 = 1$, $0 \cdot 0 = 0 \cdot 1 = 1 \cdot 0 = 0$ and $1 \cdot 1 = 1$. Otherwise, we violate axiom (3), or the Claim 2.3.2 above. But what should be $1 + 1$? If 0 and 1 are the only elements in our universe, then 2 is not available to us, and hence setting $1 + 1 = 2$ is out of the question. We must therefore set $1 + 1$ to be either 0 or 1. However, if we set $1 + 1 = 1$, then axiom (4) fails, as 1 won't have a negative (i.e., $1 + w$ is never zero). Therefore, we must set $1 + 1 = 0$. We summarize addition and multiplication in F with the following two tables.

$+$	0	1
0	0	1
1	1	0

$\cdot$	0	1
0	0	0
1	0	1

  As strange as it may seem, the set $F = \{0, 1\}$, with the addition and multiplication defined above, does satisfy the field axioms, and hence **is a field**. This is **a field with two elements**, and it is widely used in mathematics and applications.

  There are many other fields that have only finitely many elements. We call them **finite fields**.

Building addition and multiplication tables for finite field requires some effort, and more advanced knowledge. Nevertheless, we present a proof involving a field with four elements.

Claim 2.3.3. Let $F = \{0, 1, a, b\}$ be a field with four elements. Then $a \cdot b = 1$.

Proof. We prove this claim using an elimination strategy. As F is a field, any two elements can be multiplied, and the product must be also one of the field elements. In particular, the product $a \cdot b$ is equal to either $0, 1, a$ or b. We eliminate the options $0, a$ and b, from which it follows that $a \cdot b = 1$.

- If $a \cdot b = a$, then we can "divide" both sides by a, to obtain $b = 1$, which is impossible (as $0, 1, a, b$ are four distinct elements).

  In fact, we should be more careful, and make sure that "dividing by a" can be justified by the field axioms:

$$a \cdot b = a \quad \Rightarrow \quad a^{-1} \cdot (a \cdot b) = a^{-1} \cdot a \quad \Rightarrow \quad (a \cdot a^{-1}) \cdot b = a \cdot a^{-1} \quad \Rightarrow \quad 1 \cdot b = 1$$

which implies that $b = 1$, and that is impossible. Note that only the field axioms were used in the steps above (can you specify the axioms that were used in each step?).

- A similar argument shows that $a \cdot b = b$ implies $a = 1$, which is impossible.
- Finally, if $a \cdot b = 0$, we can multiply both sides by b^{-1}, and get

$$(a \cdot b) \cdot b^{-1} = 0 \cdot b^{-1} \quad \Rightarrow \quad a \cdot (b \cdot b^{-1}) = b^{-1} \cdot 0 \quad \Rightarrow \quad a \cdot 1 = 0$$

and hence $a = 0$, which is impossible. Note how the field axioms and Claim 2.3.2 were used.

As the options a, b and 0 were all eliminated, we conclude that $a \cdot b = 1$. □

The Real Numbers System.

In Section 1.4 we mentioned that the set of all rational and irrational numbers on the number line form the set of real numbers. From a formal point of view, this is not a satisfactory definition of the reals (can you figure out why?). However, we just learned that the real numbers satisfy the field axioms. Can we perhaps define the real numbers as a set, equipped with two operations (addition and multiplication), that satisfies the field axioms? Well, not quite.

The examples above show that there are other fields out there, some of which are very different from the real numbers we are so used to (for instance: finite fields, the rational numbers, rational functions, etc.). To fully characterize the reals, we will need to add a few assumptions to the field axioms, that distinguish the real numbers from other possible fields.

One way to do that, is to add the **order** and **completeness** axioms.

The **order axioms** outline a few basic properties regarding ordering of numbers, allowing us to work with inequalities, and make sense of statements such as $a < b$ or $c \geq d$. We have decided not to spend more time on outlining the axioms and their implications. Some details are provided in Exercise 2.5.55.

The **completeness axiom** is probably the most complicated one. Roughly speaking, it says that there are no holes or gaps on the number line (as we have for the set of rational numbers). The precise statement is beyond the scope of these notes, and you may learn more about it in a higher-level analysis course.

The main point we want to make here, is that once we group the field axioms with the order and completeness axioms, we do obtain a full precise characterization of the real number system, and can formally define the reals by means of axioms. Once done, any other property of the real numbers can be derived from the axioms (such as existence of square roots and the fact that $1 > 0$).

2.4 Appendix: Well-defined Functions

According to Definition 2.2.1, three pieces of information need to be specified in order to describe a function:

- A set A (the **domain**).
- Another set B (the **codomain**).
- A rule assigning to **each** element of A **exactly one** element from B.

If any of the three ingredients above are not clearly specified (or understood from the context), or "the rule" is ambiguous (or problematic and cannot be applied to some elements of A), we say that the function is **not well-defined**. Here are a few examples.

Examples.

(a) The function $f\colon \mathbb{N} \to \mathbb{N}\ ,\ f(n) = n - 10$ is **not well defined**, as the formula for f does not produce a natural number for all $n \in \mathbb{N}$. For instance, if $n = 6$, then $n - 10 = 6 - 10 = -4$ is not an element of the codomain. If we enlarge the codomain of f, to include, for instance, all integers, then our function will be well-defined.

(b) The function $g\colon \mathbb{R}^2 \to \mathbb{R}\ ,\ g(x,y) = \dfrac{1}{x^2+y^2}$ is also **not a well-defined function**, as the given formula cannot be used to compute the image of the pair $(0,0)$ (which is an element of our domain $\mathbb{R}^2$), as division by zero is not allowed. If we restrict the domain of g to $\mathbb{R}^2 \setminus \{(0,0)\}$, the function will be well-defined. Another option is do define $g(0,0)$ separately, as follows:

$$g\colon \mathbb{R}^2 \to \mathbb{R} \qquad , \qquad g(x,y) = \begin{cases} \dfrac{1}{x^2+y^2} & \text{if } \ (x,y) \neq (0,0) \\ -2 & \text{if } \ (x,y) = (0,0) \end{cases}.$$

This way, we obtain a **well-defined function**.

(c) Often, the notion of **well-defined functions** refers to cases, where a function is defined **through representatives of elements in its domain**. Consider, for instance, the following definition:

$$h\colon \mathbb{Q} \to \mathbb{R} \qquad , \qquad h\left(\frac{a}{b}\right) = a + b.$$

The function h, whose domain is $\mathbb{Q}$, is supposed to assign to every rational number $\frac{a}{b}$, a real number. However, there are multiple ways to represent a rational number. For example, what would be $h(0.25)$? Well, we can write 0.25 as $\frac{1}{4}$, and conclude that

$$h(0.25) = h\left(\frac{1}{4}\right) = 1 + 4 = 5.$$

On the other hand, we can also conclude that

$$h(0.25) = h\left(\frac{3}{12}\right) = 12 + 3 = 15.$$

We see that the definition of h produced **multiple images** for 0.25, that depend on the way we represent 0.25 as a fraction. This violates Definition 2.2.1, and hence h is **not a well-defined function**.

(d) The function $p : \mathbb{Q} \to \mathbb{R}$, given by $p\left(\frac{a}{b}\right) = \frac{3a}{2b}$ **is well-defined**, as the formula for p does not depend on the way we represent a rational number as a fraction (can you see why?).

(e) Consider the two functions $F, G : \mathbb{R} \to \mathbb{R}$ defined as follows:

$$F(x) = \begin{cases} x^2 + 1 & \text{if } x \geq 3 \\ 3x - 5 & \text{if } x \leq 3 \end{cases} \qquad G(x) = \begin{cases} \sqrt{x} + 3 & \text{if } x \geq 4 \\ 2x - 3 & \text{if } x \leq 4 \end{cases}.$$

For both functions, the two cases in their definition overlap. For F, if $x = 3$, we should be able to use any of the two formulas to compute $F(3)$. However,

$$x^2 + 1 = 3^2 + 1 = 10 \qquad \text{and} \qquad 3x - 5 = 3 \cdot 3 - 5 = 4,$$

and hence F is **not a well-defined function**. On the other hand, the two ways to compute $G(4)$ yield the same number. Indeed, if $x = 4$, we get:

$$\sqrt{x} + 3 = \sqrt{4} + 3 = 5 \qquad \text{and} \qquad 2x - 3 = 2 \cdot 4 - 3 = 5,$$

and so G is a **well-defined function**.

2.5 Exercises for Chapter 2

2.5.1. Let $S = \{(x, y) \in \mathbb{N}^2 : \ (2 - x)(2 - y) < 2(4 - x - y)\ \}$.
Prove that $S = T$, where $T = \{(1, 1), (1, 2), (2, 1), (1, 3), (3, 1)\}$.

2.5.2. Let $S = \{2, 3\} \times \{-2, -1, 0\}$, and let T be the set of all ordered pairs $(x, y) \in \mathbb{Z} \times \mathbb{Z}$ such that $-2 \leq x + 2y \leq 3$. Prove that $S \subseteq T$. Does equality hold? Explain.

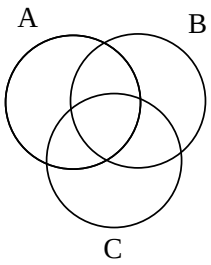
2.5.3. Write either $\subseteq$ or $\not\subseteq$ in the space provided. Explain your answer briefly.

ϕ _____ $\mathbb{N}$ $\qquad$ $\{\phi\}$ _____ $\mathbb{R}$ $\qquad$ $\mathbb{Z}$ _____ $\mathbb{N} \cup \mathbb{Q}$ $\qquad$ $\mathbb{Z}$ _____ $\mathbb{N} \cap \mathbb{Q}$

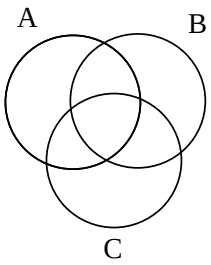
2.5.4. Describe the 'intervals' $[a, b]$ and (a, b), in the case where $a = b$.

2.5.5. In the following Venn diagrams, shade the region that corresponds to the given set.

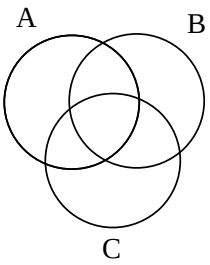
(a) $(A \cup B) \setminus C$



(b) $C \setminus (A \setminus B)$

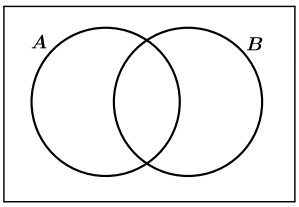


(c) $(B \setminus C) \cap A$

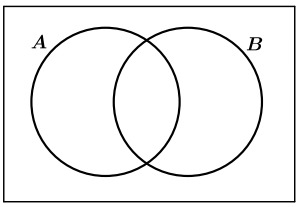


2.5.6. In the following Venn diagrams, shade the region that corresponds to the given set.

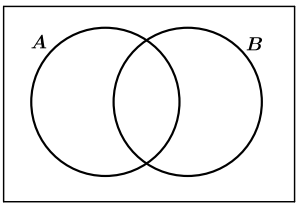
(a) $A \cup B^c$



(b) $B \setminus A^c$



(c) $A^c \cap B^c$



2.5.7. (a) The equality $[0.5, 7.5] \cap \mathbb{N} = [1, 7]$ is **incorrect**. Why?

(b) The equality $\{4\} \times \{5\} = \{20\}$ is **incorrect**. Why?

(c) The equality $\{a, b, c\} \cup \phi = \{a, b, c, \phi\}$ is **incorrect**. Why?

2.5.8. Let $A = [-1, 1]$, $B = (-\pi, \pi)$, $C = [2, \infty)$ and $U = \mathbb{R}$ be the universal set.
Find the sets $A \cap C$, $A^c \cap B$, $(B \cap C) \cap \mathbb{Z}$ and $B \setminus A$.
Use the interval notation, and the symbols { , }, $\cup$, ϕ, ∞ and π only.

2.5.9. For each statement, decide whether it is true or false. Justify your answer briefly.

(a) $\{(x, y) \colon x, y \in \mathbb{R} \quad \text{and} \quad x - 1 = 0\} \subseteq \{(x, y) \colon x, y \in \mathbb{R} \quad \text{and} \quad x^2 - x = 0\}$

(b) $\{x \in \mathbb{R} : x^3 - 2x = 0\} \subseteq \mathbb{Q}$

(c) $\mathbb{N} \in \mathbb{R}$

(d) $\mathbb{Z} \times \mathbb{R} \subseteq \mathbb{R} \times \mathbb{Z}$

(e) $\mathbb{N} \times \mathbb{Z} \subseteq \mathbb{Q} \times \mathbb{R}$

2.5.10. For each statement, decide whether it is true or false (for any sets A, B, C). Draw a Venn diagram to support your answer.

(a) $A \setminus B \subseteq A$

(b) $A \setminus B \subseteq B$

(c) $(A \cup B) \cap C = A \cup (B \cap C)$

(d) $(A \cap B) \cap C = A \cap (B \cap C)$

(e) $A \subseteq A \cap B$

(f) $(A \setminus B)^c = A^c \setminus B^c$

2.5.11. If A and B are two sets satisfying $A \setminus B = B \setminus A$, what can we conclude about A and B? Explain.

2.5.12. Is it true that $(A \times A) \setminus (B \times B) = (A \setminus B) \times (A \setminus B)$ for any two sets A, B?
If it is true, prove it. Otherwise, find a counterexample.

2.5.13. (a) Given that $A = \left\{-3, -1, \frac{1}{2}, 10, \sqrt{2}\right\}$ and $B = \{2, 4\}$, list all the elements in the set $C = B \times (A \cap \mathbb{Z})$.

(b) Given the intervals $I = [-1, 8]$ and $J = (3,5)$, list all the elements in the set $B = (I \setminus J) \cap \mathbb{N}$.

(c) Write the set $\{x \in \mathbb{R} : \ 0 < x^2 \leq 25 \ \}$ as a **union of two intervals**.

(d) Express the set $A = \{\, x : \ 1 < x^2 < 4 \ \}$ both as a union of two intervals, and as a difference of two intervals.

2.5.14. Give an example of a set A, for which $A \cap [1, 4] = A \cap \mathbb{N}$ and $A \setminus \mathbb{Z} \neq \phi$.

2.5.15. Consider the following two subsets of $\mathbb{R}^2$: $A = [0, 2] \times [0, 2]$ and $B = [-1, 1] \times [-1, 1]$.
Draw the sets A, B, $A \cap B$ and $A \setminus B$ in the plane. Use a **solid** or a **dotted line** to indicate whether the boundary is or is not part of the set.

2.5.16. Consider the sets $S = \{(x, y) \ : \ y \geq x^2 \ \}$ and $T = \{(x, y) \ : \ y \leq x + 2 \ \}$.
Draw the sets S, T, and $S \cap T$ in the two-dimensional plane.

2.5.17. Consider the following two subsets of $\mathbb{R}^2$: $T = [-1, 1] \times [-1,1]$ and $S = \left\{ (x, y) : \ x^2 + y^2 \leq 4 \right\}$.
Sketch the set $S \setminus T$.

2.5.18. Let $D = [1, 3] \cup \{4\}$ (note that this is a subset of $\mathbb{R}$). Sketch the set $D \times D$ (in $\mathbb{R}^2$).

2.5.19. Let a, b, c, d be real numbers with $a < b < c < d$. Express the set $[a, b] \cup [c, d]$ as a difference of two sets.

2.5.20. Let $S = \{x \in \mathbb{R} \ : \ (x - 2)(x + 3) < 0\}$, T the interval $(-4\,, 2)$ and U the interval $(-3\,, 5)$.
Use set operations to write a simple relation between the sets S, T and U.

2.5.21. Prove the following set identities.

(a) $(A \cap B)^c = A^c \cup B^c$

(b) $A \setminus (B \cup C) = (A \setminus B) \cap (A \setminus C)$

(c) $A \setminus (B \setminus C) = (A \setminus B) \cup (A \cap C)$

(d) $(A \cap B) \setminus (B \cap C) = A \cap (B \setminus C)$

2.5.22. Let A, B, C be three sets. Prove that if $A \setminus B \subseteq C$, then $A \setminus C \subseteq B$.

2.5.23. Let A, B be two subsets of some universal set U. Prove that if $(A \cup B)^c = A^c \cup B^c$, then $A = B$.

2.5.24. Let A, B, C and D be four sets.
Prove that if $A \cup B \subseteq C \cup D$, $A \cap B = \phi$ and $C \subseteq A$, then $B \subseteq D$.

2.5.25. Find the images of the following functions. **Explain** your answer briefly.

(a) $f\colon \mathbb{R} \to \mathbb{R}$, $f(x) = |x|$.

(b) $r\colon \mathbb{R} \to \mathbb{R}$, $r(x) = \dfrac{1}{x^2 + 2}$

(c) $g\colon \mathbb{N} \to \mathbb{R}$, $g(n) = (-1)^n$.

(d) $h\colon \mathbb{Z} \to \mathbb{Z}$, $h(k) = 3k + 1$.

2.5.26. Consider the function $f\colon \mathbb{N} \times \mathbb{N} \to \mathbb{R}$, $f(m, n) = m - n$.

(a) Find $f(3, 5)$ and $f(5, 10)$.

(b) Find two pairs (m, n) for which $f(m, n) = 9$.

(c) What is the image of f ?

2.5.27. What is the image of the function $f\colon \mathbb{Z} \times \mathbb{N} \to \mathbb{R}$, $f(a, b) = \dfrac{a}{b}$? Explain.

2.5.28. What is the image of the function $f\colon \mathbb{Z} \times \mathbb{Z} \to \mathbb{R}$, $f(a, b) = \dfrac{a + b}{2}$? Explain.

2.5.29. A function $f\colon \mathbb{R} \to \mathbb{R}$ is said to be **bounded**, if there is a positive number M, such that $|f(x)| \leq M$ for all $x \in \mathbb{R}$. Which of the following statements are true (for any two functions $f, g : \mathbb{R} \to \mathbb{R}$)? Give a **proof** or a **counterexample**.

(a) If f and g are bounded, then $f + g$ is bounded.

(b) If f and g are bounded, then $f^2 - g^2$ is bounded.

(c) If $f + g$ is bounded, then $f - g$ is bounded.

(d) If f and g are bounded, then $f \cdot g$ is bounded.

(e) If $f \cdot g$ is bounded, then both f and g are bounded.

(f) If $|f| + |g|$ is bounded, then both f and g are bounded.

2.5.30. Is the function $f\colon \mathbb{R} \to \mathbb{R}$, $f(x) = \dfrac{|x+5|}{|x|+5}$ bounded (see Exercise 2.5.29)? Explain.
Do **not** use calculus.

2.5.31. Consider the following subsets of $\mathbb{R}$:

$$A = [1\,,4] \qquad , \qquad B = (-3\,,3) \qquad , \qquad D = \left\{1\,,5\,,\frac{3}{7}\,,\frac{11}{2}\right\} \qquad ,$$

and the function $g\colon \mathbb{R} \to \mathbb{R}$, $g(x) = 2x + 1$.
Compute the following sets (you can use symbols like $\{,\}$, ϕ and the interval notation).

(a) $A \cup B$

(b) $A^c \cap B$

(c) $A \setminus B$

(d) $D \cap \mathbb{Z}$

(e) $D \setminus \mathbb{N}$

(f) $g(D)$

(g) $g(B \cap D)$

(h) $g(A) \cap D$

2.5.32. Consider the intervals $A = (-1, 2)$ and $B = [1, 3]$, and the function $f : \mathbb{R} \to \mathbb{R}$, $f(x) = |x|$.
Compute the following sets (you can use symbols like $\{,\}$, ϕ and the interval notation).

(a) $A \cap B$

(b) $A \cup B$

(c) $A \setminus B$

(d) $(A \cap \mathbb{Z}) \times (B \cap \mathbb{Z})$

(e) $A \cap B \cap \mathbb{Z}^c$
($\mathbb{Z}^c$ is the complement of $\mathbb{Z}$ in $\mathbb{R}$)

(f) $f(A)$

(g) $f(B \cap \mathbb{Z})$

(h) $f(A \cup B) \cap \mathbb{Z}$

2.5.33. Prove that the image of the function $f\colon \mathbb{R} \to \mathbb{R}$, $f(x) = \dfrac{x^2}{1+x^2}$ is the interval $[0\,,1)$.
Do **not** use calculus.

2.5.34. What is the image of the function $f : (0, \infty) \to \mathbb{R}$, $f(x) = \dfrac{4x}{x+1}$?
Prove your answer (and do **not** use calculus).

2.5.35. Prove (without using calculus), that the image of the function $f\colon (0, \infty) \to \mathbb{R}$, $f(x) = x + \dfrac{1}{x}$ is the interval $[2, \infty)$.

2.5.36. Let $f\colon A \to B$ be a function.

(a) Prove that for any two sets $C, D \subseteq A$, we have $f(C \cap D) \subseteq f(C) \cap f(D)$.

(b) Give an example of a function f, and sets C, D, for which $f(C \cap D) \neq f(C) \cap f(D)$.

2.5.37. Is it true that for any function $f\colon A \to B$, and $C, D \subseteq A$, if $C \cap D = \phi$, then $f(C) \cap f(D) = \phi$? Give a proof or a counterexample.

2.5.38. Let $f\colon A \to B$ be a function.

(a) Prove that for any two sets $C, D \subseteq A$, we have $f(C) \setminus f(D) \subseteq f(C \setminus D)$.

(b) Give an example of a function f, and sets C, D, for which $f(C) \setminus f(D) \neq f(C \setminus D)$.

2.5.39. Let $f : X \to Y$ be a function, and $A, B \subseteq X$ two subsets.

(a) Must $f(A \cap B) \subseteq f(A) \cup f(B)$? Why?

(b) Must $f(A \cup B) \subseteq f(A) \cap f(B)$? Why?

2.5.40. Let $f\colon A \to B$ and $C\,, D \subseteq A$.

(a) Is it necessarily true that if $C \subseteq D$, then $f(C) \subseteq f(D)$? Why?

(b) Is it necessarily true that if $f(C) \subseteq f(D)$, then $C \subseteq D$? Why?

2.5.41. Is the set of natural numbers $\mathbb{N}$ (with the usual addition and multiplication of numbers) a field? How about the interval $[0, \infty)$?

2.5.42. Is the set $\mathbb{Q} \cup [-1, 1]$, with the usual addition and multiplication, a field? Explain.

2.5.43. Prove that in a field F, reciprocals are unique. Namely, show that if x is a nonzero element of F, and $x \cdot r_1 = x \cdot r_2 = 1$, then $r_1 = r_2$.

2.5.44. Let F be a field. Prove the following statements. Justify each step in your proofs, and make sure to use only the field axioms, or claims that have been already proved.

(a) For any $x \in F$, $(-1) \cdot x = -x$. (Hint: It is enough to prove that $x + (-1) \cdot x = 0$.)

(b) If $x, y \in F$, and $x \cdot y = 0$, then $x = 0$ or $y = 0$.

(c) If $x, y, z \in F$, and $x + z = y + z$, then $x = y$.

2.5.45. Let $F = \{0, 1, x\}$ be a field with three elements.

(a) Prove that $x + 1 = 0$, and use it to conclude that $x + x = 1$.

(b) Prove that $x \cdot x = 1$.

(c) Draw the addition and multiplication table for F.

2.5.46. Let $F = \{0, 1, a, b\}$ be a field with four elements.

(a) Prove that $a^2 = b$, and that $b^2 = a$.

(b) What are a^3 and b^3?

(c) (**Harder**!) Prove that $1 + 1 = 0$.
(Hint: Assume that $1+1 = a$. What does that imply about $1+a$ and $1+b$? Show that the assumption leads to $a^2 = 0$, which is impossible.)

2.5.47. Here is the addition and multiplication table for a field with **four** elements.

$+$	0	1	a	b
0	0	1	a	b
1	1	0	b	a
a	a	b	0	1
b	b	a	1	0

$\cdot$	0	1	a	b
0	0	0	0	0
1	0	1	a	b
a	0	a	b	1
b	0	b	1	a

Complete with either $0, 1, a$ or b.

$a + b =$ _____ $-b =$ _____ $b^{-1} =$ _____ $a \cdot (1 + b) =$ _____

2.5.48. Is the set $\mathbb{R}^2$, with addition and multiplication defined below a field? Explain.

$$(a, b) + (c, d) = (a + c, b + d) \qquad (a, b) \cdot (c, d) = (ac, bd)$$

2.5.49. Let $A = \{0, 1, x, y\}$ be a set with four elements. Explain why the set A, with addition and multiplication defined by the two tables below, is **not** a field.

$+$	0	1	x	y
0	0	1	x	y
1	1	x	y	0
x	x	y	0	1
y	y	0	1	x

$\cdot$	0	1	x	y
0	0	0	0	0
1	0	1	x	y
x	0	x	0	x
y	0	y	x	1

2.5.50. Show that in any field F, the equation $x^2 = 1$ can have **at most** two solutions.
Can you think of a field in which the equation $x^2 = 1$ has exactly one solution?

2.5.51. Let F be a field in which $1 + 1 = 0$ (there are many such fields, some of which are infinite).
Prove that for any $x \in F$, we have $x = -x$ (i.e., any element in F equals its own negative).

2.5.52. Let $F \subseteq \mathbb{R}$ be a subset, which is also a field (with the usual addition and multiplication inherited from $\mathbb{R}$). We say that F is a **subfield** of $\mathbb{R}$.

(a) Explain why the numbers $3, -4, \frac{1}{2}$ and $\frac{2}{5}$ must be elements of F, while $\sqrt{2}$ and π need not be in F.

(b) Prove that $\mathbb{Q} \subseteq F$.

2.5.53. (**Harder!**) Let F be a subfield of $\mathbb{R}$ (see Exercise 2.5.52).
Prove that if $\sqrt{2} + \sqrt{3} \in F$, then both $\sqrt{2}$ and $\sqrt{3}$ are in F.
(Hint: What is the reciprocal of $\sqrt{2} + \sqrt{3}$?)

2.5.54. Let $K = \{\text{All real numbers of the form } a + b \cdot \sqrt{2}, \text{ where } a, b \in \mathbb{Q}\}$. In this exercise, we show that K is a field (with respect to the usual addition and multiplication of numbers). It is often denoted by $\mathbb{Q}(\sqrt{2})$.

(a) Verify that the field axioms (1), (2), (3) and (5) hold for K. Why is it so easy to check these axioms?

(b) Verify axiom (0).

(c) The hardest part is to check axiom (4). If $x = a + b\sqrt{2}$ is an element of K, what would you expect its negative to be? How can you check your conjecture?

Also, if $x \neq 0$, we would need to show that $x^{-1} = \dfrac{1}{a + b\sqrt{2}}$ is in K. To do that, multiply the numerator and denominator by $a - b\sqrt{2}$. How does that help?

2.5.55. The Order Axioms. Read the following definition and answer the questions below.

Definition: An **ordered field** is a field F, and a subset $P \subseteq F$, called a **positive set**, such that...

(i) If $x, y \in P$, then $x + y \in P$.

(ii) If $x, y \in P$, then $x \cdot y \in P$.

(iii) For any $x \in F$, **exactly one** of the following must hold true: $x \in P$, $-x \in P$ or $x = 0$.

(a) Prove that $1 \in P$.

(b) Prove that if $x \neq 0$, then $x^2 \in P$.

(c) We **define** $a < b$ to mean $b - a \in P$.

Prove that for any $a, b, c, d \in F$, if $a < b$ and $c < d$, then $a + c < b + d$.

(d) Prove that if $c > 0$ and $a < b$, then $ca < cb$.

(e) Prove that if $c < 0$ and $a < b$, then $ca > cb$.

2.5.56. In each part, decide whether the given rule describes a well-defined function. Explain your decision.

(a) $g\colon \mathbb{Z} \times \mathbb{Z} \to \mathbb{Q} \quad , \quad g(a, b) = \dfrac{a}{b}$.

(b) $h\colon \mathbb{R} \to \mathbb{R} \quad , \quad h(x) = |\sqrt{x} - 2|$.

(c) $r\colon \mathbb{R} \to \mathbb{R} \quad , \quad r(x) = \sqrt{|x|} - 2$.

(d) $f\colon \mathbb{Q} \setminus \{0\} \to \mathbb{Q} \quad , \quad f\left(\dfrac{a}{b}\right) = \dfrac{a^2 + b^2}{ab}$.

(e) $p\colon \mathbb{R} \to \mathbb{R} \quad , \quad p(x) = \begin{cases} \sin x & \text{if } \ x \leq 0 \\ |x| & \text{if } \ x \geq 0 \end{cases}$.

(f) $q\colon \mathbb{R} \to \mathbb{R} \quad , \quad q(x) = \begin{cases} \ln(1 - x^2) & \text{if } \ |x| < 1 \\ \ln(x^2 - 1) & \text{if } \ |x| > 1 \end{cases}$.

2.5.57. Suppose that $f\colon [0, 2] \to \mathbb{R}$ and $g\colon [1, 3] \to \mathbb{R}$ are two functions, and define:

$$h\colon [0, 3] \to \mathbb{R} \qquad , \qquad h(x) = \begin{cases} f(x) & \text{if } \ 0 \leq x \leq 2 \\ g(x) & \text{if } \ 1 \leq x \leq 3 \end{cases} .$$

Under what condition(s) will h be a well-defined function? Explain.

Chapter 3

Informal Logic and Proof Strategies

In this chapter, we take a step back to discuss, more generally, the language of mathematics, and some proof techniques and strategies. In the previous chapters, we have seen numerous mathematical notions, theorems, proofs and examples. As you have probably noticed, communicating mathematical arguments and ideas, in a coherent and precise way, is at the core of the subject.

In all our proofs, we used the spoken language (English, in our case), together with mathematical terms and symbols, and we will continue to do so. There is, however, an important difference between using the spoken language in mathematics and in everyday life. In mathematics, we are held to higher standards of precision and accuracy, and cannot tolerate words or phrases that have multiple meanings or are too vague and imprecise.

For instance, suppose you get a phone call from FedEx. The clerk says:

"Your package has arrived. You can come and pick it up next week, on Monday **or** Thursday morning."

What does the clerk mean by '**or**'? It means you'll need to pick one of the two days offered (Monday and Thursday), and come and get your package on the morning of that day. Clearly, picking up the package on **both** days is not a feasible option.

On the other hand, at a family dinner, you might be asked:

"What would you like for dessert? A piece of cake **or** fruit salad?"

This time, you may choose to say: "I'll have a little bit of both.", and no one would think that your answer is unreasonable. Here, the interpretation of the word '**or**' is 'one of the two options, or both', while in the FedEx example, it meant 'exactly one of the options, and not both'.

In mathematics, we cannot allow words to have multiple meanings, and so an interpretation of 'or' must be chosen, and used consistently throughout mathematics. As we have already mentioned before, '**or**' in mathematics means '**one of the two options, or both**'.

There are many other words which are commonly used in mathematics (such as 'and', 'not', 'if-then'), but may have multiple interpretations in real-life. We must clear any possible ambiguity by assigning a single interpretation, to be used everywhere in mathematics. These words are called connectives, and play a fundamental role in communicating mathematical ideas.

However, before going into details regarding connectives and their mathematical interpretation, we discuss another important concept - mathematical statements.

3.1 Mathematical Statements and their Building Blocks

Definition 3.1.1. A **mathematical statement** (or **proposition**) is a sentence that can be either **true** or **false** (in a given context).

In other words, a mathematical statement is a phrase for which **it makes sense to ask whether it is true or false**. It may contain words, symbols, or both.

Examples.

- "The square root of 9 is 3." is a **true** statement.
- "The set $\{\phi\}$ is empty." is a **false** statement (why?).
- "The function $f\colon \mathbb{R} \to \mathbb{R}$, $f(x) = x^2$ is a bounded function." is also a **false** statement (see Exercise 2.5.29).
- "$\mathbb{Z} \subseteq \mathbb{Q}$" is a **true** mathematical statement (since every integer is also a rational number).
  Note how this statement contains only symbols (and no words).
- The phrase "$\mathbb{Z} \cup \left\{\frac{1}{2}, \frac{1}{3}\right\}$" is **not** a mathematical statement (it is just a set of numbers).
  Clearly, it does not make sense to ask whether "$\mathbb{Z} \cup \left\{\frac{1}{2}, \frac{1}{3}\right\}$" is true or false.
- "$5 + 4$" is also **not** a mathematical statement. However, both "$5 + 4 = 9$" and "$5 + 4 < 6$" are statements (the former is **true**, and the latter is **false**).
- "1+1=0" is a **false** statement in the context of real numbers, but is **true** in a field with two elements. This shows how important it is to make sure that the context (for a given statement) is clearly understood.

- Here is an interesting example. Is "$x^2 > 0$" a mathematical statement?

  No, it is not. However, once we assign a value to x, it becomes a statement: If $x = 2$, then $x^2 > 0$ is **true**, and if $x = 0$, then $x^2 > 0$ is **false**.

  Such phrases are called **predicates**, and can be thought of as templates for creating mathematical statements. There is another way to turn predicates into statements (without assigning a value to the variable). This is done using **quantifiers**, discussed below.

Naturally, you might ask yourself, why is the notion of a mathematical statement so important? And why would we ever consider false statements? The answer to the first question is simple. Every claim, proposition and theorem is an example of a mathematical statement. Mathematicians spend most of their time trying to prove (or disprove) statements, and therefore it is crucial that we are able to identify what is a statement, and what is not. It would be meaningless to try and prove a phrase which is not a statement (such as $(a + b)(a - b)$).

To answer the second question, keep in mind that when mathematicians attempt to prove a new claim or theorem, they do not know (yet) whether that statement is true or false. This is why we must be prepared to work with statements that are (potentially) false. There are many other reasons for considering false statements. False statements are often used in proofs and in negating phrases in mathematics. We will touch on some of these ideas later.

Quantifiers.

As mentioned before, the phrase "$x^2 > 0$" is **not** a mathematical statement (unless we assign a value to x). However, we can turn it into a statement as follows.

(i) "**For any** real number x, we have $x^2 > 0$." or (ii) "$x^2 > 0$ **for some** real number x."

Each of the phrases (i) and (ii) are mathematical statements. Statement (i) is false, as not every real number satisfies $x^2 > 0$ (if $x = 0$, then $x^2 = 0$). On the other hand, statement (ii) is true, as $x^2 > 0$ does hold for some real numbers.

Words such as "for any", "every", "for all", "there is", "there exists", "for some", etc. are called **quantifiers**, and are used to turn predicates (sentences with variables) into statements. Quantifiers play an essential role in building mathematical statements. As we have just seen, replacing one quantifier by another can change the meaning of the statement.

Connectives.

A connective is a word (or words) that connects two statements to form a new **compound statement**.

Here is a list of commonly used connectives in mathematics, their "official" name, and examples. Can you figure out which of the examples are true statements?

Connective	**Name**	**Example**
and	conjunction	The smallest natural number is 1 **and** there is no largest natural number.
or	disjunction	For any two sets A and B, we have $A \subseteq B$ **or** $B \subseteq A$.
if-then	implication	For any function $f\colon \mathbb{R} \to \mathbb{R}$, **if** f is bounded, **then** f^2 is also bounded.
iff[1]	equivalence	149 is a prime number **if and only if** 147 is a prime number.
not	negation	**Not** every real number is rational.

[1] iff stands for "if and only if".

Note that the **negation connective** 'not' is used with **one** statement (and not two). Nevertheless, we still refer to it as a connective.

As you can imagine, a compound statement can include multiple connectives and quantifiers. Can you identify them all in the following (true) statement?

"For any $m \in \mathbb{Z}$, if m is not divisible by 3, then there exists a $k \in \mathbb{Z}$, such that $m = 3k+1$ or $m = 3k+2$."

3.2 The Logic Symbols

In mathematics, we often use special symbols to denote logical phrases. These symbols allow us to write compound statements, involving connectives and quantifiers, without using the spoken language. Mathematical statements or predicates are often denoted by an uppercase Latin letter (such as P, Q, R or $P(x), Q(x, y)$ for phrases that have variables). The following table describes logic symbols associated to commonly used connectives and quantifiers.

Connective / Quantifier	**Name**	**Symbol**	**How to use?**
for all / for any / every	the universal quantifier	$\forall$	$\forall x P(x)$ means "For any x, $P(x)$".
there is / there exists / for some	the existential quantifier	$\exists$	$\exists x P(x)$ means "There is an x for which $P(x)$".
not	negation	$\neg$	$\neg P$ means "Not P".
and	conjunction	$\wedge$	$P \wedge Q$ means "P and Q".
or	disjunction	$\vee$	$P \vee Q$ means "P or Q".
if-then	implication	$\Rightarrow$	$P \Rightarrow Q$ means "If P, then Q".
iff (if and only if)	equivalence	$\Leftrightarrow$	$P \Leftrightarrow Q$ means "P if and only if Q".

Examples.

(a) Our first example is non-mathematical. Assume that a set B represents a group of parents, and the set A represents the group of their children. Denote by $P(x, y)$ the phrase "x is y's child". Then we can form the following two statements.

$$\text{(i)} \quad (\forall x \in A)(\exists y \in B)P(x, y) \qquad\qquad \text{(ii)} \quad (\exists y \in B)(\forall x \in A)P(x, y)$$

Statement (i) reads "For every element x in A, there is an element y in B, such that $P(x, y)$". In the context of children and their parents, the statement becomes "For every child in A, there is a person in B, who is a parent of that child". This is a true statement, as every child in A has a parent in B (this follows from the way A and B were formed).

What about statement (ii)? This statement reads "There is an element y in B, such that for any x in A, $P(x, y)$". In our context, it means that "There is a person in B, who is the parent of every child in A", which might be false or true (depending on the sets A and B).

We therefore conclude, that even though statements (i) and (ii) look quite similar, they have different meanings (as (i) is true, and (ii) might be false). The only difference between the two statements is the **order of quantifiers**, and as we can see, this is enough to change their meaning.

(b) Consider the mathematical statement "Any integer is either even or odd". How can we write this statement using the logic symbols?

First, let $E(x)$ denote the phrase "x is even", and $O(x)$ denote "x is odd". Now, the statement can be written as $(\forall x \in \mathbb{Z})(E(x) \vee O(x))$. Note how the disjunction connective is used.

The brackets are added to help with the reading of the statement.

(c) Our last example is a bit more challenging. We wish to represent the following statement with the logic symbols.

P="There is no smallest real number."

We start by representing the statement Q="There is a smallest real number". To do so, we rewrite the statement, without using the word "smallest". This can be done as follows:

Q="There is a real number x that is smaller than or equal to any real number."

Note that the words "there is" and "any" suggest that two quantifiers should be used here. Using only the logic symbols, Q can be represented as

$$Q \;=\; (\exists x \in \mathbb{R})(\forall y \in \mathbb{R})(x \leq y) \; .$$

As P is simply the negation of Q, we have

$$P \;=\; \neg Q \;=\; \neg[(\exists x \in \mathbb{R})(\forall y \in \mathbb{R})(x \leq y)] \; .$$

We will see shortly, how such complex statements can be often made much shorter and simpler.

Notice how quantifiers symbols are used. A quantifier will always appear **before** the variable is mentioned. For instance, a statement like "$|a+1| \leq |a|+1$ for every real number a" must be written as

$$(\forall a \in \mathbb{R})(|a+1| \leq |a|+1) \; .$$

3.3 Truth and Falsity

As we already know, a mathematical statement can be either **true** or **false**. We denote the **truth value** of a statement by T (for true) or F (for false). When a statement is built from other (elementary) statements, through the use of connectives, its truth value will depend on the truth values of the elementary statements.

To clarify, let us examine an explicit case. Assume that P and Q are two statements (that we refer to as 'elementary statements'). Now let R be the (compound) statement $P \vee Q$. What is the **truth value** of R? Is R a **true** or a **false** statement? Obviously, we cannot answer this question without knowing P and Q, and their truth values. In fact, knowing the truth values of P and Q is enough to determine the truth value of $R = P \vee Q$. In other words, the truth value of R will depend **only** on the truth values of P and Q (regardless of the actual content of P and Q).

Since the symbol $\vee$ means 'or', the statement $P \vee Q$ is true when either P, Q or both are true. If P and Q are both false, then $P \vee Q$ must be false as well. This information can be summarized in a **truth table**, as follows:

P	Q	$P \vee Q$
T	T	T
T	F	T
F	T	T
F	F	F

In the first two columns, we list all possible combinations of truth values for the elementary statements P and Q. In the third column, we write the corresponding truth value of $P \vee Q$. For instance, the second last row says that when P is false and Q is true, then $P \vee Q$ is a true statement. The last row says that if P and Q are both false, then $P \vee Q$ is false as well.

A truth table is a convenient way to summarize all possible truth values of a statement, **as a function of the truth values of its building blocks** (the elementary statements). If many elementary statements are used, the table will be of course longer.

Moreover, the truth table above can be seen as **defining the meaning** of the disjunction connective in mathematics.

We can easily construct truth tables for most of the other connectives. Here are the tables for the negation, conjunction and equivalence connectives.

P	$\neg P$
T	F
F	T

P	Q	$P \wedge Q$
T	T	T
T	F	F
F	T	F
F	F	F

P	Q	$P \Leftrightarrow Q$
T	T	T
T	F	F
F	T	F
F	F	T

Note how the truth table for negation has only two rows, as negation is applied to a single statement (instead of two).

There is one connective that has not been included above - the implication (if-then) connective, as its truth table may seem a little bit weird at first. We discuss it separately below.

Implications.

Before presenting the truth table for the implication connective, we discuss the following example:

R="If it is snowing, then the temperature is less than or equal to $0°C$."

This statement is an implication. Its structure is $P \Rightarrow Q$, where P is the phrase "it is snowing", and Q - "the temperature is less than or equal to $0°C$".

Under most circumstances, the above statement is true, and snow is seen at freezing temperature, or below. However, it turns out that under very dry conditions, this statement might be **false**.

What does it mean for R to be false? Well, if snowing **does not** imply temperature of $0°C$ or less, then we could see snow at above-freezing temperatures. In other words, claiming that R is false means that it can snow (namely, P is true), while the temperature is greater than $0°C$ (Q is false).

P	Q	$R = (P \Rightarrow Q)$
T	F	F

What about the rest of the truth table? Is there any other scenario that would make R false? If P and Q are both true (i.e., it snows **and** the temperature is $0°C$ or less), then R holds true. Moreover, if it does

not snow, then regardless of the temperature, the implication is not violated. We conclude that when P is false, the implication $P \Rightarrow Q$ holds true, no matter what the truth value of Q is, and so the full truth table for implications is the following.

P	Q	$P \Rightarrow Q$
T	T	T
T	F	F
F	T	T
F	F	T

It might be hard for you to accept (at first) the last two rows of the truth table, so let us look at a few examples.

Example. The following (or a similar) phrase is often included in rental lease agreements.

"If the tenant severely damages the property, the landlord has the right to terminate the lease."

Again, this is an if-then statement of the form $P \Rightarrow Q$, where P is "the tenant severely damages the property" and Q is "the landlord has the right to terminate the lease".

Can the landlord terminate the lease even when the tenant does not severely damage the property? Yes he can, for other reasons (such as not paying rent). The implication is **not** violated when P is false and Q is true. Clearly, having both P and Q true or both false (e.g., the tenant damages and the lease is terminated, or the tenant does not damage and the lease is not terminated) is consistent with the above phrase.

The only possible scenario in which the above clause is not followed, is when the tenant does damage the property (P is true), but the landlord cannot terminate the agreement (Q is false). In all other cases, $P \Rightarrow Q$ is true, which is consistent with the truth table for implications.

Example. In everyday language, we often use the combination if-then to describe **cause and effect**. For instance:

"If you do not brush your teeth, you will have many cavities."

This sentence describes cause and effect: Not brushing your teeth will have a direct effect on them. But cause and effect make little sense in mathematics. Consider, for example, the following statement:

"If money grows on trees, then cats have five legs."

This is a **true** statement (both P, the hypothesis, and Q, the conclusion, are false), but there is no cause and effect involved here. It is the logical structure of the statement, and the truth table above, that make it true. We often say that if-then statements, in which the hypothesis is false, are **vacuously true**.

Example. We end this discussion with one last example. Suppose you are borrowing your friend's bicycle for a one-day biking trip. When you pick up the bike, your friend says:

"If you damage my bike, you will have to pay for it!"

Again, this statement has the structure $P \Rightarrow Q$, with P being "you damage my bike", and Q - "you will have to pay for it!". Now imagine that when you come back from the trip, and return the undamaged bike to your friend, he says "Now I want you to pay for it!". Surprised, you reply "How come? There isn't even a scratch on your bike. What do I need to pay for?". Your friend keeps insisting that you pay, arguing that according to the truth table above, asking you to pay even when the bike is undamaged is consistent with the if-then statement (as if P is false, and Q is true, the implication $P \Rightarrow Q$ is considered true!). What is going on here?

The problem lies in the way we use the words "if-then" in everyday language. Occasionally, "if-then" would actually mean "if-and-only-if" in ordinary day-to-day language. When your friend said the above phrase, you probably interpreted (as most people would) as follows:

"If you damage my bike, you will have to pay for it!
But if there is no damage, you will not need to pay at all."

In other words, you interpreted the phrase as an if-and-only-if statement ("You will have to pay for my bike if and only if you damage it!"), and consequently were very surprised when your friend demanded that you pay.

This kind of confusion can cause serious problems in mathematics. Precision and accuracy are crucial in mathematical arguments and proofs, and so we have to make sure that all interpret statements in the same way. Therefore, **we must use "if-then" when we actually mean if-then, and "if-and-only-if" when our statement describes an equivalence**.

A Remark on Quantifiers.

In this section we provided truth tables for the logical connectives, which allow us to find the truth value of a compound statement. But what about statements that involve quantifiers? For instance, can we construct a truth table for the following statement?

"For any prime number p greater than 3, $p^2 - 1$ is divisible by 24."

This statement has the form $(\forall p \in A)S(p)$, where A is the set of prime numbers greater than 3, and $S(p)$ is the phrase "$p^2 - 1$ is divisible by 24". We might be tempted to construct a truth table for this statement, by evaluating the truth value of $S(p)$ for every prime number p greater than 3. The first few rows of such a table will look as follows.

p	$S(p)$	Truth Value of $S(p)$
5	$5^2 - 1$ is divisible by 24	T
7	$7^2 - 1$ is divisible by 24	T
11	$11^2 - 1$ is divisible by 24	T
$\vdots$	$\vdots$	$\vdots$

However, it is impossible to create a complete table, as there are infinitely many primes greater than 3 (a fact that we will prove soon). Therefore, truth tables are rarely used to break down statements that involve quantifiers, and other techniques must be used instead (such as general arguments and proofs). We need to remember though, the way quantifiers are interpreted in mathematics.

$(\forall x)P(x)$	is true when $P(x)$ holds true **for any** value of x.
$(\exists x)P(x)$	is true when $P(x)$ holds true **for some** value (or values) of x.

3.4 Truth Tables and Logical Equivalences

The truth tables for the logical connectives are often referred to as the **elementary truth table**. These tables define the **meaning** of the connectives in mathematics. However, truth tables can be created for longer and more complex statements, involving multiple quantifiers. Here is an example.

Example. Consider the following statement R:

$$R \;=\; [P \wedge (\neg Q)] \Rightarrow [(\neg P) \vee Q] \;.$$

Here, P and Q are elementary statements (whose content or structure is not given), and R is a compound statement, built from P, Q, and logical connectives. The brackets are used to indicate the order in which the connectives should be applied. The truth value of R will depend on the truth values of P and Q, and we would like to describe this dependency in a table. Our truth table will have four rows, one for each possible combination of truth values for P and Q.

P	Q	R
T	T	?
T	F	?
F	T	?
F	F	?

Our task is to complete the third column of the above table, namely - to find the truth values of R. However, as R is a relatively complex statement, with several connective, finding the truth values requires some work. We must carefully analyze the various parts of R and the way they are put together, and we can do so by adding a few more columns (that we call 'helpers') to our truth table.

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

Without too much effort, we can complete the 'helper' columns (in order, from left to right), by referring to the elementary truth tables of the connectives $\neg$, $\wedge$ and $\vee$. For instance, the truth values of $\neg Q$ are obtained by reversing those of Q. Once we have the truth values for $P \wedge (\neg Q)$ and $(\neg P) \vee Q$, we can use them, together with the implication's truth table, to find the desired truth values of R. Here is the completed table.

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

We see that R is false when P is true and Q is false. In all other cases, R is a true statement. Note that the 'helper' columns are optional, and we add as many as we need in order to complete the rightmost column.

Example. Let us construct the truth table of $S = Q \vee [(\neg P) \Leftrightarrow (\neg Q)]$.
We add the columns for $\neg P$, $\neg Q$ and $(\neg P) \Leftrightarrow (\neg Q)$ as helpers.

P	Q	$\neg P$	$\neg Q$	$(\neg P) \Leftrightarrow (\neg Q)$	S
T	T	F	F	T	T
T	F	F	T	F	F
F	T	T	F	F	T
F	F	T	T	T	T

Note that the two statements above (R and S) have the same truth tables. Namely, R and S always have the same truth value regardless of the values of P and Q. Such statements are said to be **logically equivalent**.

Definition 3.4.1. Two statements are said to be **logically equivalent** if they always have the same truth value. In particular, two statements with the same truth table are logically equivalent.

Roughly speaking, two logically equivalent statements are statements that have the same meaning (they say the same thing, though in a different way). This is much like two equivalent algebraic expression, such as $4x^2 - 6y$ and $(2x)^2 + (-2y) \cdot 3$. These expressions, although different, always produce the same value (for any choice of numbers x and y), and so we treat them as being equal (or equivalent) to each other.

Logical equivalences can be used to simplify statements (the same way algebraic identities are used to simplify expressions). The following is a list of important and commonly used logical equivalences.

Proposition 3.4.2. Let P and Q represent two statements. Then the following pairs are logically equivalent.

(a) $\neg(P \wedge Q)$ and $(\neg P) \vee (\neg Q)$

(b) $\neg(P \vee Q)$ and $(\neg P) \wedge (\neg Q)$

(c) $\neg(P \Rightarrow Q)$ and $P \wedge (\neg Q)$

(d) $P \Leftrightarrow Q$ and $(P \Rightarrow Q) \wedge (Q \Rightarrow P)$

(e) $P \Rightarrow Q$ and $(\neg Q) \Rightarrow (\neg P)$

Proof (partial). Proving the proposition is straightforward. We simply create truth tables for each pair, and verify that they are indeed the same.
For instance, here are the truth tables (including one helper column in each) for the statements in part (c).

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

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

The rightmost columns in the truth tables are the same, which proves the equivalence of $\neg(P \Rightarrow Q)$ and $P \wedge (\neg Q)$. To prove part (e), we construct the following two truth tables. Note that the table on the left is just the truth table for the if-then connective.

P	Q	$P \Rightarrow Q$
T	T	T
T	F	F
F	T	T
F	F	T

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

And again, as the rightmost columns are identical, the statements $P \Rightarrow Q$ and $(\neg Q) \Rightarrow (\neg P)$ are logically equivalent.

The proofs of the other parts are left as an exercise.

□

Remarks.

- Equivalence of statements can be proved by constructing truth tables and comparing them, and that is what we did to prove parts (c) and (e) of the proposition above. However, it might be useful to try and create real-life (non-mathematical) examples, that illustrate the equivalence. This cannot serve as a mathematical proof, but it can strengthen our intuition (and confidence) as to why an equivalence is valid.

  For instance, denote by P the phrase "being rich" and by Q the phrase "being happy". Then the statement $\neg(P \Rightarrow Q)$ becomes "Being rich **does not** imply being happy". The statement $P \wedge (\neg Q)$, however, translates into "One can be rich and unhappy". It is quite evident (at least informally), that the two statements have the same meaning. This illustrates the equivalence of the statements in part (c) of Proposition 3.4.2.

  For part (e), consider the following if-then statement

  "If a person has a driver's license, s/he is at least 16 years old."

The statement has the structure $P \Rightarrow Q$, where P is "a person has a driver's license", and Q is "being at least 16 years old". Now, if we convert $(\neg Q) \Rightarrow (\neg P)$ into words, we get

"If a person is not at least 16 years old, then s/he does not have a driver's license."

Again, it is not hard to observe the equivalence of the two statements (they both say the exact same thing), which supports part (e) of the proposition.

- The equivalence in part (e) of Proposition 3.4.2 is of great importance, as it provides us with a proof technique, that is commonly used in mathematics. The statement $(\neg Q) \Rightarrow (\neg P)$ is called the **contrapositive** of $P \Rightarrow Q$, and as we have seen, it is equivalent to $P \Rightarrow Q$. This means that proving an if-then statement can be done by proving its contrapositive. Here is an example.

  Example. Let $n \in \mathbb{Z}$. If n^3 is odd, then n is odd.

  How can we prove this statement? An odd number has the form $2k+1$ (for some integer k), so we can start by writing $n^3 = 2k+1$. Our task is to show that n is odd, so we might try to solve for n, which gives $n = \sqrt[3]{2k+1}$. However, it seems like writing n as $\sqrt[3]{2k+1}$ is not going to be of any help, and we are stuck!

  Fortunately, the contrapositive of the implication "If n^3 is odd, then n is odd." is "If n is even, then n^3 is even." (an integer which is not odd is even), and is much easier to prove.

  Proof. We prove the contrapositive "If n is even, then n^3 is even.". An even number is a multiple of 2, and so $n = 2k$ for some $k \in \mathbb{Z}$. Therefore,

  $$n^3 = (2k)^3 = 8k^3 = 2 \cdot 4k^3 \ .$$

  We now see that n^3 is an integer multiple of 2, and hence is even, as needed. □

  This example shows how sometimes, implications which are hard to prove directly (or even impossible), have a simple proof for their contrapositive. We will see more examples of using the contrapositive method later in this chapter.

- The following two equivalences involve quantifiers, and so we are unable to verify them with a truth table. We state the equivalences without proof.

- ⋆ The statements $\neg[(\forall x)P(x)]$ and $(\exists x)[\neg P(x)]$ are logically equivalent.
- ⋆ The statements $\neg[(\exists x)P(x)]$ and $(\forall x)[\neg P(x)]$ are logically equivalent.

These equivalence will be used in the next section, and even though we do not provide a formal proof, they are quite intuitive. Try to say them in words. Can you see why each pair has the exact same meaning?

We end this section with the definition of a tautology and a contradiction.

Definition 3.4.3. A **tautology** is a statement that is always true (regardless of the truth values of its elementary statements). A **contradiction** is a statement that is always false.

Examples.

(a) The statement $(P \wedge Q) \Rightarrow (P \vee Q)$ is a tautology (which makes sense: If P and Q are both true, then P or Q is true is well). To prove it, we construct the truth table, and observe that the rightmost column contains only T's.

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

(b) The statement $P \Leftrightarrow (\neg P)$ is a contradiction, as the following (short) truth table suggests.

P	$\neg P$	$P \Leftrightarrow (\neg P)$
T	F	F
F	T	F

(c) Tautologies and contradictions may include quantifiers. In these cases, we may not be able to use truth tables. For instance, the statement $(\forall x)[P(x) \vee [\neg P(x)]]$ is a tautology (which follows from the fact that $Q \vee (\neg Q)$ is a tautology).

3.5 Negation

The negation of a statement is its opposite, usually obtained by adding the word '**not**'.

Definition 3.5.1. The negation of a statement P is the statement $\neg P$ ('**not** P').

For instance, the negation of "all cats are black" is "**not** all cats are black", and the negation of "being tall implies having back problems" is "being tall **does not** imply having back problems". Mathematical statements can also be negated by adding the word 'not' (or 'no'). For example, the negation of

$$P = \text{"There is a set } A\text{, for which } A \in A\text{."}$$

is the statement

$$\neg P = \text{"There is } \textbf{no} \text{ set } A\text{, for which } A \in A\text{."}$$

In mathematics, we need to work with negations and use them in various arguments and proofs regularly. However, forming a negation by simply adding the word 'not' (or by putting the negation symbol $\neg$ in front of it) may be insufficient, and mathematicians often try to simplify negations and make them as explicit as possible. For instance, the statement "not all cats are black" can be restated as "some cats are not black". Here is a more mathematical example.

Example 3.5.2. Let $U = \mathbb{R}$ be the set of all real numbers, and consider the following statement:

$$Q = \text{"For all } a, b \in U\text{, if } a \cdot b = 0\text{, then } a = 0 \text{ or } b = 0\text{."}$$

This is a true statement. In fact, the statement can be proved from the field axioms, and thus remains true as long as U is a field. There are, however, mathematical 'universes[1]' in which Q is false, and hence its negation $\neg Q$ is true.

In other words, there are cases in which the statement

$$\neg Q = \text{"}\textbf{Not} \text{ for all } a, b \in U\text{, if } a \cdot b = 0\text{, then } a = 0 \text{ or } b = 0\text{."}$$

is valid. To better understand the negated statement, we simplify it as much as we can. Ideally, we try to restate the negation without using the word 'not' or 'no'. To begin with, we observe that if not all a, b satisfy the if-then phrase, then at least one pair of a, b must violate it. We can therefore restate the negation as

[1] We do not attempt to define precisely what we mean by 'universes'. In a ring, an algebraic structure you may encounter in a more advanced mathematics course, the statement Q may be false.

$$\neg Q = \text{“There exist } a, b \in U, \text{ for which } a \cdot b = 0 \text{ does } \textbf{not} \text{ imply } a = 0 \text{ or } b = 0.\text{”}$$

Moreover, if $a \cdot b = 0$ does not imply $a = 0$ or $b = 0$, then $a \cdot b$ can be zero, while none of a, b are zero (an implication is false when the hypothesis is true and the conclusion is false). Hence, we can write:

$$\neg Q = \text{“There exist } a, b \in U, \text{ such that } a \cdot b = 0, \ a \neq 0 \text{ and } b \neq 0.\text{”}$$

Note how we were able to restate $\neg Q$ without using any words of negation (such as ‘not’, ‘no’, ‘it is false that’, etc.). By doing so, we have made the negation explicit, simpler, and easier to use in arguments and proofs.

In the example above, we started with an unsimplified negation, and gradually replaced it by equivalent statements, which are simpler. This process can be done more mechanically, by referring to the logical equivalences mentioned in the previous section. We illustrate this procedure by re-doing Example 3.5.2.

Example 3.5.3. We negate and simplify the statement

$$Q \ = \ (\forall a \in \mathbb{R})(\forall b \in \mathbb{R})[(a \cdot b = 0) \Rightarrow ((a = 0) \vee (b = 0))].$$

Using the equivalence of $\neg[(\forall x)P(x)]$ and $(\exists x)[\neg P(x)]$, we have

$$\neg Q \ = \ (\exists a \in \mathbb{R})(\exists b \in \mathbb{R})[\, \neg[(a \cdot b = 0) \Rightarrow ((a = 0) \vee (b = 0))] \,].$$

Next, we apply the equivalence of $\neg(P \Rightarrow Q)$ and $P \wedge (\neg Q)$ (see part (c) of Proposition 3.4.2), to get

$$\neg Q \ = \ (\exists a \in \mathbb{R})(\exists b \in \mathbb{R})[(a \cdot b = 0) \wedge [\, \neg[(a = 0) \vee (b = 0)] \,]].$$

Now, we apply the equivalence of $\neg(P \vee Q)$ and $(\neg P) \wedge (\neg Q)$ on the “$a = 0$ or $b = 0$” part:

$$\neg Q \ = \ (\exists a \in \mathbb{R})(\exists b \in \mathbb{R})[\, (a \cdot b = 0) \wedge [\neg(a = 0)] \wedge [\neg(b = 0)] \,].$$

Finally, we replace $\neg(a = 0)$ and $\neg(b = 0)$ by $a \neq 0$ and $b \neq 0$ (after all, $\neq$ means ‘not equal’):

$$\neg Q \ = \ (\exists a \in \mathbb{R})(\exists b \in \mathbb{R})[\, (a \cdot b = 0) \wedge (a \neq 0) \wedge (b \neq 0) \,].$$

As you can see, we managed to express the negation of Q without using the negation symbol ‘$\neg$’. We obtained the exact same statement as the one we got in Example 3.5.2, except that it is written with the logic symbols (instead of words).

An important use of negation is with definitions, as we illustrate in the next example. Showing that an object **does not** satisfy a definition is the same as showing that its negation holds true.

Example 3.5.4. Recall the following definition of a bounded function:[2]

"A function $f\colon \mathbb{R} \to \mathbb{R}$ is **bounded**, if there is an $M \in \mathbb{R}$, such that $|f(x)| \leq M$ for all $x \in \mathbb{R}$."

Using the logic symbols, we can rewrite the definition as follows.

"A function $f\colon \mathbb{R} \to \mathbb{R}$ is **bounded**, if $(\exists M \in \mathbb{R})(\forall x \in \mathbb{R})(|f(x)| \leq M)$."

Now suppose we want to prove that a function f is **unbounded** (i.e., not bounded). To do so, we need to show that the above definition does not hold (or, equivalently, that its negation is satisfied). Using the negation symbol, we can write:

"A function $f\colon \mathbb{R} \to \mathbb{R}$ is **unbounded**, if $\neg(\exists M \in \mathbb{R})(\forall x \in \mathbb{R})(|f(x)| \leq M)$."

However, by applying the logical equivalences involving quantifiers (see page 67, we can avoid the negation symbol, and restate the above phrase as follows.

"A function $f\colon \mathbb{R} \to \mathbb{R}$ is **unbounded**, if $(\forall M \in \mathbb{R})(\exists x \in \mathbb{R})(|f(x)| > M)$."

This is an explicit and more practical definition for an unbounded function. Note how the inequality $|f(x)| \leq M$ was replaced by its negation, $|f(x)| > M$.

We end this section with one final note. Negating a statement reverses its truth value, and so a statement and its negation will never be both true or both false. Keep that fact in mind, and use it to check your negations.

Examples.

(a) The (simplified) negation of $R = (\exists x \in (-3,2) \cap \mathbb{Z})\left(x^2 < \frac{1}{2}\right)$ is $\neg R = (\forall x \in (-3,2) \cap \mathbb{Z})\left(x^2 \geq \frac{1}{2}\right)$. As $0^2 < \frac{1}{2}$, the statement R is true, and $\neg R$ is false.

(b) The negation of $P = (\forall x \in \mathbb{R})[(x^2 + 1 < 0) \Rightarrow (15 < 5)]$ is $\neg P = (\exists x \in \mathbb{R})[(x^2 + 1 < 0) \wedge (15 \geq 5)]$. For any real number x, $x^2 + 1 > 0$, and hence the implication $(x^2 + 1 < 0) \Rightarrow (15 < 5)$ is **true**. We conclude that P is true, and $\neg P$ is false.

3.6 Proof Strategies

These notes are meant to introduce the notion of a mathematical proof, and provide examples of proofs, and guidance and tools for creating them. However, we have not discussed yet (at least, not explicitly) the

[2] Note that $|f(x)| \leq M$ is equivalent to $-M \leq f(x) \leq M$. Geometrically, f is bounded if its graph lies between the horizontal lines $y = M$ and $y = -M$ (for some M).

general question of how to **construct** (or generate) mathematical proofs. It would be nice to have a clear set of guidelines, procedures or algorithms that we can use to generate proofs of mathematical statements. In Chapters 1 and 2 we have seen quite a few proofs, but with very little similarities. For instance, the proofs of the quadratic formula and inequalities were quite computational, while proofs involving sets were mostly done by showing two inclusions. Proving statements about fields required careful use of the field axioms, and sometimes an elimination strategy (such as in Claim 2.3.3).

How would one know the right way to tackle a claim, theorem or proposition? Is there a recipe that can be executed to generate a proof? We do have algorithms for solving quadratic equations and systems of linear equations, and for finding extreme values of a function. Is there also an algorithm for generating proofs?

Unfortunately, no. In fact, I would rather say: **Fortunately**, no!

If there was an algorithm, or an explicit set of guidelines for generating proofs, mathematics would become easier, but also a mechanical and procedural field. We could then probably have computers generate proofs for many mathematical statements, and theoretical mathematics will become as dull as adding fractions (or multiplying two-digit numbers). Many are attracted to mathematics because of the **creative nature** of this field. Some compare mathematics to the arts, and see beauty in the process of discovering and creating mathematical proofs. It is **the lack** of an algorithm, or procedures, for creating proofs that makes mathematics so interesting (and often challenging).

For this reason, proving theorems can be difficult, at times frustrating, and be a long process. But the journey can nevertheless be exciting, enlightening, and transformative. Discovering a proof, especially after working hard and implementing new ideas, can be extremely rewarding. The need to be creative, produce original ideas, and looking at things in a unique way is at the heart of mathematics. For this reason, proving mathematical statements is far from applying a cook-book recipe, and requires effort, persistence, and risk-taking.

However, even without step-by-step instructions, there are some proof strategies that you should be aware of. These strategies are not algorithms, and cannot be applied blindly to create proofs (even after a suitable strategy has been chosen). Instead, they can be seen as commonly-used approaches, or 'modes of thinking'. Choosing an appropriate strategy for proving a given statement can still be challenging, and even after doing so, implementing that strategy can by tricky. With experience and practice, you will become better in deciding how to tackle mathematical statements and proofs.

Direct Proof.

Most statements in mathematics can be regarded as implications; that is, as having the form $P \Rightarrow Q$. P

can be seen as the assumptions (or hypotheses), while Q is what needs to be proved. Some mathematical statements are 'if-and-only-if' statements, but these can be naturally regarded as two implications (as $P \Leftrightarrow Q$ is logically equivalent to $(P \Rightarrow Q) \wedge (Q \Rightarrow P)$).

In a direct proof, we simply assume that our hypotheses P hold true, and derive the statement Q, while relying on P and previously established theorems (and definitions). Here is a direct proof of a divisibility test you might have seen in your elementary school years.

Example 3.6.1. Let n be a three-digit positive integer (i.e., $n \in \mathbb{N}$ and $100 \leq n \leq 999$). If the sum of digits of n is divisible by 3, then n is divisible by 3.

Proof. In this example, the hypothesis (P) is the assumption that n is a three-digit positive number, whose sum of digits is divisible by 3. Denote the digits of n by x, y and z (for instance, if $n = 729$, then $x = 7$, $y = 2$ and $z = 9$). We assume that $x + y + z$ is divisible by 3.

Our task is to prove that n itself is divisible by 3 (this is our Q). To do so, we observe that

$$n = 100x + 10y + z$$

(as x is the hundreds digit, y is the tens, and z is the ones). Consequently, we can write

$$n = (99x + x) + (9y + y) + z = (99x + 9y) + (x + y + z) = 3 \cdot (33x + 3y) + (x + y + z) \ .$$

We are told that $x + y + z$ is divisible by 3, that is, $x + y + z = 3k$ for some integer k (recall Definition 1.4.1). Therefore,

$$n = 3 \cdot (33x + 3y) + 3k = 3 \cdot (33x + 3y + k) \ ,$$

which implies that n is divisible by 3, as needed.

□

Proof by Contrapositive.

We briefly recall the method of proof by contrapositive, that has already been mentioned on page 66.

The statement $P \Rightarrow Q$ is logically equivalent to $(\neg Q) \Rightarrow (\neg P)$, which is called **the contrapositive** of $P \Rightarrow Q$. Consequently, an implication can be proved by proving its contrapositive. Using this strategy, we assume that the conclusion Q is **false**, and use it to prove that the hypothesis P is **also false**.

In some cases, proving the contrapositive of an implication is much easier than proving the original implication. Here is an example.

Example 3.6.2. Consider the function $f(x) = \dfrac{x}{x+1}$, and let $a, b \neq -1$. Prove that if $a \neq b$, then $f(a) \neq f(b)$.

Here, our hypothesis P is $a \neq b$, and we need to prove that $f(a) \neq f(b)$ (or, more explicitly, that $\frac{a}{a+1} \neq \frac{b}{b+1}$). In this case, producing a direct proof would mean that we begin with the inequality $a \neq b$, and manipulate it to get $\frac{a}{a+1} \neq \frac{b}{b+1}$. This is possible to do, but there are some advantages to using the contrapositive method here. First, the contrapositive will involve **equalities** (rather than inequalities), as the negation of $x \neq y$ is $x = y$. Secondly, the contrapositive will require us to **simplify an equation**, a process that we are well familiar with.

Proof. We prove the contrapositive:

$$\text{“If } f(a) = f(b), \text{ then } a = b.\text{”}$$

If $f(a) = f(b)$, then $\frac{a}{a+1} = \frac{b}{b+1}$. We cross multiply and simplify, to get

$$a(b+1) = b(a+1) \qquad \Rightarrow \qquad ab + a = ba + b \qquad \Rightarrow \qquad a = b \ ,$$

and the proof is completed.

□

Proof by Contradiction.

Proof by contradiction is one of the most important proof methods in mathematics. It is widely used in all areas of mathematics, and at all levels. To prove a statement by contradiction, we assume that **what needs to be proved is false**, and show that this assumption leads to a contradiction. In other words, to prove an implication $P \Rightarrow Q$ by contradiction, we assume that both P and $\neg Q$ are true, and try to derive a contradiction. What is it that we can contradict? Anything that is known to be true, such as a definition, a previously established theorem, or the hypothesis P.

Informally, when we prove a statement by contradiction, we begin by wondering “what if the thing we need to prove is wrong?”. The proof itself will sort of answer this question, by showing that if what needs to be proved is false, then some other fact, that is known to be true, is also false. As contradictions are not acceptable in mathematics, we conclude that what needs to be proved must be true.

Let us clarify by discussing some examples.

Example 3.6.3. Prove that the equation $3x^2 - 7x + 1 = 0$ has no **rational** solutions.

A natural way to prove, in mathematics, that something does not exist, is to assume that it does, and derive a contradiction. Note that the quadratic $3x^2 - 7x + 1 = 0$ does have real solutions (according to Theorem 1.1.1). We need to show that there are **no rational solutions** to this equation.

Proof. Assume, by contradiction, that there exists a rational solution, r, to the given equation:

$$3r^2 - 7r + 1 = 0 \ .$$

That is, r is a number of the form $\frac{m}{n}$, with m and n, integers, and $n \neq 0$. We assume that the fraction $\frac{m}{n}$ is in lowest terms, meaning that it is completely simplified (for instance, $\frac{2}{3}$ and $\frac{5}{12}$ are in lowest terms, while $\frac{14}{21}$ and $\frac{18}{51}$ are not). More formally, we assume that m and n have no common divisors other than ± 1.

As $r = \frac{m}{n}$, we can replace r with $\frac{m}{n}$ in the equation above, and multiply by n^2, to get

$$3\left(\frac{m}{n}\right)^2 - 7\left(\frac{m}{n}\right) + 1 = 0 \qquad \Rightarrow \qquad 3m^2 - 7mn + n^2 = 0 \ .$$

Since both m and n are integers, each can be either an even or an odd number. Let us check all possible combinations of even/odd.

- If both m and n are odd, then the numbers $3m^2$, $7mn$ and n^2 are all odd (any product of odd numbers is odd). Therefore, the left-hand side, $3m^2 - 7mn + n^2$ must be an odd number (as the sum of three odd numbers is odd). This contradicts the fact that $3m^2 - 7mn + n^2$ is zero, which is an even number.

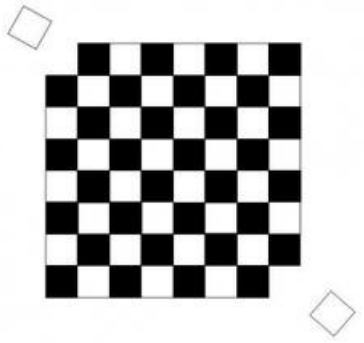
- If m is even and n is odd, then both $3m^2$ and $7mn$ are even numbers, while n^2 is odd. Again we conclude that $3m^2 - 7mn + n^2$ must be odd, which leads to a contradiction. A similar argument shows that if n is even and m is odd, we also get a contradiction.

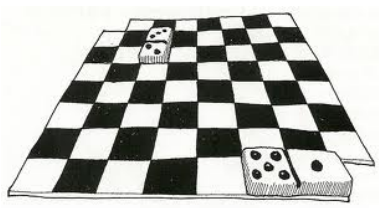
- If m and n are both even, it contradicts the fact that the fraction $\frac{m}{n}$ is in lowest terms.

We see that any possible scenario leads to a contradiction, which means that our initial assumption - the existence of a rational solution to the given quadratic, must be false. Therefore, there is no rational solution to $3x^2 - 7x + 1 = 0$, and the proof is complete. □

The next is a well-known (and 'less mathematical') example.

Example 3.6.4. Is it possible to cover a regular checkerboard, with two opposite corners removed, with regular dominoes? Each domino can be placed either horizontally or vertically, and covers exactly two squares (see diagram).





You might want to try and experiment a bit before continue reading.

There are two possible answers to this question. If it is possible to cover the checkerboard, then one can prove it by **demonstrating** such a covering. But what if it is not possible? How can one prove that? Trying for hours (or days) and failing to find a cover cannot serve as a proof. On the other hand, checking any possible way to place dominoes on the checkerboard may take forever.

We claim that it is indeed impossible to produce a covering for the checkerboard, and we prove it by contradiction. Namely, we start by assuming that we can cover the board with dominoes, and show that this assumption leads to a contradiction. The main observation we use in the proof, is that a single domino, when placed on the board, must cover exactly one white and one black square.

Answer: No, it is not possible to cover the checkerboard with dominoes.

Proof. Assume there is a way to cover the board with dominoes. Since our board has 62 squares (as two corners were removed), 31 dominoes are needed for the covering. Each domino must cover a black and a white square, and so overall, 31 black and 31 white squares would be covered. But this is impossible! The two removed corner-squares have the same color, and hence the number of white squares is not the same as the number of black squares (we have 30 squares of one color, and 32 of the other). This is a contradiction, and thus there is no way to cover our board with dominoes. □

Here is another non-existence proof example.

Example 3.6.5. There are no **natural numbers** x, y for which $x^2 - y^2 = 1$.

Proof. Again, we assume, by contradiction, that there are $x, y \in \mathbb{N}$ for which $x^2 - y^2 = 1$.

We then factor the left-hand side, and write the equation as $(x + y)(x - y) = 1$. The numbers $x + y$ and $x - y$ are integers, whose product is 1, which means that both $x + y$ and $x - y$ are equal to 1 (note that $x + y > 0$, which rules out the possibility of $x + y = x - y = -1$).

But the only solution to $x + y = x - y = 1$ is $x = 1$ and $y = 0$, which contradicts the fact that y is a natural number. Therefore, the equation $x^2 - y^2 = 1$ has no natural solutions. □

We end this section with a proof of a famous theorem – the infinitude of prime numbers – and we show Euclid's proof of the theorem by contradiction. Recall that a prime number p is a natural number, greater than one, whose only positive divisors are 1 and p.

Theorem 3.6.6. *There are infinitely many prime numbers.*

Proof. Suppose, by contradiction, that there are finitely many prime numbers, and denote them by $p_1, p_2, p_3, \ldots, p_k$ (that is, we assume that there are exactly k prime numbers). Define

$$M = p_1 \cdot p_2 \cdot p_3 \cdot \cdots \cdot p_k + 1 \ .$$

The number M is greater than any of the p_i's, and hence is not a prime number itself. However, any natural number (greater than 1) is divisible by a prime number (a fact that will be proved in the next chapter), and so M must be divisible by one of the p_i's. Now, as

$$M - p_1 \cdot p_2 \cdot p_3 \cdot \cdots \cdot p_k = 1 \ ,$$

we get that 1 is divisible by one of the p_i's, which is a contradiction (a natural number cannot be divisible by a larger number). We thus conclude that there are infinitely many prime numbers. □

3.7 Exercises for Chapter 3

3.7.1. Given a real number x, let A be the phrase "$\frac{1}{2} < x < \frac{5}{2}$", B the phrase "$x \in \mathbb{Z}$", C the phrase "$x^2 = 1$", and D the phrase "$x = 2$". Which of the following are true for all $x \in \mathbb{R}$? Explain.

(a) $A \Rightarrow C$

(b) $(A \wedge B) \Rightarrow C$

(c) $D \Rightarrow [A \wedge B \wedge (\neg C)]$

3.7.2. For each statement, write the meaning in English and decide whether it is true or false (x and y represent real numbers). Explain your decision briefly.

(a) $\forall x \forall y (x \geq y)$

(b) $\exists x \exists y (x \geq y)$

(c) $\exists y \forall x (x \geq y)$

(d) $\forall x \exists y (x \geq y)$

(e) $\forall x \exists y (x^2 + y^2 = 1)$

(f) $\exists x \forall y (x^2 + y^2 = 1)$

3.7.3. Express the following statements using the logic symbols, and decide, for each, whether it is true or false. Explain your decision briefly.

(a) There is a smallest positive real number.

(b) Every integer is a product of two integers.

(c) The equation $x^2 + y^2 = 1$ has a solution (x, y) in which both x and y are natural numbers.

(d) Every real number can be written as a difference of two positive real numbers.

3.7.4. Let S, T and U be three sets.

Then the statement $S \cap T \subseteq U$ can be written, using the logic symbols, as follows.

$$(\forall x)[((x \in S) \wedge (x \in T)) \Rightarrow (x \in U)] \quad .$$

(a) Write the statement $S \cap T \not\subseteq U$ using the logic symbols (but without the symbol '$\neg$').

(b) Write the statement $S \subseteq T \cup U$ and its negation using the logic symbols.

3.7.5. Let $P(x)$ be the assertion "x is positive", and let $Q(x)$ the assertion "$x^2 > x$".

(a) Is the statement $(\forall x \in \mathbb{R})[P(x) \Rightarrow Q(x)]$ true or false? Why?

(b) Is the statement $[(\forall x \in \mathbb{R})P(x)] \Rightarrow [(\forall x \in \mathbb{R})Q(x)]$ true or false? Why?

3.7.6. Consider the following two statements:

$R =$ "For any real number x, there is a real number y, such that $x+y<1$."

$S =$ "There is a real number y, such that for all real numbers x, we have $x+y<1$."

(a) Write both statements using the logic symbols.

(b) Write the negations of R and S. Use the logic symbols, but do not use the symbols '$\neg$' or '$\not<$'.

(c) Is R a **true** or a **false** statement? Is S a **true** or a **false** statement? Explain.

3.7.7. For what **value (or values)** of $x \in \mathbb{R}$ is the following statement false? Why?

" If $|x - 3| = 1$, then $|x - 2| = 2$. "

3.7.8. Construct the truth tables for the following statements.

(a) $(P \wedge Q) \vee (\neg Q)$

(b) $P \Rightarrow (P \Rightarrow Q)$

(c) $(P \Rightarrow Q) \Rightarrow P$

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

(e) $(P \wedge Q) \Leftrightarrow (P \vee Q)$

(f) $[P \vee (\neg Q)] \Rightarrow [Q \wedge (\neg P)]$

3.7.9. Find a statement R (in terms of P and Q) with the following truth table:

P	Q	R
T	T	F
T	F	T
F	T	T
F	F	T

3.7.10. (a) Let P, Q, and R be three statements. If $P \vee (Q \Rightarrow (\neg R))$ is a **false** statement, what must be the truth values of P, Q, and R? Why?

(b) Let P, Q, R and S be four statements. If $[(P \wedge Q) \vee R] \Rightarrow (R \vee S)$ is a **false** statement, what must be the truth values of P, Q, R and S? Why?

3.7.11. Let P, Q and R be three statements.

(a) If P is **false**, Q is **false**, and R is **true**, is the statement $[P \wedge (\neg Q)] \Rightarrow (R \vee Q)$ true or false?

(b) If P, Q and R are **all true**, is the statement $[P \wedge (\neg Q)] \Rightarrow (R \vee Q)$ true or false?

3.7.12. Prove parts (a), (b) and (d) in Proposition 3.4.2. Also provide real-life examples that illustrate the equivalence of the statements.

3.7.13. (a) Find a statement that is equivalent to $\neg(P \wedge (\neg Q))$, which is a disjunction (i.e., includes the symbol $\vee$).

(b) Write a statement that is equivalent to $P \Rightarrow Q$, using only the connectives $\vee$ and $\neg$.

3.7.14. Explain why $(P \vee Q) \Rightarrow R$ and $(P \Rightarrow R) \wedge (Q \Rightarrow R)$ are logically equivalent statements (try to avoid truth tables).

3.7.15. (a) Is the statement $P \Rightarrow Q$ logically equivalent to $(\neg P) \Rightarrow (\neg Q)$? Explain.

(b) Is the statement $(P \wedge Q) \vee R$ equivalent to $P \wedge (Q \vee R)$? Explain.

(c) Is the statement $(P \Rightarrow Q) \Rightarrow R$ equivalent to $P \Rightarrow (Q \Rightarrow R)$? Explain.

(d) Is the statement $(\neg P) \Leftrightarrow (\neg Q)$ the **negation** of $P \Leftrightarrow Q$? Explain.

3.7.16. Write the **contrapositive** of the following sentences. Use words or the logic symbols, and simplify your answer. Try to avoid using the negation symbol, or words of negation.

(a) If k is a prime and $k \neq 2$, then k is odd.

(b) If I do my assignments, I will get a good mark in the course.

(c) If $x^2 + y^2 = 9$, then $-3 \leq x \leq 3$.

(d) If $a^2 + b^2 = 0$, then $a = 0$ and $b = 0$.

(e) If Anna is failing both history and psychology, then Anna is not graduating.

3.7.17. Which of the following are tautologies? Which are contradictions? Explain.

$$P \wedge (\neg P) \qquad P \Leftrightarrow (\neg P) \qquad P \vee (\neg P) \qquad P \Rightarrow (\neg P) \qquad (P \wedge Q) \Rightarrow Q$$

3.7.18. In the following cartoon, the dog concludes that he is a cat. Find the flaw in his argument. Which connectives and quantifiers are used? Can you relate this to any of the truth tables discussed in the chapter?

3.7.19. Write the statement "There is no set A, for which $A \in A$", without using words of negation (e.g., 'no', 'not').

3.7.20. Negate the following statements. You may use words or the logic symbols in your answers. Simplify the negations as much as you can.

(a) For all $x \in A$ there is a $b \in B$ such that $b > x$.

(b) For any positive real number x, there is a natural number n, for which $\dfrac{1}{n} < x$.

3.7.21. Write the negation of the following statements **without** using the negation symbol $\neg$.
Also, for each statement, decide whether it is **true** or **false**. Explain your answer briefly.

(a) $(\forall x \in \mathbb{R})(\exists y \in \mathbb{R})(x^2 > y^2)$

(b) $(\exists x \in \mathbb{Z})\left[\left(x^2 = (x+1)^2\right) \Rightarrow (x^3 \in \mathbb{Z})\right]$

(c) $(\forall n \in \mathbb{N})[(n-1)^3 + n^3 \neq (n+1)^3]$

(d) $[(\forall x \in \mathbb{R})(x > 0)] \Rightarrow [(\forall x \in \mathbb{R})(x = x + 1)]$

(e) $(\forall x \in \mathbb{R})[\,(x^2 \leq -1) \Rightarrow [(x+1)^2 = x^2 + 1]\,]$

(f) $(\forall x \in \mathbb{R})[(x > 0) \Rightarrow (\exists n \in \mathbb{N})(n \cdot x > 1)]$

(g) $(\forall x \in \mathbb{R})(\exists y \in \mathbb{R})[(x+y)^2 = x^2 + y^2]$

(h) $(\exists y \in \mathbb{R})(\forall x \in \mathbb{R})(|x+y| = |x| + |y|)$

(i) $(\forall x \in \mathbb{Q})(\exists n \in \mathbb{N})(n \cdot x \in \mathbb{Z})$

(j) $(\forall x \in \mathbb{R})(\forall y \in \mathbb{R})\,[((x + y \leq 7) \wedge (xy = x)) \Rightarrow (x < 7)]$

3.7.22. For each statement below, write it and its negation using the logic symbols. Make sure to simplify the negation as much as you can. Also, decide whether the given statement is true or false. **Explain** your decision briefly.

(a) There exists an integer M, such that $x^2 \leq M$ for all real numbers x .

(b) There is a real number y, such that $|x - y| = |x| - |y|$ for any real number x.

(c) For all real numbers x, $(x-6)^2 = 4$ implies $x = 8$.

(d) For all real numbers x, y, if $x^2 - y^2 = 9$, then $|x| \geq 3$.

(e) For any real number x, If $(x-1)(x-3) = 3$, then $x - 1 = 3$ or $x - 3 = 3$.

3.7.23. Write the **negation** of the following statement in words.

"For any field F , and any $a \in F$, if $a^3 = 1$ then $a = 1$."

Is this statement true or false? **Explain**.

3.7.24. Let P be the statement $(\forall x \in \mathbb{Z})[x(x-1) \geq 0]$.

(a) It is a common error to believe that the negation of P is $(\forall x \notin \mathbb{Z})[x(x-1) < 0]$. Why is this wrong? What is the correct way to negate the statement?

(b) Which statement is true: P or $\neg P$? Explain.

3.7.25. Write **in words** the definition of an **unbounded** function from $\mathbb{R}$ to $\mathbb{R}$ (see Exercise 2.5.29). Interpret this definition geometrically. You may add a diagram to support your answer.

3.7.26. Let (a_n) be a sequence of real numbers: $a_1, a_2, a_3, \ldots$.

(a_n) is said to be **increasing**, if $a_n < a_{n+1}$ for all $n \in \mathbb{N}$.

Using words, write the statement "(a_n) is **not** increasing.". **Do not** use any words of negation.

3.7.27. Review Example 3.6.1 before attempting this exercise.

(a) Prove that a four-digit positive integer, whose sum of digits is divisible by 3, is also divisible by 3.

(b) Prove that if n is a three-digit positive integer that is divisible by 3, then its sum of digits is also divisible by 3. Can you combine this observation and Example 3.6.1 into one if-and-only-if statement?

(c) Generalize our previous discoveries to prove that "A natural number n is divisible by 3 if and only if its sum of digits is divisible by 3".

3.7.28. Prove that a natural number is divisible by 9 if and only if its sum of digits is divisible by 9.

3.7.29. Is the following statement true or false? Justify your answer with a proof or a counterexample.

"For any $n \in \mathbb{N}$, we have $(n-1)^2 + n^3 = (n+1)^3$."

3.7.30. Use **contrapositive** to prove the following statements.

(a) Let x be an integer. If $x^2 - 1$ is **not** divisible by 8, then x is even.

(b) Let m and n be two integers. Prove that if $m^2 + n^2$ is divisible by 4, then both m and n are even numbers.

(c) Let $x, y \in \mathbb{R}$. If x and y are both positive, then $\sqrt{x+y} \neq \sqrt{x} + \sqrt{y}$.

(d) Let $x, y \in \mathbb{R}$. If $x \neq y$, then $\dfrac{x}{\sqrt{x^2+1}} \neq \dfrac{y}{\sqrt{y^2+1}}$.

(e) Let $x \in \mathbb{R}$. If $x^3 + 5x = 40$, then $x < 3$.

3.7.31. Prove that the following equations have **no rational** solutions.

(a) $x^3 + x^2 = 1$

(b) $x^3 + x + 1 = 0$

(c) $x^5 + 3x^3 + 7 = 0$

(d) $x^5 + x^4 + x^3 + x^2 + 1 = 0$

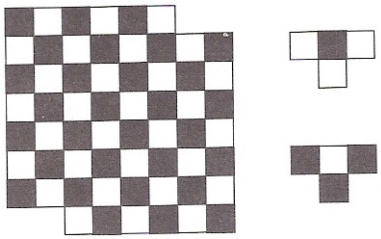
3.7.32. Prove that the following equations have **no natural** solutions.

(a) $x^2 - 4y^2 = 7$

(b) $x^2 - y^2 = 10$

3.7.33. Prove that the equation $x^2 + x + 1 = y^2$ has no natural solutions.
(**Hint**: Multiply by 4 and complete the square.)

3.7.34. Two squares from each of two opposite corners are deleted from a checkerboard, as shown below. Prove that the remaining squares cannot be fully covered using copies of the "T-shapes" and their rotations.



3.7.35. Read the proof of Theorem 3.6.6. The proof argues that if $p_1, \ldots, p_k$ are prime numbers, then $M = p_1, \ldots, p_k + 1$ must be either a new prime number, or a number that is divisible by a prime other than $p_1, \ldots, p_k$.

(a) Verify this argument by assuming that $k = 4$, $p_1 = 2$, $p_2 = 3$, $p_3 = 5$ and $p_4 = 7$. What is M? Is it a prime number?

(b) Repeat with $p_1, \ldots, p_k$ be the primes $7, 11, 13, 19$. And again with $2, 5, 11, 19, 23$. Use a computing device to calculate M.

Chapter 4

Mathematical Induction

4.1 The Principle of Mathematical Induction

Mathematical induction is a proof technique, used to prove that an infinite sequence of statements is true (or, equivalently, that a statement holds true for all natural numbers). When dealing with infinitely many statements, there is no way we can prove them all by proving each statement individually. Induction is a tool we can often use to bypass this difficulty.

Mathematical induction is an extremely powerful proof technique in mathematics, for several reasons. First, it is not restricted to specific areas of mathematics, and thus can be used to prove statements in algebra, geometry, number theory, analysis, etc. Secondly, induction is useful at all levels of mathematics. It can be used to prove elementary statements about numbers, as well as advanced statements in, say, topology and modern algebra.

Examples. Here are a few motivating examples.

(a) In Chapter 1 we proved the Arithmetic-Geometric Mean Inequality (See Proposition 1.2.3). This inequality can be generalized to more than two numbers, as follows.

$$\sqrt[n]{x_1 \cdot x_2 \cdot \ldots \cdot x_n} \leq \frac{x_1 + x_2 + \cdots + x_n}{n} \qquad (\text{for} \quad x_1, x_2, \ldots, x_n \geq 0).$$

How would one prove such an inequality? We have already proved it for the case where $n = 2$, and in Exercise 1.5.19, even for $n = 3$ and $n = 4$. We can spend more time and construct proofs for $n = 5$ and $n = 6$, but that would not justify the inequality for all n's. Mathematical induction can be used to construct a proof for **any** natural number n (see Exercise 4.6.33).

(b) The Triangle Inequality (Proposition 1.3.3) can also be generalized to more then two numbers. We

will prove soon that for any $x_1, \ldots, x_n \in \mathbb{R}$, we have

$$|x_1 + x_2 + \cdots + x_n| \leq |x_1| + |x_2| + \cdots + |x_n| \ .$$

We have already proved the inequality for the case $n = 2$, and it is not too hard to prove it for other values of n. However, proving it for **any** number of x's (i.e., for every n) requires induction.

(c) Finally, take a look at the following equalities:

$$1 + 3 = 4 \quad , \quad 1 + 3 + 5 = 9 \quad , \quad 1 + 3 + 5 + 7 = 16 \quad , \quad 1 + 3 + 5 + 7 + 9 = 25 \ .$$

Can you notice a pattern? It looks like the sum of consecutive odd natural numbers (starting at 1) always results in a square number. But how can we be sure? We can keep checking more and more sums (and with computers, we can even check thousands of sums in a split of a second), but that would still not cover all cases. To prove that $1 + 3 + 5 + \cdots + (2n - 1) = n^2$ for **any** $n \in \mathbb{N}$, we will need induction.

The Principle of Mathematical Induction (PMI).

Assume that $P(1), P(2), P(3), \ldots$ is an infinite sequence of mathematical statements.

If (1) $P(1)$ is true, and

(2) For any $k \in \mathbb{N}$, $P(k)$ implies $P(k+1)$,[1]

then all the statements in the sequence are true.

Remarks.

- Condition (1) is often called **the base case**, and condition (2) – **the induction step**. In most cases, proving the base case is relatively easy and quick (but must be done nevertheless), and most of the work is put in proving the induction step.

- The principle of mathematical induction is often taken as an axiom for the natural numbers, or is proved from other axioms. As we did not follow an axiomatic approach to defining the natural numbers, we will accept the principle without proof. In words, the principle says that if the first statement in the sequence is true, and if every statement implies the next, then all the statements are true.

$$P(1) \Rightarrow P(2) \Rightarrow P(3) \Rightarrow P(4) \Rightarrow \cdots \Rightarrow P(k) \Rightarrow P(k+1) \Rightarrow \ldots$$

[1] condition (2) can be written as $(\forall k \in \mathbb{N})[P(k) \Rightarrow P(k+1)]$.

- Some like to compare the above principle to falling dominoes. If you line up dominoes in a straight line, then pushing the first domino creates a chain effect that eventually brings down all the dominoes. However, to make it work, two conditions must be met: (1) Someone has to push the first domino in line, and (2) the dominoes must be close enough to each other, so that a falling domino pushes the next one in line. These two conditions are analogous to the conditions stated in the principle of mathematical induction.

We proceed with some examples.

Examples.

(a) We start by confirming that the sum of the first n odd natural numbers is n^2, for any n.

Claim. For any $n \in \mathbb{N}$, we have $1 + 3 + 5 + \cdots + (2n - 1) = n^2$.

Note that the equality above can be interpreted as an infinite sequence of statements $P(1), P(2), P(3), \ldots$.

$$\begin{array}{lll} P(1) \text{ is } & 1 = 1^2 & (n = 1) \\ P(2) \text{ is } & 1 + 3 = 2^2 & (n = 2) \\ P(3) \text{ is } & 1 + 3 + 5 = 3^2 & (n = 3) \\ P(4) \text{ is } & 1 + 3 + 5 + 7 = 4^2 & (n = 4) \\ \vdots & \vdots & \vdots \end{array}$$

It is straightforward to verify each equality in the sequence, but our task is to prove them all! Using induction, we construct a proof, as follows.

Proof. Verifying the **base case** (condition (1) in the principle of mathematical induction) is immediate. As $1 = 1^2$, $P(1)$ is a true statement, as needed.

Next, we need to prove that for any $k \in \mathbb{N}$, $P(k)$ implies $P(k + 1)$. In other words, we assume that

$$1 + 3 + 5 + \cdots + (2k - 1) = k^2 \qquad \text{for } \textbf{some } \ k \in \mathbb{N}$$

(this is called the **induction hypothesis**), and we need to prove that

$$1 + 3 + 5 + \cdots + [2(k + 1) - 1] = (k + 1)^2.$$

Observe that the last term in the sum $1+3+5+\cdots+[2(k+1)-1]$ is $2k+1$, and hence the preceding term must be $2k - 1$. We can now prove the $(k + 1)$-case, as follows.

$$\begin{aligned} 1 + 3 + 5 + \cdots + [2(k + 1) - 1] &= 1 + 3 + 5 + \cdots + (2k - 1) + (2k + 1) \\ &= [1 + 3 + 5 + \cdots + (2k - 1)] + (2k + 1) \\ &= k^2 + (2k + 1) = (k + 1)^2. \end{aligned}$$

Note how in the third step, the induction hypothesis was used to replace $1+3+5+\cdots+(2k-1)$ with k^2.

We managed to prove $P(k+1)$ from $P(k)$, which completed the induction step of our proof. By the principle of mathematical induction, the equality $1+3+5+\cdots+(2n-1)=n^2$ is valid for any $n\in\mathbb{N}$. □

(b) We now prove an inequality, known as **Bernoulli's Inequality**.

Claim. For any $n\in\mathbb{N}$ and $x\in\mathbb{R}$, with $x\geq -1$, we have $(1+x)^n\geq 1+nx$.

Proof. We proceed by induction on n (while x is treated as a fixed, but arbitrary, real number, greater than or equal to -1).

For $n=1$, the inequality becomes $(1+x)^1\geq 1+1\cdot x$, which is valid, and hence the base case holds.

Next, we assume that $(1+x)^k\geq 1+kx$ for **some** natural number k, and prove the $(k+1)$-case.

$$(1+x)^{k+1}=(1+x)^k\cdot(1+x)\geq(1+kx)(1+x)=1+kx+x+kx^2=1+(k+1)x+kx^2.$$

The induction hypothesis, $(1+x)^k\geq 1+kx$, was used in the second step, together with the fact that $1+x\geq 0$. Also, as $kx^2\geq 0$, we conclude that

$$1+(k+1)x+kx^2\geq 1+(k+1)x.$$

Overall, we have proved that $(1+x)^{k+1}\geq 1+(k+1)x$, which is Bernoulli's inequality for $n=k+1$. By the principle of mathematical induction, Bernoulli's inequality holds true for any $n\in\mathbb{N}$ (and $x\geq -1$). □

(c)

Theorem (The Triangle Inequality). For any n real numbers $x_1,x_2,\ldots,x_n$, we have

$$|x_1+x_2+\cdots+x_n|\leq|x_1|+|x_2|+\cdots+|x_n|\ .$$

The Theorem generalizes Proposition 1.3.3 from two to arbitrarily many numbers.

Proof. We perform induction on n (i.e., on the number of x's). In other words, we prove by induction that for any $n\in\mathbb{N}$, the inequality $|x_1+x_2+\cdots+x_n|\leq|x_1|+|x_2|+\cdots+|x_n|$ holds true for all real numbers $x_1,\ldots,x_n$.

If $n=1$ (the base case), we get $|x_1|\leq|x_1|$, which holds true for any real number x_1.

Now assume that the triangle inequality is valid for $n = k$ numbers (this is the induction hypothesis), and consider $k + 1$ real numbers $x_1, x_2, \ldots, x_k, x_{k+1}$.

Using Proposition 1.3.3, with $x = x_1 + \cdots + x_k$ and $y = x_{k+1}$, we have

$$|x_1 + x_2 + \cdots + x_k + x_{k+1}| = |(x_1 + x_2 + \cdots + x_k) + x_{k+1}| \leq |x_1 + x_2 + \cdots + x_k| + |x_{k+1}| \ ,$$

and by the induction hypothesis,

$$\leq |x_1| + |x_2| + \cdots + |x_k| + |x_{k+1}| \ ,$$

as needed. This proves the inequality for $n = k + 1$, and hence, by induction, for all n's. □

(d) **Claim.** For any $n \in \mathbb{N}$, $2^{6n} + 3^{2n-2}$ is divisible by 5.

Note that there is no equality or inequality to prove here. This statement involves the notion of divisibility.

Proof. For $n = 1$, we get $2^6 + 3^0 = 64 + 1 = 65$, which is divisible by 5, and so the base case holds.

Assume that $2^{6k} + 3^{2k-2}$ is divisible by 5 for some $k \in \mathbb{N}$ (this is the induction hypothesis). We need to prove that $2^{6(k+1)} + 3^{2(k+1)-2}$ is also divisible by 5. To do so, we write

$$2^{6(k+1)} + 3^{2(k+1)-2} = 2^{6k+6} + 3^{2k-2+2} = 2^6 \cdot 2^{6k} + 3^2 \cdot 3^{2k-2} = 64 \cdot 2^{6k} + 9 \cdot 3^{2k-2} =$$
$$= (55 + 9) \cdot 2^{6k} + 9 \cdot 3^{2k-2} = 55 \cdot 2^{6k} + 9 \cdot (2^{6k} + 3^{2k-2}) \ .$$

The term $55 \cdot 2^{6k}$ is divisible by 5, since 5 divides 55, and the term $9 \cdot (2^{6k} + 3^{2k-2})$ is divisible by 5, by the induction hypothesis. We conclude that $2^{6(k+1)} + 3^{2(k+1)-2}$ is divisible by 5. This proves the claim for $n = k + 1$, and hence, by induction, the claim is valid for all $n \in \mathbb{N}$. □

(e) Our next example is somewhat different from the previous ones. It involves **counting sets**. We begin with the following question.

Question: Given a finite nonempty set S, with n elements, how many subsets does S have (including, of course, the empty set and the set S itself)?

Note that the question we posed is **not** a mathematical statement, and thus there is nothing we can prove (yet). We would like, however, to answer this question, and we start by looking at some special cases.

If $n = 1$ (i.e., S has only one element), then there are only **two** subsets: ϕ and S.

If S has two elements, say $S = \{a, b\}$, then we get **four** subsets: $\phi, \{a\}, \{b\}, \{a, b\}$.

For $n = 3$, the set $S = \{a, b, c\}$ has eight subsets: $\phi, \{a\}, \{b\}, \{c\}, \{a, b\}, \{a, c\}, \{b, c\}, \{a, b, c\}$.
And for $n = 4$, we obtain **sixteen** subsets (check!).

By examining the pattern $2, 4, 8, 16, \ldots$, we conjecture that a set with n elements has 2^n subsets. This is a mathematical statement, that we can prove by induction.

Claim 4.1.1. A finite set, with n elements, has 2^n subsets.[2]

Proof. A set S, with one element, has only two subsets: S and ϕ, and thus the claim holds true for $n = 1$ (the base case).
Assume that the claim holds true for some natural number k (i.e., for $n = k$), and consider a set S with $k + 1$ elements:

$$S = \{x_1, x_2, \ldots, x_k, x_{k+1}\}.$$

Denote by $\tilde{S}$ the set obtained from S by removing the element x_{k+1}: $\tilde{S} = \{x_1, x_2, \ldots, x_k\}$. By our assumption, sets with k elements have 2^k subsets, and so $\tilde{S}$ has 2^k subsets, which we denote by $A_1, A_2, \ldots, A_{2^k}$.

Remember that our task is to prove the claim for $n = k + 1$, i.e., to count the number of subsets of S (and not $\tilde{S}$). However, as $\tilde{S} \subseteq S$, every subset of $\tilde{S}$ is also a subset of S.

This means that $A_1, A_2, \ldots, A_{2^k}$ are also subsets of S. Are there any more? Of course there are. None of the subsets $A_1, \ldots, A_{2^k}$ contains the element x_{k+1}. We create more subsets of S by adding x_{k+1} to each of the existing subsets, as follows:

$$B_1 = A_1 \cup \{x_{k+1}\} \quad , \quad B_2 = A_2 \cup \{x_{k+1}\} \quad , \quad B_3 = A_3 \cup \{x_{k+1}\} \quad , \quad \ldots \quad , \quad B_{2^k} = A_{2^k} \cup \{x_{k+1}\} \ .$$

This way, we obtained a **complete list** of all subsets of S: $A_1, A_2, \ldots, A_{2^k}, B_1, B_2, \ldots, B_{2^k}$.
Every subset of S is listed precisely once: If a subset does not contain x_{k+1}, it must be one of the A_i's. Otherwise, it must be one of the the B_j's.

So how many subsets are there in total for S? We have 2^k A_i's, and 2^k B_j's. Overall, there are $2^k + 2^k = 2 \cdot 2^k = 2^{k+1}$ subsets, which proves the claim for $n = k + 1$. By induction, the claim holds true for all n's. □

[2] Note that the claim is also valid for $n = 0$, as an empty set has $2^0 = 1$ subsets.

4.2 Summation and Product Notation

In mathematics, we often use induction to prove identities or inequalities involving sums and products (such as the triangle inequality). We introduce below the commonly used Pi and Sigma notation for writing long sums and products in a more efficient (or 'compact') way.

Definition 4.2.1. Suppose that $m, n \in \mathbb{Z}$ with $m \leq n$, and $a_m, a_{m+1}, \ldots, a_n$ are real numbers. We define

$$\sum_{i=m}^{n} a_i = a_m + a_{m+1} + a_{m+2} + \cdots + a_n \qquad \text{(\textbf{Sigma notation} for sums)}$$

$$\prod_{i=m}^{n} a_i = a_m \cdot a_{m+1} \cdot a_{m+2} \cdot \ldots \cdot a_n \qquad \text{(\textbf{Pi notation} for products)}$$

Note that the index i serves as a counter (and other letters may be used instead). In both cases, we evaluate the summand a_i for $i = m, m+1, m+2, \ldots, n$, and then add or multiply the resulting elements.

Examples.

(a) $\displaystyle\sum_{i=1}^{10} \frac{1}{i} = \frac{1}{1} + \frac{1}{2} + \frac{1}{3} + \cdots + \frac{1}{10}$

(b) $\displaystyle\sum_{i=3}^{10} (2i) = 6 + 8 + 10 + \cdots + 20$

(c) $\displaystyle\prod_{k=5}^{15} 3^{k+2} = 3^7 \cdot 3^8 \cdot 3^9 \cdot \ldots \cdot 3^{17}$

(d) $\displaystyle\prod_{k=1}^{n} k = 1 \cdot 2 \cdot 3 \cdot \ldots \cdot n$ (This product is also denoted as $n!$, and reads "n factorial".)

(e) If c is a constant number, and $n \in \mathbb{N}$, then

$$\sum_{j=1}^{n} c = c + c + \cdots + c = n \cdot c \qquad \text{and} \qquad \prod_{j=1}^{n} c = c \cdot c \cdot \ldots \cdot c = c^n \ .$$

(f) $\displaystyle\sum_{j=3}^{78} \left(\frac{1}{j} - \frac{1}{j+1}\right) = \left(\frac{1}{3} - \frac{1}{4}\right) + \left(\frac{1}{4} - \frac{1}{5}\right) + \left(\frac{1}{5} - \frac{1}{6}\right) + \cdots + \left(\frac{1}{78} - \frac{1}{79}\right) = \frac{1}{3} - \frac{1}{79} = \frac{76}{237}$

(note how most terms in the sum cancel each other, which allowed us to simplify it to a single fraction.)

Known properties of sums and products can be re-written using the Sigma and Pi notation.

Proposition 4.2.2. The following identities hold (c, a_i and b_i denote arbitrary real numbers).

(a) $\displaystyle\sum_{i=m}^{n}(c\cdot a_i) = c\cdot\sum_{i=m}^{n} a_i$

(b) $\displaystyle\sum_{i=m}^{n}(a_i+b_i) = \left(\sum_{i=m}^{n} a_i\right)+\left(\sum_{i=m}^{n} b_i\right)$

(c) $\displaystyle\prod_{i=m}^{n}(c\cdot a_i) = c^{n-m+1}\cdot\prod_{i=m}^{n} a_i$

(d) $\displaystyle\prod_{i=m}^{n}(a_i\cdot b_i) = \left(\prod_{i=m}^{n} a_i\right)\cdot\left(\prod_{i=m}^{n} b_i\right)$

We can easily justify these properties by writing sums and products explicitly, without the Sigma or Pi notation. For instance, to justify part (a), we write

$$\sum_{i=m}^{n}(c\cdot a_i) = c\cdot a_m + c\cdot a_{m+1} + \cdots + c\cdot a_n = c\cdot(a_m + a_{m+1} + \cdots + a_n) = c\cdot\sum_{i=m}^{n} a_i \ .$$

As an exercise, try to justify parts (b), (c) and (d) using a similar strategy. The following proposition is a generalization of Bernoulli's Inequality (see Example (b) on page 86).

Proposition 4.2.3.
If $x_1, x_2, \ldots, x_n$ are real numbers in the interval $[0,1]$, then $\displaystyle\prod_{i=1}^{n}(1-x_i) \geq 1-\sum_{i=1}^{n} x_i$.

(**Remark**: The inequality can be written explicitly, as

$$(1-x_1)\cdot(1-x_2)\cdot\ldots\cdot(1-x_n) \geq 1-(x_1+x_2+\cdots+x_n)\ .\ )$$

Proof. We perform induction on n (the number of x's). If $n=1$, the inequality becomes $(1-x_1)\geq(1-x_1)$, which is clearly valid (for any real number x_1). Next, we assume that the inequality holds true for some $n=k$, and prove it for $n=k+1$. Suppose $x_1, x_2, \ldots, x_k, x_{k+1}$ are real numbers in the interval $[0,1]$.

$$\prod_{i=1}^{k+1}(1-x_i) = \left[\prod_{i=1}^{k}(1-x_i)\right]\cdot(1-x_{k+1}) \geq \left(1-\sum_{i=1}^{k} x_i\right)\cdot(1-x_{k+1}) =$$

Note how the induction hypothesis was used in the second step (together with the fact that $1-x_{k+1}\geq 0$). We continue by expanding the brackets:

$$= 1-\left(\sum_{i=1}^{k} x_i\right) - x_{k+1} + x_{k+1}\cdot\sum_{i=1}^{k} x_i = 1-\left(\sum_{i=1}^{k+1} x_i\right) + x_{k+1}\cdot\sum_{i=1}^{k} x_i \geq$$

Finally, we observe that $x_{k+1} \cdot \sum_{i=1}^{k} x_i$ is a nonnegative number (as all the the x_i's are), and hence we get

$$\geq 1 - \left(\sum_{i=1}^{k+1} x_i \right) .$$

Putting it all together, we have shown that

$$\prod_{i=1}^{k+1} (1 - x_i) \geq 1 - \sum_{i=1}^{k+1} x_i ,$$

which is our inequality for $n = k + 1$. This completes the proof of the proposition. □

4.3 Variations

In previous sections, we used mathematical induction to prove that a given statement is valid for **all** natural numbers. However, in some cases, we might want to show that a statement is valid for only **some** $n \in \mathbb{N}$. Can we still use induction? Yes, we can, as long as we use the appropriate variation of the original PMI. Here is an example.

Example 4.3.1. For which $n \in \mathbb{N}$ do we have $2^n \geq (n+1)^2$?

Clearly, the inequality is **not** valid for all $n \in \mathbb{N}$. If $n = 1$, we get $2 \geq 4$, which is **false**. Nevertheless, we know (at least informally) that exponential functions grow faster than quadratics, and hence we expect the inequality to be valid, as long as n is large enough. But what do we mean by "large enough"? Can we make this statement precise? And moreover, can we prove it?

Let us start by trying some small values for n:

For $n = 1$:	$2^1 \geq 2^2$	False!
For $n = 2$:	$2^2 \geq 3^2$	False!
For $n = 3$:	$2^3 \geq 4^2$	False!
For $n = 4$:	$2^4 \geq 5^2$	False!
For $n = 5$:	$2^5 \geq 6^2$	False!
For $n = 6$:	$2^6 \geq 7^2$	**True!**
For $n = 7$:	$2^7 \geq 8^2$	**True!**
For $n = 8$:	$2^8 \geq 9^2$	**True!**

As we can see, the inequality is valid for $n = 6, 7, 8$, and it looks like it remains valid, as long as $n \geq 6$. We now state this fact and prove it by induction. However, we use $n = 6$ as our base case, instead of $n = 1$.

Claim. If $n \in \mathbb{N}$ and $n \geq 6$, then $2^n \geq (n+1)^2$.

Proof. The base case $n = 6$ is easily verified, as $2^6 = 64$ is greater than $(6+1)^2 = 49$.
Assume that $2^k \geq (k+1)^2$ for some natural number $k \geq 6$, and compute:

$$2^{k+1} = 2^k \cdot 2 \geq (k+1)^2 \cdot 2 = 2k^2 + 4k + 2 = k^2 + 4k + 4 + k^2 - 2 = (k+2)^2 + (k^2 - 2) \geq (k+2)^2 \ .$$

Note how the induction hypothesis is used in the second step. In the last step, we omitted the term $k^2 - 2$, which must be positive, as $k \geq 6$. Overall, we proved the inequality for $n = k+1$, and hence, by induction, for all $n \geq 6$. □

In the proof above, we used the following variation of the Principle of Mathematical Induction.

Variation 1
Suppose that $P(1), P(2), \ldots$ is a sequence of statements, and let ℓ be a natural number.
If $P(\ell)$ is true, and $P(k) \Rightarrow P(k+1)$ for any $k \geq \ell$, then $P(n)$ is true for all $n \geq \ell$.

Other variations can be formed. For instance, Variations 2 and 3 below can be used to prove that a statement is valid for all **even** (respectively **odd**) natural numbers.

Variation 2
Let $P(1), P(2), \ldots$ be a sequence of statements.
If $P(2)$ is true, and $P(k) \Rightarrow P(k+2)$ for any **even** $k \in \mathbb{N}$, then $P(n)$ is true for all **even** n's.

Variation 3
Let $P(1), P(2), \ldots$ be a sequence of statements.
If $P(1)$ is true, and $P(k) \Rightarrow P(k+2)$ for any **odd** $k \in \mathbb{N}$, then $P(n)$ is true for all **odd** n's.

Example. For all **even** $n \in \mathbb{N}$, $n(n^2 + 3n + 2)$ is divisible by 24.

Proof. We use Variation 2 to prove the above statement.

For $n = 2$, the base case, we have $n(n^2 + 3n + 2) = 2 \cdot (4 + 6 + 2) = 24$, which is, of course, divisible by 24. Now assume that $k(k^2 + 3k + 2)$ is divisible by 24 for some even $k \in \mathbb{N}$, and consider the case $n = k + 2$:

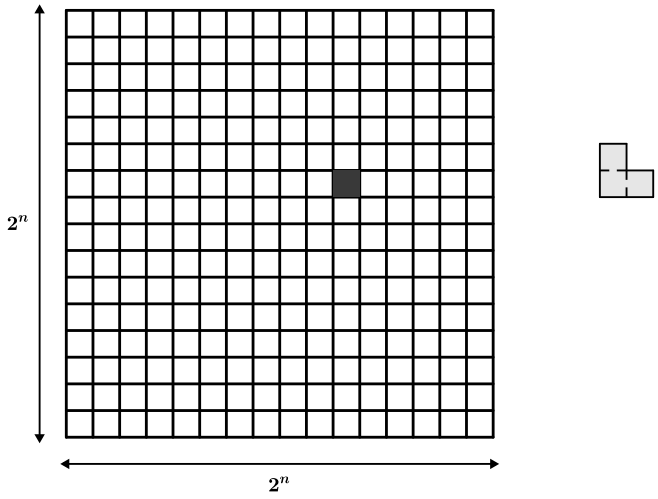
$$\begin{aligned}(k+2) \cdot [(k+2)^2 + 3(k+2) + 2] &= (k+2) \cdot (k^2 + 4k + 4 + 3k + 6 + 2) = \\ &= (k+2) \cdot [(k^2 + 3k + 2) + (4k + 10)] = \\ &= k \cdot (k^2 + 3k + 2) + k \cdot (4k + 10) + 2 \cdot (k^2 + 3k + 2 + 4k + 10) = \\ &= k \cdot (k^2 + 3k + 2) + (6k^2 + 24k + 24)\end{aligned}$$

The term $k \cdot (k^2+3k+2)$ is divisible by 24, by the induction hypothesis. The second term, $6k^2+24k+24$, is also divisible by 24, since k is even (and hence k^2 is divisible by 4). We proved the statement for $n = k+2$, and by PMI (Variation 2), for all even $n \in \mathbb{N}$. □

4.4 Additional Examples

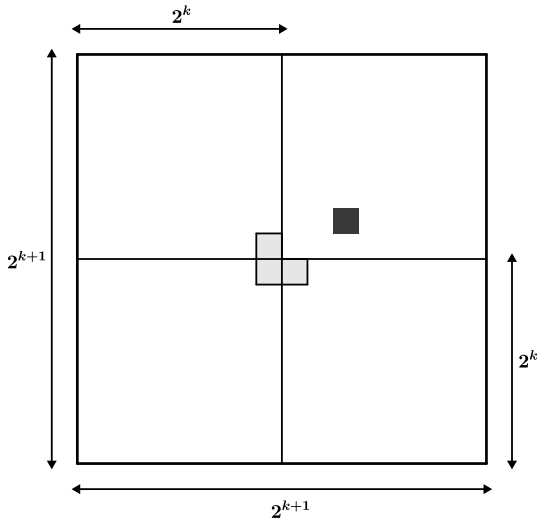
We present a few more examples of problems that can be solved by induction, and we begin with a "less mathematical" one.

Example 4.4.1. Prove that for any $n \in \mathbb{N}$, a $2^n \times 2^n$ grid, with one cell occupied, has an L-tiling. That is, it can be covered by L-shapes and their rotations. Each L-shape covers exactly three cells (see diagram).



Proof. For $n = 1$, we have a 2×2 grid, with one cell occupied, which can be clearly covered with a single L-shape. Assume that for some $k \in \mathbb{N}$, any $2^k \times 2^k$ grid, with one cell occupied, has an L-tiling. We need to show that a $2^{k+1} \times 2^{k+1}$ grid, with one cell occupied, has an L-tiling.

To be able to use the induction hypothesis, we divide our $2^{k+1} \times 2^{k+1}$ grid into four equal parts, each having dimensions $2^k \times 2^k$. Moreover, we place one L-shape in the center, so that it occupies three cells, one in each of the parts that do not contain the occupied cell (see diagram).



Each of the smaller $2^k \times 2^k$ grids has one occupied cell, so by the induction hypothesis, has an L-tiling. The original claim follows by induction. □

Recursion.

Our next example is about recursion. A recursive definition of a sequence is a rule for generating its elements from previous entries. Elements of a sequence are often denoted by a lower case letter and subscript (or index). For instance, the sequence (a_n) is the sequence of numbers

$$a_1, a_2, a_3, a_4, \dots \ .$$

The first element of the sequence is a_1, the second is a_2, the third - a_3, and so on. In general, a_n is the n-th entry of the sequence (a_n). Here is an example of a recursively defined sequence.

$$\begin{cases} a_1 = 6 \\ a_{n+1} = 5 \cdot a_n + 1 \qquad \text{for} \quad n \in \mathbb{N} \end{cases} .$$

The definition provides the first element of the sequence $a_1 = 6$, and then a rule for computing further elements. The equation $a_{n+1} = 5 \cdot a_n + 1$ tells us how to compute the $(n+1)$-st element from the n-th element. Namely, we multiply by 5 and add 1. Consequently, the first few numbers in the sequence are

$$6, 31, 156, 781, 3906, \dots$$

In some cases, recursive definitions are more elegant or natural than an explicit formula in terms of n (such as $b_n = 4n^3 - 5n$), but they can also be less convenient. In the sequence (a_n) above, we cannot find the one-hundredth element a_{100} unless we first compute all the preceding 99 entries $a_1, a_2, \dots, a_{99}$.

Are there methods for converting a recursive definition of a sequence into an explicit formula? Well, there is no general algorithm, but there are methods (some quite advanced) that can be applied in special cases. However, checking that a particular formula works can often be done by induction.

Example 4.4.2. Consider the recursive sequence $\begin{cases} a_1 = 6 \\ a_{n+1} = 5 \cdot a_n + 1 \qquad \text{for} \quad n \in \mathbb{N} \end{cases}$.

Prove that $a_n = \dfrac{5^{n+1}-1}{4}$ for all $n \in \mathbb{N}$.

Proof. Note that the recursive definition is given, and we may use it throughout the proof. It is the explicit formula that we now prove by induction.

If $n = 1$, we get

$$\frac{5^{n+1}-1}{4} = \frac{5^2-1}{4} = 6 \ ,$$

which equals a_1, and so the base case holds true.

Assume that $a_k = \dfrac{5^{k+1}-1}{4}$ for some $k \in \mathbb{N}$. To show that the explicit formula is valid for $n = k+1$, we use the recursive definition, and the induction hypothesis, as follows:

$$a_{k+1} = 5 \cdot a_k + 1 = 5 \cdot \frac{5^{k+1}-1}{4} + 1 = \frac{5 \cdot 5^{k+1} - 5 + 4}{4} = \frac{5^{k+2}-1}{4} \ .$$

We proved the case $n = k+1$, and hence, by induction, our proof is completed. □

A Fallacy.

Induction must be used carefully. Both the base case and the induction step need to be properly checked, and should not be treated as mechanical procedures. A careless application of the induction principle can lead to incorrect proofs.

For instance, the following claim is clearly wrong, but can you find the mistake in the proof?

"**Claim:**" In any group of n people (where $n \in \mathbb{N}$), all must have the same gender.

"*Proof.*" For $n = 1$, the claim holds true, since in a group with one person, there is indeed only one gender. Now assume that the claim is valid for some $k \in \mathbb{N}$, and consider a group with $k+1$ people. We call that group S, and denote its elements by $x_1, \ldots, x_{k+1}$:

$$S = \{x_1, x_2, \ldots, x_k, x_{k+1}\} \ .$$

To prove that all people in S have the same gender, we define the following subsets of S:

$$A = \{x_1, x_2, \ldots, x_k\} \qquad , \qquad B = \{x_2, \ldots, x_k, x_{k+1}\} \ .$$

In other words, A is obtained from S by removing the element x_{k+1}, and B is obtained by removing x_1. Both A and B are sets with k elements, and so by the induction hypothesis, in each of the groups, all have the same gender. Moreover, A and B overlap, as $A \cap B = \{x_2, \ldots, x_k\}$, and hence the gender of those in

A and B must be the same. We therefore conclude that all people in $S = A \cup B$ have the same gender, which proves the claim for $n = k + 1$. By induction, the claim holds true for all $n \in \mathbb{N}$. □

What is going on here? Where is the mistake in the proof?

The base case seems fine. The sets A and B indeed have k elements each, so applying the induction hypothesis on both is legitimate (the hypothesis can be applied as many times as needed). The problem arises when we claim that "A and B overlap". It is true that when S has three elements or more, the intersection $A \cap B$ is nonempty. However, if $S = \{x_1, x_2\}$, then $A = \{x_1\}$ and $B = \{x_2\}$, which implies that $A \cap B = \phi$, and the argument breaks.

The fact that an argument, that is supposed to hold for all $k \in \mathbb{N}$, fails for a single value of k, is enough to invalidate the whole proof!

4.5 Strong Induction

We devote the last section of this chapter to another variation of the principle of mathematical induction.

The Principle of Strong Mathematical Induction (PSMI).

Assume that $P(1), P(2), P(3), \ldots$ is an infinite sequence of mathematical statements.

If (1) $P(1)$ is true, and

(2) For any $k \in \mathbb{N}$: $P(1), P(2), \ldots, P(k)$ imply $P(k+1)$,

then all the statements in the sequence are true.

As we can see, the only difference between the two principles (PMI and PSMI), is in the induction step (i.e., in condition (2)). In PMI, we need to show that for any $k \in \mathbb{N}$, $P(k) \Rightarrow P(k+1)$ (i.e., that the k-th statement implies the $(k+1)$-st). With strong induction, we assume, in the induction step, that all the statements $P(1), P(2), \ldots, P(k)$ are true, and prove, from that assumption, the statement $P(k+1)$. In other words, our induction hypothesis is **stronger**, and allows us to use any of the statements for $n = 1, \ldots, k$ in our proof.

As an example, we prove the following theorem.

Theorem 4.5.1. *Every natural number $n \geq 2$ can be written as a product of prime numbers.*

(Note: If n is itself a prime number, we can still view it as a product of prime numbers, by allowing our "products" to have a single factor.)

Proof. We use strong induction in our proof, and $n = 2$ as our base case.

If $n = 2$, the theorem is valid, as 2 is a prime number. Assume that the theorem holds true for $n = 2, 3, 4, \ldots, k$ (for some natural number $k \geq 2$), and consider $n = k + 1$. If $k + 1$ happens to be a prime number, the theorem applies. Otherwise, $k + 1$ is composite, and is divisible by some natural number $2 \leq m \leq k$. Equivalently, $k + 1 = m \cdot \ell$, where m, ℓ are natural numbers between 2 and k.

Both m and ℓ are natural numbers, greater than 1 and smaller than $k + 1$, and hence covered by the induction hypothesis. Therefore, m and ℓ are products of prime numbers, and consequently, so is $k + 1 = m \cdot \ell$.

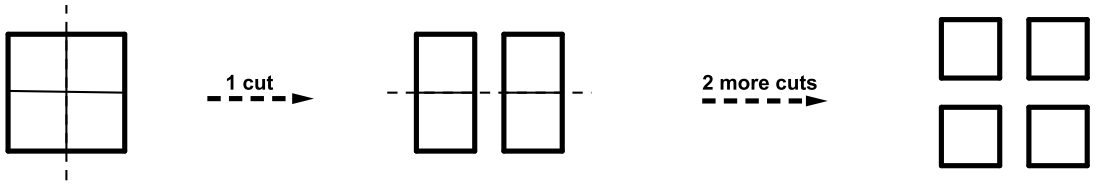
By PSMI, the theorem is valid for all natural numbers $n \geq 2$. □

Pay close attention to the proof of the Theorem. Why was strong induction needed here? Could we carry the above argument with "usual" induction?

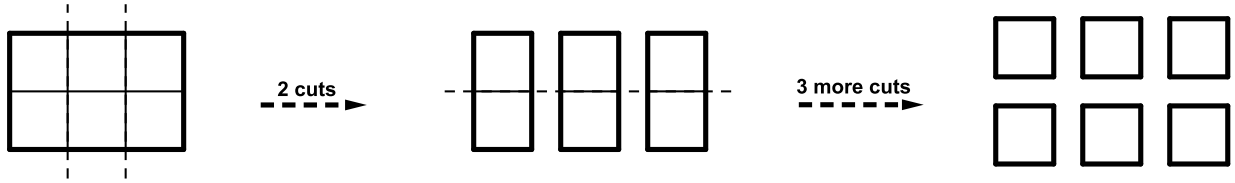
We proceed with two more examples.

Example 4.5.2. Consider a rectangular chocolate bar with n squares. How many cuts are needed to break the bar into 1×1 squares?
(Cuts are performed either horizontally or vertically, and a cut can be applied to one piece at a time.)

Let us start by checking a couple of special cases.
If we have a square chocolate bar, with $n = 4$ squares, then 3 cuts are needed.



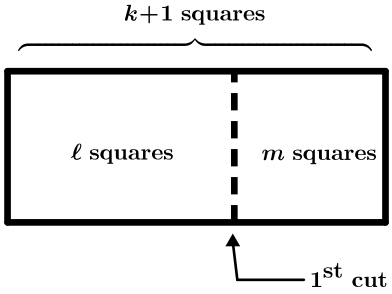
For a 2×3 bar (i.e., $n = 6$), we need 5 cuts altogether.



After experimenting a little more, it seems like the number of cuts needed is one less than the number of squares. We can now formulate a claim, and prove it by induction.

Claim 4.5.3. $n-1$ cuts are needed to break a rectangular chocolate bar, with n squares, into 1×1 squares.

Proof. If $n = 1$, then our bar consists of a single square, and no cuts are needed. As $n - 1 = 1 - 1 = 0$, the claim holds true, and the base case is verified. Assume that the claim is valid for $n = 1, 2, \ldots, k$ (for some $k \in \mathbb{N}$), and consider a chocolate bar with $k+1$ squares. To be able to use the induction hypothesis, we perform an arbitrary first cut, which breaks the bar into two separate pieces, say with ℓ and m squares (and so $\ell + m = k + 1$).



Note that both ℓ and m must be smaller than $k + 1$, and thus covered by the induction hypothesis. Therefore, $\ell - 1$ cuts are needed for the part with ℓ squares, and $m - 1$ for the part with m squares. Overall, the number of cuts needed to break our original bar, with $k + 1$ squares, into 1×1 pieces is

$$1 + (\ell - 1) + (m - 1) = \ell + m - 1 = (k + 1) - 1 = k$$

(the first '1' represents the initial cut), which proves the claim for $n = k + 1$. By strong induction, the claim is true for all $n \in \mathbb{N}$. □

Our last example is a theorem, related to **binary representation of numbers** (i.e., representing numbers with the digits 0 and 1 only). This important notion has many applications in mathematics, computer science and digital electronics (that will **not** be discussed in these notes).

Theorem 4.5.4. *Every natural number n can be expressed as a sum of distinct nonnegative integer powers of 2.*[3]

Let us first explain the theorem. A nonnegative integer is either 0 or a natural number, and so the nonnegative integer powers of 2 are

$$2^0 = 1\ ,\quad 2^1 = 2\ ,\quad 2^2 = 4\ ,\quad 2^3 = 8\ ,\quad 2^4 = 16\ ,\ldots$$

[3] Actually, this can be done in a **unique way**, but we will not prove this fact.

The theorem states that every natural number can be written as a sum of such powers, **without repetitions** (hence 'distinct'). These "sums" can have one or more summands. For instance:

$$17 = 2^0 + 2^4$$
$$128 = 2^7$$
$$42 = 2^1 + 2^3 + 2^5$$
$$65 = 2^0 + 2^6$$
$$312 = 2^3 + 2^4 + 2^5 + 2^8$$

Proof. The base case is easily verified, as $1 = 2^0$.

Let $k \in \mathbb{N}$, and assume that the theorem holds true for $n = 1, 2, 3, \ldots, k$. We need to prove that $k+1$ can be expressed as a sum of distinct nonnegative integer powers of 2, and we do that by looking at the following two possible cases.

- **Case 1: k is even.**

  By assumption, the theorem applies to $n = k$, and so we can write

  $$k = 2^{a_1} + 2^{a_2} + \cdots + 2^{a_m},$$

  where $a_1, \ldots, a_m$ are distinct **positive** integers. As k is even, the term 2^0 does not appear in the sum. Consequently,

  $$k + 1 = 2^0 + 2^{a_1} + 2^{a_2} + \cdots + 2^{a_m},$$

  and we have expressed $k+1$ in the required form.

- **Case 2: k is odd.**

  The argument used in Case 1 won't work here (why?), so we use a different approach. As k is odd, $k+1$ is even, and we can write $k+1 = 2m$, for some $m \in \mathbb{N}$. As m is smaller than $k+1$, the induction hypothesis applies, and we have

  $$m = 2^{a_1} + 2^{a_2} + \cdots + 2^{a_m} \ ,$$

  for some nonnegative distinct integers $a_1, \ldots, a_m$ (this time, one of the a_i's may be zero!).

  We conclude that

  $$k + 1 = 2m = 2^{a_1+1} + 2^{a_2+1} + \cdots + 2^{a_m+1} \ ,$$

  as needed. Note that since $a_1, a_2, \ldots, a_m$ are distinct, so are $a_1 + 1, a_2 + 1, \ldots, a_m + 1$.

We proved the theorem for $n = k + 1$, and hence, by strong induction, for any $n \in \mathbb{N}$. □

4.6 Exercises for Chapter 4

4.6.1. Prove the following equalities for all $n \in \mathbb{N}$.

(a) $1^2 + 3^2 + 5^2 + \cdots + (2n-1)^2 = \dfrac{4n^3 - n}{3}$

(b) $1^2 + 2^2 + 3^2 + \cdots + n^2 = \dfrac{1}{6}n(1+n)(1+2n)$

4.6.2. Prove the following inequalities by induction for all $n \in \mathbb{N}$.

(a) $5^n + 5 < 5^{n+1}$

(b) $1 + 2 + 3 + \cdots + n \leq n^2$

(c) $\dfrac{1}{\sqrt{1}} + \dfrac{1}{\sqrt{2}} + \dfrac{1}{\sqrt{3}} + \cdots + \dfrac{1}{\sqrt{n}} \geq \sqrt{n}$

(d) $\dfrac{1}{\sqrt{1}} + \dfrac{1}{\sqrt{2}} + \dfrac{1}{\sqrt{3}} + \cdots + \dfrac{1}{\sqrt{n}} \leq 2\sqrt{n}$

4.6.3. (**Harder!**) Let $0 < a < 1$. Prove that for any $n \in \mathbb{N}$, $(1-a)^n < \dfrac{1}{1+n\cdot a}$.

4.6.4. Prove that for every $n \in \mathbb{N}$, $3^{4n+2} + 1$ is divisible by 10.

4.6.5. Prove that for any $n \in \mathbb{N}$, $n^3 + 2n$ is divisible by 3.

4.6.6. Prove that for all $n \in \mathbb{N}$, $4^{2n} - 1$ is divisible by 5.

4.6.7. Let a be an integer different than 1.

Prove, **by induction**, that for any $n \in \mathbb{N}$, $a^n - 1$ is divisible by $a - 1$.

4.6.8. Prove that $\dfrac{1}{3}n^3 + \dfrac{1}{2}n^2 + \dfrac{1}{6}n$ is **an integer** for any $n \in \mathbb{N}$.

4.6.9. Compute the following expressions (obtain a single number).

(a) $\displaystyle\sum_{k=1}^{100} \left[k \cdot (-1)^k\right]$

(b) $\displaystyle\sum_{k=2}^{200} \left(\frac{1}{k} - \frac{1}{k+1}\right)$

(c) $\displaystyle\prod_{k=1}^{69} 2^{k-35}$

(d) $\displaystyle\prod_{i=10}^{99} \frac{i}{i+1}$

4.6.10. Which of the following are true for any $a_1, \ldots, a_n, b_1, \ldots, b_n \in \mathbb{R}$? Justify your answer briefly.

Provide counterexamples for the false statements.

(a) $\left(\displaystyle\sum_{k=1}^{n} a_k\right) \cdot \left(\sum_{k=1}^{n} b_k\right) = \sum_{k=1}^{n} (a_k \cdot b_k)$

(b) $\sum_{k=1}^{n} a_k - \sum_{k=1}^{n} b_k = \sum_{k=1}^{n} (a_k - b_k)$

(c) $\prod_{k=1}^{n} a_k - \prod_{k=1}^{n} b_k = \prod_{k=1}^{n} (a_k - b_k)$

(d) $\left(\prod_{k=1}^{n} a_k\right) / \left(\prod_{k=1}^{n} b_k\right) = \prod_{k=1}^{n} \frac{a_k}{b_k}$, assuming that $b_1, b_2, \ldots, b_n$ are all nonzero.

4.6.11. Prove the following identities.

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

(b) $\sum_{i=1}^{n} (3i - 2) = \frac{n(3n-1)}{2}$ (for $n \geq 1$).

(c) $\prod_{i=2}^{n} \left(1 - \frac{1}{i^2}\right) = \frac{n+1}{2n}$ (for $n \geq 2$).

4.6.12. Let $P(1), P(2), \ldots$ be a sequence of statements. Write down variations (in the style of those on page 92), for proving that $P(n)$ is true for...

(a) all n's which are a multiple of 3.

(b) $n = 2, 5, 8, 11, \ldots$.

(c) $n = 11, 13, 15, 17, \ldots$.

4.6.13. On page 90, we mentioned that Proposition 4.2.3 is a special case of Bernoulli's Inequality (Example (b) on page 86). Show how Bernoulli's Inequality can be indeed obtained from Proposition 4.2.3.

4.6.14. Prove, by induction, that $10^n - 1$ is divisible by 11 for every **even** natural number n.

4.6.15. (a) Show that for any $k \in \mathbb{N}$, if $2^{3k-1} + 5 \cdot 3^k$ is divisible by 11, then $2^{3(k+2)-1} + 5 \cdot 3^{k+2}$ is also divisible by 11.

(b) Which of the following statements is **true**? Explain.

(i) For any **odd** number $n \in \mathbb{N}$, $2^{3n-1} + 5 \cdot 3^n$ is divisible by 11.

(ii) For any **even** number $n \in \mathbb{N}$, $2^{3n-1} + 5 \cdot 3^n$ is divisible by 11.

4.6.16. Let (a_n) be a sequence such that $a_1 = 1$ and $a_{n+1} = a_n + 3n(n+1)$ for $n \in \mathbb{N}$.
Prove that $a_n = n^3 - n + 1$ for $n \in \mathbb{N}$.

4.6.17. Let (a_n) be a sequence given by

$$\begin{cases} a_1 = 3 \\ a_{n+1} = a_n + 6n(n+1) \qquad (\text{for } n \in \mathbb{N}) \end{cases} .$$

Prove, by induction, that $a_n = 2n^3 - 2n + 3$ for any $n \in \mathbb{N}$.

4.6.18. (a) Prove that $3x^2 + 3 \geq (x+1)^2 + 1$ for all real numbers x (can this be done by induction?).

(b) Consider the following recursively defined sequence: $\begin{cases} a_1 = 2 \\ a_{n+1} = 3 \cdot a_n \qquad (\text{for } \ n = 1, 2, \dots) \end{cases}$.

Use part (a) to prove, by induction, that $a_n \geq n^2 + 1$ for all $n \in \mathbb{N}$.

4.6.19. Justify parts (b), (c) and (d) of Proposition 4.2.2, by writing sums and products explicitly (i.e., without Sigmas or Pi's).

4.6.20. Let (x_n) be a sequence given by the following recursion formula:

$$x_1 = 3\ , \quad x_2 = 7 \qquad \text{and} \qquad x_{n+1} = 5 \cdot x_n - 6 \cdot x_{n-1} \qquad \text{for} \quad n \geq 2.$$

Prove that for all $n \in \mathbb{N}$, $x_n = 2^n + 3^{n-1}$.

4.6.21. Consider the following sequence defined recursively:

$$\begin{cases} a_1 = 5, \qquad a_2 = 8, \\ a_{n+1} = 2a_n - a_{n-1} + 2 \qquad \text{for} \quad n \geq 2 \end{cases}$$

Prove that $a_n = n^2 + 4$ for all $n \in \mathbb{N}$.

4.6.22. The sequence (a_n) is defined recursively by $\begin{cases} a_1 = 6 \quad , \quad a_2 = 8 \\ a_n = 4 \cdot a_{n-1} - 4 \cdot a_{n-2} \qquad \text{for} \ \ n > 2 \end{cases}$.

Prove that $a_n = (4 - n) \cdot 2^n$ for all $n \in \mathbb{N}$.

4.6.23. Let (a_n) be a sequence satisfying $a_1 = a_2 = 1$ and $a_n = \frac{1}{2}\left(a_{n-1} + \frac{2}{a_{n-2}}\right)$ for $n \geq 3$.
Prove that $1 \leq a_n \leq 2$ for all $n \in \mathbb{N}$.

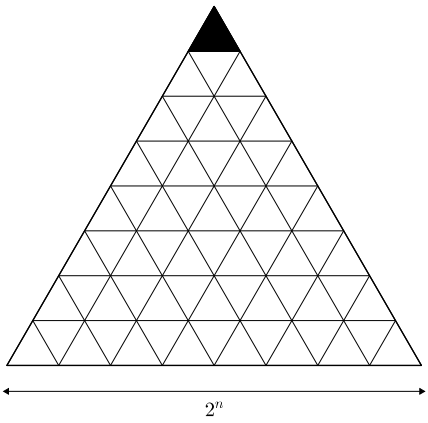
4.6.24. Consider the following recursively defined sequence:

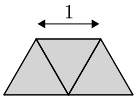
$$a_1 = \frac{5}{2} \qquad , \qquad a_{n+1} = \frac{1}{2} \cdot (a_n + 2) \qquad (\text{for } n \in \mathbb{N}) \ .$$

Prove that $a_n > \frac{1}{2^n}$ for all $n \in \mathbb{N}$.

4.6.25. Let (a_n) be a sequence satisfying $a_n = 2a_{n-1} + 3a_{n-2}$ for $n \geq 3$.
Given that a_1, a_2 are odd, prove that a_n is odd for $n \in \mathbb{N}$.

4.6.26. Consider a regular triangular board of side length 2^n. The board consists of 4^n equilateral triangles of side length 1. Show that, if one of the corner triangles is removed, then the remaining board can be tiled by isosceles trapezoids (as in the diagram), each covering exactly three triangles.





4.6.27. In the proof of Theorem 4.5.4, why couldn't we use the argument in Case 1 for the case where k is odd? Also, why did we have to use strong induction? Which part of the argument cannot be carried out with 'usual' induction?

4.6.28. Prove that every $n \in \mathbb{N}$ can be written as a product of an odd integer and a nonnegative integer power of 2. For instance: $36 = 9 \cdot 2^2$, $80 = 5 \cdot 2^4$, $17 = 17 \cdot 2^0$, etc...
Hint: Use strong induction on n. In the induction step, treat the cases 'k even' and 'k odd' separately.

4.6.29. Let x be a nonzero real number, such that $x + \frac{1}{x}$ is an **integer**.
Prove that for all $n \in \mathbb{N}$, the number $x^n + \frac{1}{x^n}$ is also an integer.

4.6.30. Find the mistake in the following "proof":

> "**Claim**": The numbers $0, 1, 2, 3, \ldots$ are all even.
>
> "**Proof**": We use strong induction to prove the statement 'n is even' for $n = 0, 1, 2, 3, \ldots$.
>
> Base case: $n = 0$ is an even number, hence the statement is true for $n = 0$.
>
> Assume that the statement is true for $n = 0, 1, 2, \ldots, k$, and consider $n = k + 1$.
>
> By assumption, both 1 and k are even numbers, and hence so is their sum $k + 1$. It thus follows that the statement holds for all $n = 0, 1, 2, 3, \ldots$.

4.6.31. Find the mistake in the following "proof":

"**Claim**": For all $n \in \mathbb{N}$, we have $5^n = 5$.

"**Proof**": For $n = 1$, we have $5^1 = 5$, which proves the base case. Now assume that the claim holds true for $n = 1, 2, \ldots, k$. Then we have

$$5^{k+1} = \frac{5^k \cdot 5^k}{5^{k-1}} = \frac{5 \cdot 5}{5} = 5$$

(we used the induction hypothesis in the second equality). This proves the $n = k+1$ case. By strong induction, the claim follows.

4.6.32. What do you think about the following proof of the statement "**Every person is bald**"?

A person with a single hair is clearly bald, which confirms our base case. Now, assume that a person with k hairs is bald. Then obviously, adding one more hair to a bald person will leave that person bald. By induction, it follows that any person with n hairs is bald. In other words, everyone is bald!

4.6.33. (**Harder!**) **The General Arithmetic-Geometric Mean Inequality**. In this exercise, we prove the general version of the AGM inequality, stating that for any $x_1, x_2, \ldots, x_n \geq 0$, we have

$$\sqrt[n]{x_1 \cdot x_2 \cdots x_n} \leq \frac{x_1 + x_2 + \cdots + x_n}{n}.$$

The proof uses induction, but is a bit tricky. We break it into several steps.

(a) Prove that if x and y are real numbers, satisfying $0 \leq x \leq 1 \leq y$, then $x + y \geq xy + 1$.
(Hint: Note that $x - 1 \leq 0$ and $y - 1 \geq 0$.)

(b) Assume that $a_1, a_2, \ldots, a_n$ are non-negative real numbers (with $n \geq 2$), whose product is 1 :

$$a_1 \cdot a_2 \cdots a_n = 1.$$

Explain why $a_i \leq 1 \leq a_j$ for some $i \neq j$.

(c) Prove, by induction on n, the following claim:
If $a_1, \ldots, a_n$ are non-negative real numbers, with $a_1 \cdots a_n = 1$, then $a_1 + \cdots + a_n \geq n$.
(Hints: To carry the induction step, note that a product of $k+1$ numbers can be thought of as a product of k numbers:

$$a_1 \cdot a_2 \cdot a_3 \cdots a_{k+1} = (a_1 \cdot a_2) \cdot a_3 \cdots a_{k+1}.$$

You will also need to use parts (a) and (b) in your proof.)

(d) Finally, prove the general AGM inequality.

To do so, note that if one of the x_i's is zero, the AGM inequality follows immediately (why?). If all the x_i's are positive, use part (c) with

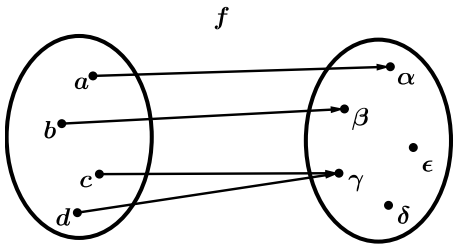
$$a_1 = \frac{x_1}{\sqrt[n]{x_1 \cdot x_2 \cdots x_n}} \quad , \quad a_2 = \frac{x_2}{\sqrt[n]{x_1 \cdot x_2 \cdots x_n}} \quad , \quad a_3 = \frac{x_3}{\sqrt[n]{x_1 \cdot x_2 \cdots x_n}} \quad , \quad \text{etc.}$$

Chapter 5

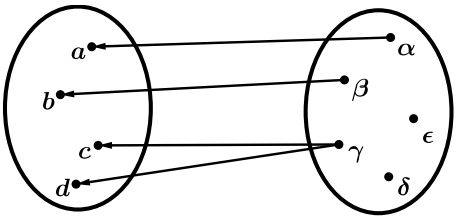
Bijections and Cardinality

5.1 Injections, Surjections and Bijections

Consider a function $f\colon \{a,b,c,d\} \to \{\alpha,\beta,\gamma,\delta,\epsilon\}$, given by the following diagram.[1]



Indeed, this diagram defines a function from the set $\{a,b,c,d\}$ to the set $\{\alpha,\beta,\gamma,\delta,\epsilon\}$ (see Definition 2.2.1). Suppose now, that we decide to reverse the arrows, and obtain the following diagram.



Does the new diagram define a function from $\{\alpha,\beta,\gamma,\delta,\epsilon\}$ to $\{a,b,c,d\}$? No, it does not, for the following reasons:

- In the new diagram, both c and d are assigned to the element γ. Functions **cannot** assign two images to an element in the domain.
- In the new diagram, no elements are assigned to δ and ϵ. However, a function **must** assign an image to **every** element in its domain.

[1] $\alpha,\beta,\gamma,\delta,\epsilon$ are the first five letters of the Greek alphabet: alpha, beta, gamma, delta and epsilon.

In order to be able to invert a function, none of these phenomena can occur, which leads to the following definitions.

Definition 5.1.1. Let $f\colon A \to B$ be a function.

(a) f is **injective** (or an **injection**, or a **one-to-one function**), if for every $y \in B$, there is **at most** one $x \in A$, for which $f(x) = y$.

(b) f is **surjective** (or a **surjection**, or an **onto function**), if for every $y \in B$, there is **at least** one $x \in A$, for which $f(x) = y$.

(c) f is **bijective** (or a **bijection**), if it is **both** injective and surjective.

Remarks.

- In the previous example, f is **neither injective nor surjective**. We have $f(c) = f(d) = \gamma$, and hence f is **not** an injection, and as δ and ϵ are not images of elements in the domain, f is also **not** a surjection.

- There are many ways to state the definition of injectivity and surjectivity. For instance:

  - f is **injective**, if for every $x_1, x_2 \in A$, $x_1 \neq x_2$ implies $f(x_1) \neq f(x_2)$ (i.e., f maps distinct elements in the domain to distinct elements in the codomain).
  - f is **injective**, if for every $x_1, x_2 \in A$, $f(x_1) = f(x_2)$ implies $x_1 = x_2$ (this is simply the contrapositive of the implication $x_1 \neq x_2 \ \Rightarrow \ f(x_1) \neq f(x_2)$).
  - f is **surjective**, if $f(A) = B$ (i.e., if the image of f is the whole codomain B).
  - f is a **bijection**, if for **any** $y \in B$, there is **exactly one** $x \in A$, for which $f(x) = y$.

  Convince yourself that these definitions are equivalent to the ones given in Definition 5.1.1. We will freely use the most convenient formulation in our proofs and examples.

In order to be able to invert a function (namely, reverse the arrows and obtain a function that goes in the opposite direction), our initial function must be both injective and surjective. It other words, it must be a bijection. Let us formalize this idea.

Definition 5.1.2. Let $f\colon A \to B$ be a bijection. Then the **inverse** of f is a function $g\colon B \to A$, that assigns to any $y \in B$, the only $x \in A$ for which $f(x) = y$. We denote the inverse function by f^{-1}.

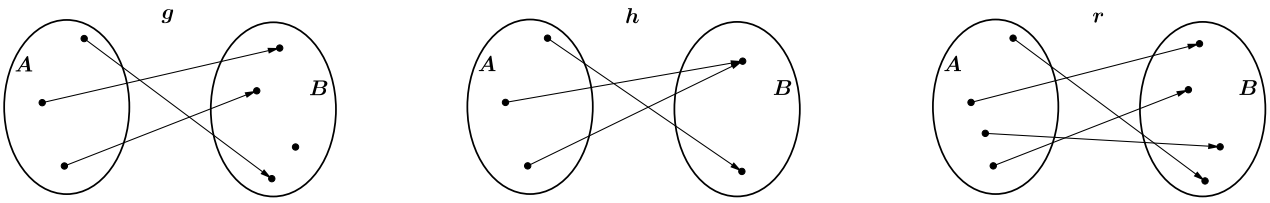
Remark. The definition of the inverse may sound confusing. All it says is that if f maps an element $a \in A$ to some $b \in B$, then f^{-1} maps b back to a.

$$f(a) = b \qquad \Leftrightarrow \qquad f^{-1}(b) = a$$

In other words, f^{-1} undoes the effect of f.

Examples.

(a) g, h and r are functions from a set A to another set B, given by the following diagrams.



The function g is **injective**, as there are no two elements in A which are mapped to the same element in B (i.e., there are no two arrows pointing to the same element in B). However, g is **not surjective** as there is a "lonely" element in B, which is not the image of an element in A.

The function h is **onto**, as every element in B is an image, but it is **not one-to-one**, since there are two arrows pointing to the same element in B.

Finally, r is both **an injection** and **a surjection**, as every element in B is the image of **exactly one** element in A. In other words, r is a **bijection** (and hence can be inverted).

(b) $f\colon \mathbb{R} \to \mathbb{R}$, $f(x) = \dfrac{1}{1+x^2}$.

f is **not an injection**, as we can easily find two x's with the same image. For instance, $f(2) = f(-2) = \frac{1}{5}$.

f is **not a surjection** either, since $f(x) > 0$ for all $x \in \mathbb{R}$, and hence there is no $x \in \mathbb{R}$ for which $f(x) = -1$.

(c) $g\colon \mathbb{N} \to \mathbb{N}$, $g(n) = n^3 + 1$.

Note that the domain of g is the set of natural numbers, and so for every n in the domain, $g(n) = n^3 + 1 \geq 1 + 1 = 2$. This means that for all $n \in \mathbb{N}$, $g(n) \neq 1$, which shows that g is **not surjective**.

g is, however, **an injection**. We prove this by showing that distinct elements in the domain are mapped to distinct images. Indeed, if $n_1 \neq n_2$ are two natural numbers, then

$$n_1^3 \neq n_2^3 \qquad \Rightarrow \qquad n_1^3 + 1 \neq n_2^3 + 1 \qquad \Rightarrow \qquad g(n_1) \neq g(n_2).$$

(d) We proceed with a less-standard example.

Let $\mathcal{P}$ denote the set of all **nonzero polynomials with real coefficients**. Recall that a polynomial is a sum of terms of the form $a \cdot x^n$, where a is a real number, and n is a nonnegative integer. For instance, $5x^3 + 8x^7 - 4$ and $x - \frac{1}{2}x^4$ are polynomials, while $\sqrt{x}$ and $\frac{1}{x} + 2x$ are not.

Every polynomial has a **degree**, which is the highest exponent appearing in the polynomial, with a nonzero coefficient. We often denote the degree of a polynomial by deg(). For instance,

$$\deg(5x^3 + 8x^7 - 4) = 7 \quad , \qquad \deg\left(x - \frac{1}{2}x^4\right) = 4 \ , \ \ \deg(2x + 7) = 1 \ , \ \ \deg((2 + x^2)^{13}) = 26 \ , \ \ \ldots.$$

Now, define a function D from the set of nonzero polynomials to the set of nonnegative integers, that assigns to each polynomial $p(x)$, its degree:

$$D\colon \mathcal{P} \to \mathbb{N} \cup \{0\} \qquad , \qquad D(p(x)) = \deg(p(x)).$$

Is D injective? Is it surjective?

Clearly, there are polynomials of the same degree. For example, $x^3 + x + 2$ and $5x^3 - 4x^2$ are both cubic polynomials, and so have degree 3:

$$D(x^3 + x + 2) = D(5x^3 - 4x^2) = 3.$$

As D assigns the same image to two distinct elements, it is **not injective**.

What about surjectivity? If n is a nonnegative integer, can we always find a polynomial with degree n? Of course we can. $p(x) = x^n$ is a polynomial of degree n. In other words, $D(x^n) = n$ for $n = 0, 1, 2, \ldots$, which shows that D **is a surjective function**.

(e) $h\colon \mathbb{R} \setminus \{-1\} \to \mathbb{R} \setminus \{1\}$, $h(x) = \dfrac{x}{x+1}$.

Again, we would like to find out whether h is injective, surjective, neither or both. One way to do that, is to draw the graph of the function (using calculus or a graphing software), and try to get some intuition regarding surjectivity and injectivity. Then, we can proceed to proving our conjectures.

In this example, h **is a bijection**. Here is the proof.

To show that h is one-to-one, we start by assuming that $h(x_1) = h(x_2)$ for some $x_1, x_2 \neq -1$. We argue that $x_1 = x_2$ as follows.

$$h(x_1) = h(x_2) \qquad \Rightarrow \qquad \frac{x_1}{x_1 + 1} = \frac{x_2}{x_2 + 1} \qquad \Rightarrow \qquad x_1x_2 + x_1 = x_2x_1 + x_2 \qquad \Rightarrow \qquad x_1 = x_2.$$

Now we show that h is surjective. Let $y \neq 1$. Our task is to show that $h(x) = y$ for some x in the domain of h. We do that by solving the equation $h(x) = y$ for x:

$$\frac{x}{x+1} = y \qquad \Rightarrow \qquad x = xy + y \qquad \Rightarrow \qquad x(1-y) = y \qquad \Rightarrow \qquad x = \frac{y}{1-y}.$$

Note that, in the final step, we could divide by $1 - y$ since $y \neq 1$. Also note that $\frac{y}{1-y} \neq -1$ (why?), and so x lies in the domain of h. Overall, we showed that any y in the codomain of h has an 'x', which implies that h is surjective.

The bijectivity of h implies that h is invertible (i.e., can be inverted). For the inverse function, the domain and codomain are switched, and so $h^{-1}\colon \mathbb{R} \setminus \{1\} \to \mathbb{R} \setminus \{-1\}$. How can we find a formula for h?

Remember that inverting a function means (informally) switching the roles of x and y, and so our goal is to express x as a function of y. In fact, we have already done that, when we showed that h is surjective:

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

We therefore conclude that $h^{-1}(y) = \dfrac{y}{1-y}$ (or $h^{-1}(x) = \dfrac{x}{1-x}$, after switching back to x).

(f) One of the commonly-used trigonometric functions is the sine function.

$$f\colon \mathbb{R} \to \mathbb{R} \qquad , \qquad f(x) = \sin x.$$

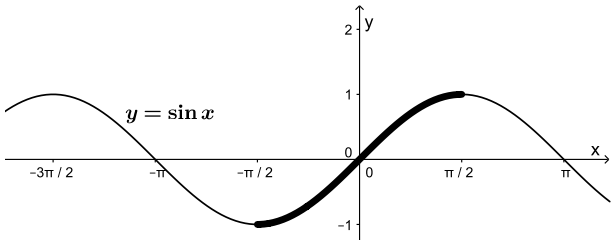
The graph of f looks like a wave, which clearly shows that f is not injective nor surjective.

More formally, $\sin(0) = \sin(\pi) = 0$, and thus f is **not one-to-one**, and $\sin(x) \neq 2$ for any x, which implies that f is **not onto either**.

However, we can easily "turn" f into a bijection, by restricting its domain and codomain to the intervals $\left[-\frac{\pi}{2}, \frac{\pi}{2}\right]$ and $[-1, 1]$, respectively. Consequently, the function

$$g\colon \left[-\frac{\pi}{2}, \frac{\pi}{2}\right] \to [-1, 1] \qquad , \qquad g(x) = \sin x$$

is a bijection (this can be derived from the geometric definition of the sine function, or seen by simply looking at the restricted graph).



The inverse function, g^{-1}, is commonly called **the inverse sine function**, or **the arcsine function**:

$$g^{-1}\colon [-1,1] \to \left[-\frac{\pi}{2}, \frac{\pi}{2}\right] \qquad , \qquad g^{-1}(x) = \sin^{-1} x = \arcsin x.$$

(g) A similar discussion can be carried out for the function $f(x) = x^2$. As a function from $\mathbb{R}$ to $\mathbb{R}$, f is neither injective nor surjective. However, if we view f as a function from $[0,\infty)$ to $[0,\infty)$, then it is a bijection, and its inverse is the square-root function. In other words, the function

$$f\colon [0,\infty) \to [0,\infty) \qquad , \qquad f(x) = x^2$$

is invertible, and its inverse is the function

$$f^{-1}\colon [0,\infty) \to [0,\infty) \qquad , \qquad f^{-1}(x) = \sqrt{x}.$$

Sometimes, it may be difficult to prove directly that a function is injective or surjective. In these cases we may use indirect methods for doing so. For instance, monotone functions are one-to-one. Here is a definition you might have seen in your calculus class.

Definition 5.1.3. Let $A \subseteq \mathbb{R}$, and $f\colon A \to \mathbb{R}$ a function.
We say that f is a **strictly increasing** (respectively **decreasing**) **function**, if for every $x_1 < x_2$ in A, we have $f(x_1) < f(x_2)$ (respectively $f(x_1) > f(x_2)$).
A function f is **strictly monotone**, if it is either strictly increasing or strictly decreasing.

Remarks.

- The wording of this definition might seem weird at first, if you are not used to the word 'respectively', which allows us to compress definitions and mathematical statements. If we read the first sentence of Definition 5.1.3 without the parentheses, we obtain the definition of a **strictly increasing** function:

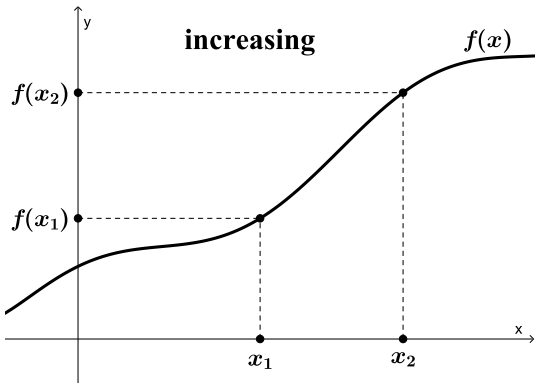
  "f is a **strictly increasing function**, if for every $x_1 < x_2$ in A, we have $f(x_1) < f(x_2)$."

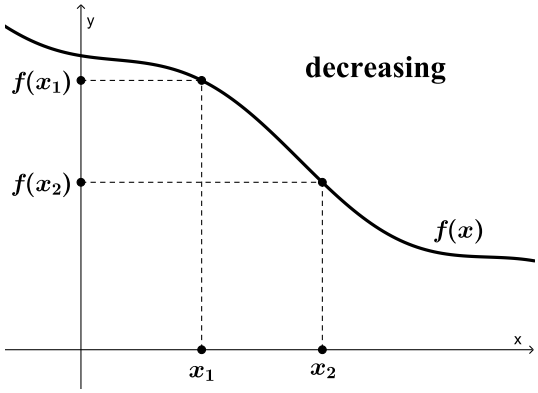
  However, if we replace 'increasing' and '$f(x_1) < f(x_2)$' with the content in parentheses, we get the definition of a **strictly decreasing** function:

  "f is a **strictly decreasing function**, if for every $x_1 < x_2$ in A, we have $f(x_1) > f(x_2)$."

- Informally, the definition says that for a strictly increasing function, the y-values **increase** as we increase the x-values. For a strictly decreasing function, y-values **decrease** as the x-values increase

(see diagrams).





We are now ready to prove the following proposition.

Proposition 5.1.4. Let $A \subseteq \mathbb{R}$. If $f\colon A \to \mathbb{R}$ is a strictly monotone function, then f is injective.

Proof. We assume that f is a strictly monotone function, and we prove that f is one-to-one, by showing that $x_1 \neq x_2$ implies $f(x_1) \neq f(x_2)$ for arbitrary $x_1, x_2 \in A$.

As $x_1 \neq x_2$, either $x_1 < x_2$ or $x_2 < x_1$. Since f is strictly monotone, either $f(x_1) < f(x_2)$ or $f(x_1) > f(x_2)$. In either case, we see that $f(x_1) \neq f(x_2)$, and hence f is injective, as needed.[2] □

Example 5.1.5. The function $f\colon (-1,1) \to \mathbb{R}$, $f(x) = \dfrac{2x}{1-x^2}$ is injective, but it may require some work to prove this fact directly from Definition 5.1.1. However, using calculus, we see that the derivative is always positive:

$$f'(x) = \frac{2(1+x^2)}{(1-x^2)^2} > 0 \qquad \text{for } -1 < x < 1.$$

We conclude that f is strictly increasing, and hence, by Proposition 5.1.4, f is one-to-one.[3]

5.2 Compositions

Any two functions f and g, with codomain $\mathbb{R}$, can be added, subtracted, multiplied and divided to create new functions. There is, however, one more important operation that can be performed on functions – we can compose them! Namely, apply one after the other, and we can do that even when their domain and codomain are **not** sets of numbers. Here is the definition.

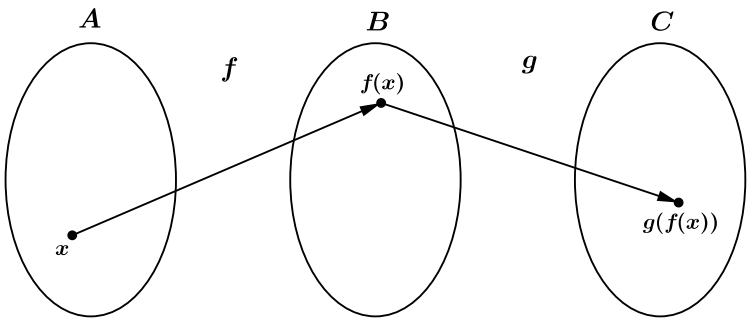
[2] This proof is extremely short. Do not let that fool you! Make sure you fully understand the argument. What did we assume? What did we prove? Which definitions did we use? Did we cover all cases?

[3] In these notes, we normally try to avoid any calculus. This example is an exception.

Definition 5.2.1. Let $f\colon A \to B$ and $g\colon B \to C$ be two functions.

The **composition** of g with f, denoted as $g \circ f$, is the function from A to C, given by

$$g \circ f(x) = g(f(x)) \qquad \text{for } x \in A.$$



Note that the only requirement for composition is that the codomain of f is equal to the domain of g.[4]

Examples.

(a) Consider the functions $f, g\colon \mathbb{R} \to \mathbb{R}$, given by

$$f(x) = \frac{1}{1+x^2} \qquad \text{and} \qquad g(x) = e^x.$$

If we compose g with f, we obtain the function $g \circ f(x) = g(f(x)) = g\left(\frac{1}{1+x^2}\right) = e^{\frac{1}{1+x^2}}$.

We can also compose f with g (i.e., apply g first), to get $f \circ g(x) = f(g(x)) = f(e^x) = \frac{1}{1+(e^x)^2}$.

As we can see, $g \circ f \neq f \circ g$, which means that when we compose two functions, **the order of composing matters**.

(b) Let f and g be the following two functions:

$$\begin{cases} f\colon \mathbb{N} \times \mathbb{N} \to \mathbb{Z} \\ f(m,n) = m - n \end{cases} \qquad \begin{cases} g\colon \mathbb{Z} \to \mathbb{R} \\ g(k) = \sqrt{|k|} \end{cases}.$$

We can compose g with f, as the codomain of f is the domain of g.

We get the function $g \circ f\colon \mathbb{N} \times \mathbb{N} \to \mathbb{R}$, given by

$$g \circ f(m,n) = g(f(m,n)) = g(m-n) = \sqrt{|m-n|}.$$

Note that the composition $f \circ g$ is **undefined**, as the codomain of g (which is $\mathbb{R}$), is different than the domain of f (which is $\mathbb{N} \times \mathbb{N}$).

[4] In fact, it is enough to require that the **image of** f is **contained** in the **domain of** g.

What is the relation between composition, surjectivity and injectivity? Does injectivity of functions imply the injectivity of their composition? What if a composition is known to be surjective? Does that mean that the functions themselves must be surjective? These sort of questions arise frequently in mathematical proofs. Our next example deals with one such question.

Example. Let $f\colon A \to B$ and $g\colon B \to C$ be two functions. Prove that if $g \circ f$ is injective, then so is f.

Proof. Note that we are given that $g \circ f$ is one-to-one. Our task is to prove that the inner function, f, is one-to-one as well. We do so by showing that $f(x_1) = f(x_2)$ implies $x_1 = x_2$ for all $x_1, x_2 \in A$.
If $f(x_1) = f(x_2)$, then, by applying the function g on both sides of this equality, we get

$$g(f(x_1)) = g(f(x_2)) \qquad \text{or} \qquad g \circ f(x_1) = g \circ f(x_2).$$

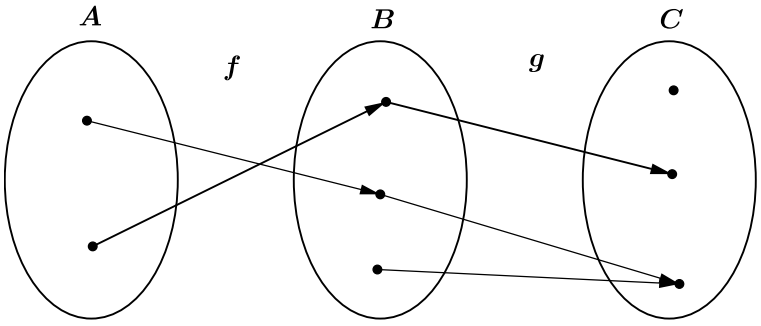
However, $g \circ f$ is known to be injective, and hence we conclude, from $g \circ f(x_1) = g \circ f(x_2)$, that $x_1 = x_2$. This completes the proof of injectivity of f, as needed. □

The last example raises the following question.

Question: If the composition $g \circ f$ is injective, must also g be injective?

The answer can be either 'yes' or 'no'. If g must be injective, we could probably prove it using an argument similar to the one we used for showing that f is injective. Otherwise, we should look for a counterexample showing that g need not be injective.

Indeed, it is not hard to construct functions f and g, such that $g \circ f$ is injective, while g is not. The following diagram describes such a counterexample. Can you see that $g \circ f$ is injective, while g is not? Therefore, the answer to the above question is: **No**.



We end this section with a proposition, stating that injectivity, surjectivity and bijectivity are properties which are preserved under composition. The last part of the proposition gives a formula for computing the inverse of a composed function.

The proposition is important for two reasons. First, once proved, it can be used in other proofs and arguments. Secondly, proving such propositions is excellent practice in using the definitions of composition, injectivity and surjectivity, and in building logical arguments.

Proposition 5.2.2.

(a) The composition of two injections is an injection.

(b) The composition of two surjections is a surjection.

(c) The composition of two bijections is a bijection.

(d) If $f\colon A \to B$ and $g\colon B \to C$ are two bijections, then $(g \circ f)^{-1} = f^{-1} \circ g^{-1}$.
(In other words, the inverse of a composition of two bijections, is the composition of their inverses, in reverse order.)

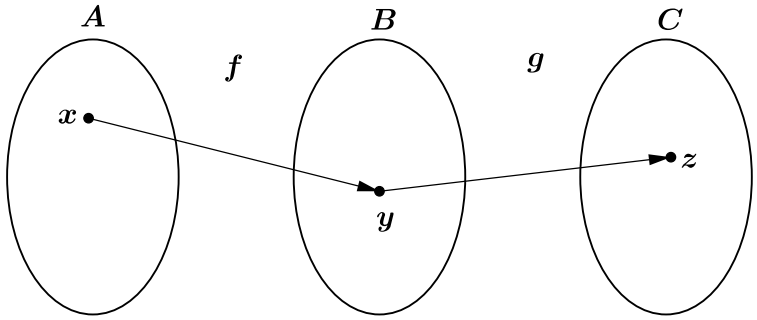
Proof. (a) The proof of this part is left as an exercise (see Exercise 5.6.26).

(b) Let $f\colon A \to B$ and $g\colon B \to C$ be two surjections. Our task is to prove that $g \circ f\colon A \to C$ is also surjective (see diagram).

Let $z \in C$ be an arbitrary element. As g is surjective, $z = g(y)$ for some $y \in B$. Similarly, f is surjective, and so $y = f(x)$ for some $x \in A$. Over all, we have

$$z = g(y) = g(f(x)) = g \circ f(x),$$

which shows that $g \circ f$ is surjective (every $z \in C$ is the image of some $x \in A$ under $g \circ f$).



(c) This part follows right away from parts (a) and (b), as bijections are functions which are both injective and surjective.

(d) Note that both functions $(g \circ f)^{-1}$ and $f^{-1} \circ g^{-1}$ are functions from C to A. Our task is to show that they are equal to each other. I.e., we need to prove that $(g \circ f)^{-1}(z) = f^{-1} \circ g^{-1}(z)$ for any $z \in C$.

Fix $z \in C$, and denote $y = g^{-1}(z)$ and $x = f^{-1}(y)$. Therefore,

$$f^{-1} \circ g^{-1}(z) = f^{-1}(g^{-1}(z)) = f^{-1}(y) = x.$$

On the other hand, as $g(y) = z$ and $f(x) = y$, we have

$$g \circ f(x) = g(f(x)) = g(y) = z \qquad \Rightarrow \qquad (g \circ f)^{-1}(z) = x.$$

We proved that $(g \circ f)^{-1}(z) = f^{-1} \circ g^{-1}(z) = x$, which completes the proof of part (d).

□

5.3 Cardinality

In everyday life, we are often required to compare the number of objects of two sets, and decide which one is larger (or whether they are of equal size). For instance, when scheduling classes at university, we need to make sure that the number of students, in a particular class, does not exceed the number of seats available in the classroom. In a parking lot, the attendant needs to ensure that the number of cars is not larger than the number of available parking spots.

Comparing the size (i.e., the number of elements) of two sets, seems to be a straightforward task. Nevertheless, in order to fully understand the notion of cardinality, it is crucial that we take a closer look at the process of comparing the number of objects in two sets.

Question: How do we compare the size of two finite sets A and B?

The quick answer is (of course): **By counting**. All we have to do is count the number of elements in A and B, and compare the two resulting numbers. If the numbers are equal, the sets have the same size. Otherwise, one set is larger than the other.

There is, however, another method for comparing sets, which is more important and useful in mathematics.

Imagine, for instance, that in a particular university classroom, every student occupies exactly one seat. If some of the seats in the room are unoccupied, we can immediately conclude that the number of seats is larger than the number of students present in class. There is no need to do any counting at all!

Here is another example. In a dance club, a group of teenagers are dancing on the dance floor. If every guy dances with exactly one girl (and no one is left alone), we can conclude right away, that the number of

guys on the dance floor equals the number of girls. Again – no counting is needed. The group of dancers form pairs, each of which consists of exactly one guy and one girl. This perfect matching implies that there is an equal number of guys and girls.

The matching method is interesting, for several reasons. In primitive cultures, where counting methods were hardly developed, people often used (and are still using) matching. A farmer may pair goats with bunches of carrots, to make sure he has enough to feed the whole herd. It is also known that children use the matching method before they learn how to count. For instance, a two year old may help his mom set up the table for dinner, by placing one spoon next to each plate, and then realize that he is short of spoons (or has too many). This task does not require the ability to count.

Our interest in the matching method is for a different reason though. In mathematics, we often need to deal with **infinite sets** (such as intervals on the number line, regions in the plane, etc.). However, for infinite sets, the counting method cannot be applied. Which set is larger, the integers ($\mathbb{Z}$) or the closed interval $[0,1]$? Both sets are infinite, and counting the number of elements in each is not possible. Fortunately, the pairing (or matching) method can be carried out, and that is how we proceed.[5]

Definition 5.3.1. Two sets A and B are said to have **the same cardinality**, if there is a bijection between them.

Remarks.

- "Cardinality" is the accurate (and more general) term used in mathematics for "the number of elements" in a set. It is more precise than "size", which may refer to length or area, and not necessarily to the number of objects.
- By "a bijection between them", we mean either a bijection from A to B, or from B to A. In fact, if there is a bijection going from A to B, then its inverse is a bijection from B to A.

As you will soon realize, it is extremely important to work closely with the definition of cardinality when facing the problem of comparing two sets, and especially when dealing with infinite sets. We have very little experience in working with infinite sets in our everyday life, and relying on our intuition is too risky.

Examples. (a) The sets

$$A = \{\text{Sunday, Monday, Tuesday, Wednesday, Thursday, Friday, Saturday}\} \quad \text{and} \quad B = \{1,2,3,4,5,6,7\}$$

[5] In fact, even the "counting method" can be seen as matching. Counting objects in a set is nothing but pairing them with the natural numbers: We assign the number 1 to an element, 2 to another, and so on.

have the same cardinality, as they both contain seven elements. More formally, we can easily construct a bijection $f\colon A \to B$ by defining

$$f(\text{Sunday}) = 1 \quad , \quad f(\text{Monday}) = 2 \quad , \ \dots \ , \quad f(\text{Saturday}) = 7 \ ,$$

and so by Definition 5.3.1, the two sets **have the same cardinality**.

(b) The sets $A = \{x, y, z\}$ and $B = \{\phi\}$ do not have the same cardinality. This is clear if the counting method is used (A has three elements, and B has only one), but we should get into the habit of relying on Definition 5.3.1 in our arguments (and not on counting).

We claim that there is no way of forming a bijection between A and B, as any function $f\colon A \to B$ must assign the element ϕ to each element in A (and so $f(x) = f(y) = f(z) = \phi$). This shows that there are no **injective** functions from A to B, and hence these sets **do not have the same cardinality**.

(c) Consider the sets $A = \{1, 2, \dots, 100\}$ and $B = \mathbb{R}$ (all real numbers). It is quite clear that A and B do not have the same cardinality, as A is finite and B is infinite. Can we use Definition 5.3.1 to confirm our intuition? If $f\colon A \to B$ is a function, then its image, $f(A)$, consists of at most 100 elements. As $\mathbb{R}$ contains more than 100 numbers, f cannot be surjective. Therefore, there is no bijection between A and B, and the two sets **do not have the same cardinality**.

(d) Let $A = \mathbb{N} = \{1, 2, 3, \dots\}$ and $B = \{-1, -2, -3, \dots\}$. Note that both sets are infinite, and so counting elements is not an option. However, it is quite easy to "pair" elements of one set with elements from the other set. Define a function $f\colon A \to B$ by $f(n) = -n$.
It is not hard to show that f is a bijection from A to B, which implies that the two sets **have the same cardinality**.

(e) This example, though similar to the previous one, may surprise you at first. Let A be the set of natural numbers, and B the set of **even** positive integers:

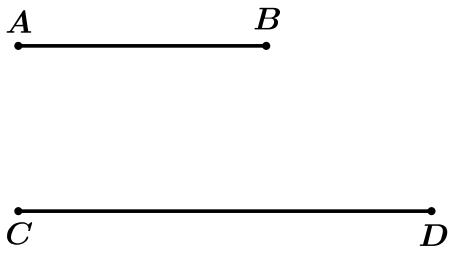
$$A = \mathbb{N} = \{1, 2, 3 \dots\} \qquad \text{and} \qquad B = \{2, 4, 6 \dots\}.$$

Do these sets have the same cardinality? Well, intuitively, it may seem like B is "smaller" than A, as it is a proper subset of A. This intuition, however, is based on our everyday experience with finite sets, and so we must be very careful in applying it to infinite sets. Referring to Definition 5.3.1, we should ask ourselves: **Can we form a bijection from A to B (or from B to A)**? And the answer is – yes, we can. If we define

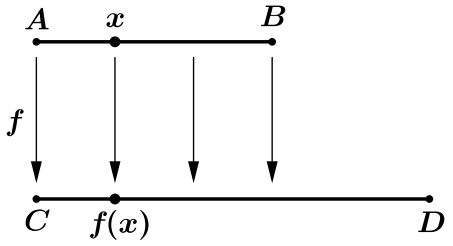
$$g\colon A \to B \qquad \text{by} \qquad g(n) = 2n,$$

then g is a bijection from A to B, which proves that A and B **do have the same cardinality**.
In other words, **there are as many natural numbers as even positive integers**! Again – this is very counterintuitive (at first). With time and practice, your intuition will adjust to align with the formal definition of cardinality. Meanwhile, consider your intuitions about counting with suspicion, coming as they do from your interactions with finite objects. Work closely with Definition 5.3.1 and you'll be safe.

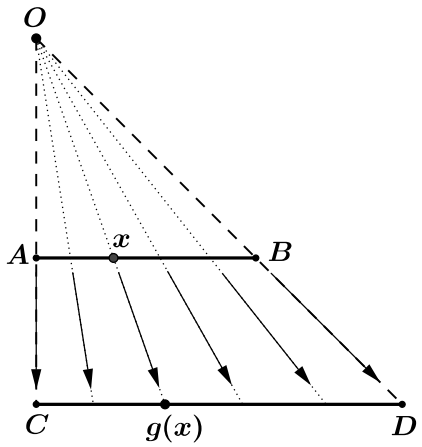
(f) Our next example is more geometric in nature. Consider the two line segments AB and CD drawn below, and let us think of them as sets of points in the plane.



Do AB and CD (as sets of points) have the same cardinality? Clearly, AB is shorter than CD. Does that imply that the line segments do not have the same cardinality (according to our definition)? One way to construct a function $f\colon AB \to CD$ is to assign, to each point in AB the point right below it on CD, as the following diagram suggests.



The function f is a one-to-one function, but not onto (why?). Does that mean that AB and CD do not have the same cardinality? No, it does not. Read carefully Definition 5.3.1 one more time. Two sets have the same cardinality if **there exists** a bijection between them. The fact that we were able to construct a function which is not bijective does not imply that a bijection cannot exist.
In fact, there is a way of forming a bijection between the two line segments! First, we draw line segments AC and BD, then extend them until they meet at some point O (see diagram).

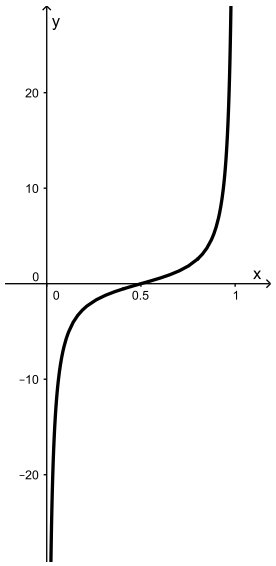


Then, for every point x on AB we assign the point $g(x)$ on CD, obtained by extending the line segment from O to x until it hits CD. This way, we obtain a bijection from AB to CD, which shows that **the two segments have the same cardinality**.

This example shows that cardinality and length are two different notions[6]. Two segments may have the same cardinality (as sets of points), even though they are of different length.

(g) Do the open interval $(0,1)$ and the set of real numbers $\mathbb{R}$ have the same cardinality? Namely, can we form a bijection between the two sets?

Here, our knowledge from calculus and pre-calculus may be handy. A bijection from $(0,1)$ to $\mathbb{R}$ will need to "go to plus and minus infinity", and this can be achieved by using vertical asymptotes. For instance, the following graph represents a bijection from $(0,1)$ to $\mathbb{R}$.



[6]We do not attempt to define precisely the notion of 'length'. This notion will be developed formally in advanced courses in real analysis and measure theory.

We can also provide a closed formula for such a bijection. Functions such as $f(x) = \dfrac{1}{1-x} - \dfrac{1}{x}$ and $g(x) = \tan[\pi(x - 0.5)]$ have a graph that is similar (on $(0, 1)$) to the one in the diagram.[7]

We conclude that the open interval $(0, 1)$ and the whole real line **have the same cardinality** (i.e., have the "same number of elements").

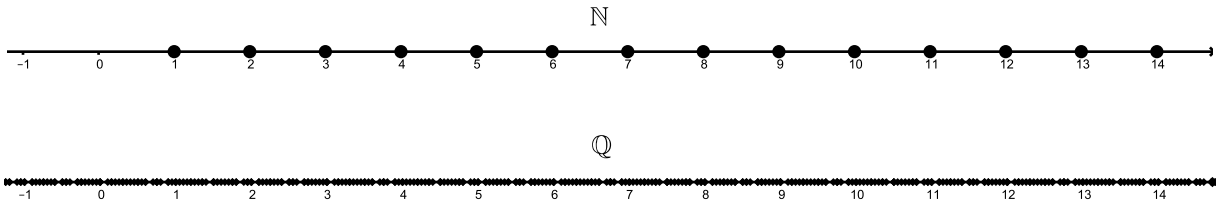
5.4 Cardinality Theorems

In this section, we present proofs for a few fundamental theorems about cardinality. First we prove that the rational and natural numbers are sets of the same cardinality. Then, we show that the real and the natural numbers do not have the same cardinality. Finally, we discuss Cantor's Theorem on cardinality of power sets.

As you will see, the proofs of the theorems are more sophisticated than previous proofs discussed in these notes, and involve clever ideas and unusual constructions. Moreover, the statements themselves are deep, often counterintuitive, and with far-reaching consequences. The power of a mathematical proof will be fully felt and demonstrated in this section.

Theorem 5.4.1. *The set of natural numbers, $\mathbb{N}$, and the set of rational numbers, $\mathbb{Q}$, have the same cardinality.*

In other words, there are as many rational numbers as natural numbers. Geometrically, the natural numbers can be thought of as an infinite sequence of dots on the number line (with consecutive dots being one unit apart), while the rationals are occupying every region on the number line (there are infinitely many rationals in every interval). We say that $\mathbb{Q}$ is **dense** in $\mathbb{R}$.[8]



Looking at the above diagrams, it may by quite difficult to believe that the two sets of numbers are of the same size. You might wonder – maybe every two infinite sets have the same cardinality? This question will be answered by the next theorem. Nevertheless, leaving intuition aside, if indeed $\mathbb{N}$ and $\mathbb{Q}$ have the

[7] The fact that f and g are bijections can be proved using calculus, and properties of the tangent function.

[8] A set of real numbers is said to be **dense**, if it has nonempty intersection with every open interval.

same cardinality, how would one go about proving it? Namely, how can we form a bijection between the rational and the natural numbers. There is no "first rational number" which can be assigned to the number 1 (and then "a second" for 2, etc.). The proof below uses a surprising and clever strategy to construct a bijection. Here it is.

Proof. The proof involves several steps.

Step 1 – Defining A_k. We define, for each $k \in \mathbb{N}$, the set

$$A_k = \left\{ \frac{a}{b} : a, b \in \mathbb{N} \quad \text{and} \quad a+b=k \right\}.$$

For instance, if $k = 4$ then A_k is the set of all fractions with positive numerator and denominator, that add up to 4:

$$A_4 = \left\{ \frac{1}{3}, \frac{2}{2}, \frac{3}{1} \right\}.$$

If $k = 7$, we get the set

$$A_7 = \left\{ \frac{1}{6}, \frac{2}{5}, \frac{3}{4}, \frac{4}{3}, \frac{5}{2}, \frac{6}{1} \right\},$$

and so on (what is A_1?). Looking carefully at a few more special cases, we conclude that each A_k contains exactly $k-1$ elements, with numerators going from 1 to $k-1$ (and corresponding denominators, going from $k-1$ to 1)

$$A_k = \left\{ \frac{1}{k-1}, \frac{2}{k-2}, \frac{3}{k-3}, \ldots, \frac{k-1}{1} \right\}.$$

Step 2 – Forming a Sequence. We arrange the elements of the sets A_k in an infinite sequence, as follows.

$$\underbrace{\frac{1}{1}}_{A_2}, \underbrace{\frac{1}{2}, \frac{2}{1}}_{A_3}, \underbrace{\frac{1}{3}, \frac{2}{2}, \frac{3}{1}}_{A_4}, \underbrace{\frac{1}{4}, \frac{2}{3}, \frac{3}{2}, \frac{4}{1}}_{A_5}, \underbrace{\frac{1}{5}, \frac{2}{4}, \frac{3}{3}, \frac{4}{2}, \frac{5}{1}}_{A_6}, \underbrace{\frac{1}{6}, \frac{2}{5}, \frac{3}{4}, \frac{4}{3}, \frac{5}{2}, \frac{6}{1}}_{A_7}, \ldots$$

Step 3 – Removing Repeated Terms. Note that this sequence contains many repeated elements. For instance, $\frac{1}{2}$, from A_3, is repeated as $\frac{2}{4}$ in A_6, and $\frac{1}{1}$ (which is just the number 1), is repeated many times, as $\frac{2}{2}, \frac{3}{3}, \frac{4}{4}$, etc. We remove, from our sequence, any rational number that has already appeared.

$$\frac{1}{1}, \frac{1}{2}, \frac{2}{1}, \frac{1}{3}, \cancel{\frac{2}{2}}, \frac{3}{1}, \frac{1}{4}, \frac{2}{3}, \frac{3}{2}, \frac{4}{1}, \frac{1}{5}, \cancel{\frac{2}{4}}, \cancel{\frac{3}{3}}, \cancel{\frac{4}{2}}, \frac{5}{1}, \frac{1}{6}, \frac{2}{5}, \frac{3}{4}, \frac{4}{3}, \frac{5}{2}, \frac{6}{1}, \frac{1}{7}, \cancel{\frac{2}{6}}, \frac{3}{5}, \cancel{\frac{4}{4}}, \frac{5}{3}, \cancel{\frac{6}{2}}, \frac{7}{1} \ldots$$

We denote the elements of the resulting sequence by $a_1, a_2, a_3, \ldots$ (for instance, $a_1 = \frac{1}{1}, a_4 = \frac{1}{3}, a_6 = \frac{1}{4}, a_{11} = \frac{5}{1}$, etc.

This sequence has **two** important features:

- There are no repeated numbers.

- Every **positive** rational number appears as an element of the sequence. For instance, $\frac{17}{35}$ appears in A_{52} (as $17 + 35 = 52$). In general, the rational number $\frac{a}{b}$, with $a, b \in \mathbb{N}$, appears in A_{a+b}.

Step 4 - Adding the Remaining Rational Numbers. We now enlarge the sequence (a_n) to include all rational numbers. We do so by inserting 0 as its first element, and negative rational numbers after each positive element, as follows.

$$0, a_1, -a_1, a_2, -a_2, a_3, -a_3, a_4, -a_4, a_5, -a_5, a_6, -a_6, \ldots$$

Or, more explicitly,

$$0, \frac{1}{1}, -\frac{1}{1}, \frac{1}{2}, -\frac{1}{2}, \frac{2}{1}, -\frac{2}{1}, \frac{1}{3}, -\frac{1}{3}, \frac{3}{1}, -\frac{3}{1}, \ldots.$$

This modified sequence contains **all the rational numbers** (positive, negative and zero). We can now finally construct the desired bijection.

Step 5 - Forming the Bijection. As we now have a sequence, containing **all the rational numbers**, and **without any repetitions**, we can define a function $f\colon \mathbb{N} \to \mathbb{Q}$ according to the following diagram.

$\mathbb{N}$	1	2	3	4	5	6	7	8	9	$\cdots$
	$\downarrow$	$\downarrow$	$\downarrow$	$\downarrow$	$\downarrow$	$\downarrow$	$\downarrow$	$\downarrow$	$\downarrow$	
$\mathbb{Q}$	0	a_1	$-a_1$	a_2	$-a_2$	a_3	$-a_3$	a_4	$-a_4$	$\cdots$

The function f sends 1 to 0, the even natural numbers $2, 4, 6, \ldots$ to the positive rational numbers $a_1, a_2, a_3, \ldots$, and the odd natural numbers $3, 5, 7, \ldots$ to the negative rational numbers $-a_1, -a_2, -a_3, \ldots$. All the rational numbers are covered, which implies that f is surjective. f is also injective, as our sequence $a_1, a_2, a_3, \ldots$ from Step 3 has no repetition.

In conclusion, we have constructed a bijection from $\mathbb{N}$ to $\mathbb{Q}$, and so these two sets have the same cardinality.

□

The approach used in the proof of the theorem can be generalized. Whenever elements of a set can be arranged in an infinite sequence (without repetitions), it has the same cardinality as $\mathbb{N}$. Such sets are said to be **countable**.

Definition 5.4.2. A set, that has the same cardinality as $\mathbb{N}$, is called **a countable set**.
An infinite set that is **not** countable, is called **an uncountable set**.

Examples. The following sets are countable.

- $\mathbb{N}$ itself (why?).
- The rational numbers $\mathbb{Q}$ (this is Theorem 5.4.1).
- The negative integers (see Example (d) from page 118).
- The even positive integers (see Example (e) from page 118).
- The set of all integers $\mathbb{Z}$ (as we can arrange them in an infinite sequence: $0, 1, -1, 2, -2, 3, -3, \dots$).
- $\mathbb{N} \times \mathbb{N}$ (see Exercise 5.6.33).

Do there exist infinite sets which are **not** countable? In general, is it possible for two infinite sets **not to have the same cardinality**? Are some infinite sets "larger" than others?

The following theorem answers these questions.

Theorem 5.4.3. *The set of real numbers $\mathbb{R}$ is **uncountable**.*

In other words, there is no way to arrange all the real numbers in one infinite sequence. How would one prove that "there is no way of doing something"? One common way is to use the method of **proof by contradiction**. Namely, we assume that there is a bijection between $\mathbb{N}$ and $\mathbb{R}$, and show that this assumption leads to a contradiction. Again, the proof of the theorem is far from being straightforward, and involves an extremely clever construction, known as **Cantor's Diagonalization Argument**.

Proof. In Example (g) (page 120), we proved that $\mathbb{R}$ and the interval $(0, 1)$ have the same cardinality. Hence, it is enough to prove that $(0, 1)$ is uncountable.

Assume, by contradiction, that $(0, 1)$ is countable, and suppose that $f\colon \mathbb{N} \to (0, 1)$ is a bijection. Then, for each $n \in \mathbb{N}$, $f(n)$ is some real number between 0 and 1, whose decimal expansion must start as $0.\_\_\_\dots$. If the decimal expansion happens to be finite (such as 0.25), zeros can be added to make it infinite ($0.250000\dots$).

We can therefore picture the sequence $f(1), f(2), f(3), \ldots$ as an infinite array of digits.[9]

$$\begin{aligned}
f(1) &= 0.____________________ \ldots \\
f(2) &= 0.____________________ \ldots \\
f(3) &= 0.____________________ \ldots \\
f(4) &= 0.____________________ \ldots \\
f(5) &= 0.____________________ \ldots \\
f(6) &= 0.____________________ \ldots \\
&\vdots
\end{aligned}$$

To prove that f is not a surjection, we now create a real number, in the interval $(0, 1)$, that is not the image of any $n \in \mathbb{N}$ under f.

Define a number $x \in (0, 1)$, with decimal expansion $0.a_1a_2a_3a_4a_5 \ldots$, as follows:[10]

$$a_k = \begin{cases} 5 & \text{if the } k\text{-th digit of } f(k) \text{ is } 4 \\ 4 & \text{otherwise} \end{cases}.$$

For instance, if the first few images of f are

$$\begin{aligned}
f(1) &= 0.\mathbf{3}9283434832 \ldots \\
f(2) &= 0.14\mathbf{4}563203947 \ldots \\
f(3) &= 0.595\mathbf{5}52551920 \ldots \\
f(4) &= 0.12049\mathbf{4}271927 \ldots \\
f(5) &= 0.10522\mathbf{2}961203 \ldots \\
f(6) &= 0.810240\mathbf{1}4459 \ldots \\
&\vdots
\end{aligned}$$

then the number x begins as $0.454544\ldots$.

In other words, we look at the digits on the **diagonal** of the above array, and pick the k-th digit of x to be either 5 (if the corresponding digit on the diagonal is 4) or 4 (in all other cases).

The number x is defined in such a way that its k-th digit is **different** than the k-th digit of $f(k)$, and therefore $x \neq f(k)$ for all $k \in \mathbb{N}$. Consequently, x is a real number in $(0, 1)$ that does not appear in the list

[9] There is a little technical issue here. Some numbers have two different decimal representations, such as $0.5000\ldots = 0.4999\ldots$, and so we have to make a choice. We follow the convention of choosing the one ending with zeros (rather than the one ending with nines).

[10] There is nothing special about the digits 4 and 5. Any two digits will do.

$f(1), f(2), f(3), \ldots$, and so f is **not surjective**. This contradicts the fact that f is a bijection, and hence $\mathbb{R}$ must be uncountable. This concludes the proof of the theorem. □

Our next theorem is a general cardinality statement. First, we present a definition.

Definition 5.4.4. Given a set X, we define its **power set**, $P(X)$, to be the set of all subsets of X:

$$P(X) = \{A \colon A \subseteq X\} \ .$$

For instance, if $X = \{a, b\}$, then $P(X) = \{\phi, \{a\}, \{b\}, \{a, b\}\}$. In Chapter 4, we proved that a set with n elements has 2^n subsets (see Claim 4.1.1 on page 88). However, the notion of power sets applies to infinite sets as well. For example, the power set of the natural numbers, $P(\mathbb{N})$, is the collection of all subsets of $\mathbb{N}$. Sets such as $\{1, 10, 22, 114\}$, $\{2, 4, 6, 8, \ldots\}$ and $\{1, 10, 100, 1000, 10000, \ldots\}$ are **elements** of $P(\mathbb{N})$, while the sets $\{-1, 0, 1, 2, 3\}$ and $\{1, \frac{1}{2}, \frac{1}{3}, \frac{1}{4}, \ldots\}$ are not. We write:

$$\{1, 10, 22, 114\} \in P(\mathbb{N})$$
$$\{2, 4, 6, 8, \ldots\} \in P(\mathbb{N})$$
$$\{1, 10, 100, 1000, 10000, \ldots\} \in P(\mathbb{N})$$
$$\{-1, 0, 1, 2, 3\} \notin P(\mathbb{N})$$
$$\{1, \frac{1}{2}, \frac{1}{3}, \frac{1}{4}, \ldots\} \notin P(\mathbb{N})$$

We can now state Cantor's famous theorem on power sets.

Theorem 5.4.5 (Cantor's Theorem)**.** *Let X be any set. Then the sets X and $P(X)$* ***do not*** *have the same cardinality.*

This theorem is quite easy to prove if we assume that X is finite. If X has n elements (where n is either 0 or a natural number), then, according to Claim 4.1.1, $P(X)$ has 2^n elements. As $n < 2^n$ for any $n \in \mathbb{N} \cup \{0\}$ (which can be shown by induction), the theorem is confirmed. This strategy, however, is not applicable to infinite sets. The proof, done by contradiction, uses a clever construction and pure logic.

Proof. We assume, by contradiction, that X and $P(X)$ do have the same cardinality, and so there is a bijection $f \colon X \to P(X)$. We define the following subset of X:

$$D = \{a \in X \colon a \notin f(a)\} \subseteq X$$

(D is the set of all elements a of X, which are **not** elements of their image $f(a)$). As D is a subset of X, it is **one of the elements** of the power set $P(X)$ (i.e., $D \in P(X)$). However, as f is a bijection, D must be the image of some element in X:

$$D = f(y) \qquad \text{for some } y \in X.$$

Now we are getting close to our contradiction. As y is an element of X, and D is a subset of X, either y is or is not an element of D. We take a close look at each of these options.

Case 1: $y \in D$**.**

According to the definition of the set D, if $y \in D$, then $y \notin f(y)$. But this implies that $y \notin D$ (as $f(y) = D$), which is impossible.

Case 2: $y \notin D$**.**

If $y \notin D$, then $y \notin f(y)$ (again, as $f(y) = D$), which means that y satisfies the requirement $a \notin f(a)$ for being in the set D. We thus conclude that $y \in D$, which is also impossible.

We therefore have a contradiction, as one of the two cases above must occur. We have no choice but to abandon our initial assumption, that X and $P(X)$ have the same cardinality, which concludes the proof of the theorem. □

Informally, the power set of a given set X has "more elements" or a "larger cardinality" than X (a notion to be made precise in the next section). Cantor's theorem implies that for every set, no matter how large, there is an even larger set (namely, its power set). This means, for instance, that in the following sequence, no two sets have the same cardinality.

$$\mathbb{N}\,,\ P(\mathbb{N})\,,\ P(P(\mathbb{N}))\,,\ P(P(P(\mathbb{N})))\,,\ \ldots.$$

Roughly speaking, we may say that there is "no largest infinity" (in the sense of cardinality).

5.5 More Cardinality and The Schröder-Bernstein Theorem

In Section 5.3 we introduced the notion of "having the same cardinality". However, we did not define what it means for one set to have a **larger** (or **smaller**) **cardinality** than another set (a notion that does exist, and is widely used, for finite sets). The following definition is a natural extension of Definition 5.3.1.

Definition 5.5.1. Let A and B be two sets. We say that...

(a) A and B have the **same cardinality**, and write $|A| = |B|$, if there is a **bijection** from A to B.

(b) A has cardinality **less than or equal to** the cardinality of B, and write $|A| \leq |B|$, if there is an **injection** from A to B.

(c) A has cardinality **greater than or equal to** the cardinality of B, and write $|A| \geq |B|$, if there is an **injection** from B to A.

We assign natural meaning to statements such as $|A| < |B|$, $|A| \neq |B|$, etc. (for instance, $|A| < |B|$ means that A has cardinality less than, but not the same as, the cardinality of B).

Examples.

- The set $\{-4, -2, 0, 2, 4\}$ has cardinality less than $\{1, 2, 3, 4, 5, 6, 7, 8, 9, 10\}$. We write

  $$|\{-4, -2, 0, 2, 4\}| \leq |\{1, 2, 3, 4, 5, 6, 7, 8, 9, 10\}| \quad \text{or} \quad |\{-4, -2, 0, 2, 4\}| < |\{1, 2, 3, 4, 5, 6, 7, 8, 9, 10\}|.$$

- $\mathbb{N}$ and $\mathbb{Q}$ are both countable sets (Theorem 5.4.1), so we write $|\mathbb{N}| = |\mathbb{Q}|$. Similarly, we have $|(0, 1)| = |\mathbb{R}|$.

- The function $f\colon \mathbb{Z} \to \mathbb{R}$, that sends every integer to itself, is an injection, and so $|\mathbb{Z}| \leq |\mathbb{R}|$. However, $\mathbb{Z}$ is countable, while $\mathbb{R}$ is not (Theorem 5.4.3), and thus $|\mathbb{Z}| < |\mathbb{R}|$.

- For any set X, the function from X to $P(X)$, sending $a \in X$ to $\{a\} \in P(X)$, is injective. Moreover, X and $P(X)$ never have the same cardinality (Cantor's Theorem). We can summarize these facts by saying that for any set X, $|X| < |P(X)|$.

A word of caution is in place. We have been using symbols such as $\leq$, $=$ and $<$ ever since elementary school, to compare real numbers, and order them (for example, $5 < 19$, $-5 > -7.5$, $\frac{1}{2} = \frac{5}{10}$). Basic properties of these relations became natural, and we use them freely without having any doubts. For instance, if $a < b$ and $b < c$, then $a < c$ (for any $a, b, c \in \mathbb{R}$). Now, that we have extended the use of these symbols to comparing cardinalities of (both finite and infinite) sets, we need to review these "obvious" features of $<$, $\leq$ and $=$, and check whether they still apply in the context of cardinality. We should ask ourselves:

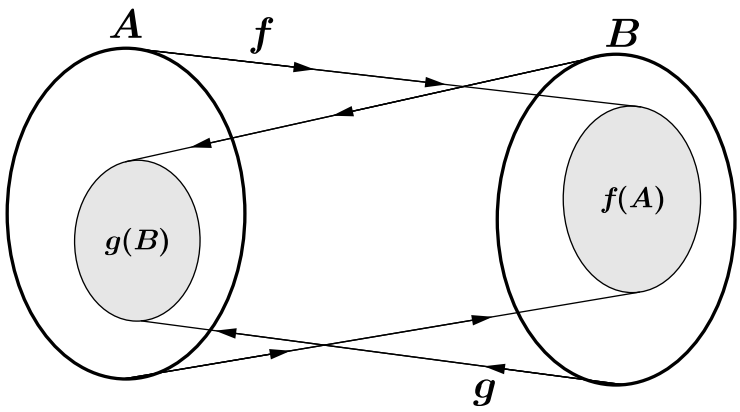
(1) If $|A| \leq |B|$ and $|B| \leq |C|$ (for sets A, B, C), is it necessarily true that $|A| \leq |C|$?

(2) If A and B are any two sets, is it necessarily true that either $|A| \leq |B|$ or $|A| \geq |B|$ (i.e., given two sets, is there always an injection going from one of them to the other?).

(3) If $|A| \leq |B|$ and $|A| \geq |B|$, can we conclude that $|A| = |B|$?

Fortunately, the answer to all the three questions is affirmative. However, the proofs can be nontrivial. Question (1) is left as a (relatively easy) exercise (see Exercise 5.6.50). To fully answer Question (2), a more advanced set-theoretic tool is needed, which is beyond the scope of these notes. Question (3) is answered by the famous Schröder-Bernstein Theorem.

Theorem 5.5.2 (Schröder-Bernstein Theorem)**.**
Let A and B be two sets. If $|A| \leq |B|$ and $|A| \geq |B|$, then $|A| = |B|$.
In other words, if there are injections $f\colon A \to B$ and $g\colon B \to A$, then there is a bijection $h\colon A \to B$.



The proof of this theorem is outlined in Exercise 5.6.52, and is quite technical. The main issue here is that f and g are **known to be injective**, but **neither of them is necessarily a surjection**. A bijection from A to B needs to be constructed from f and g, and doing so requires some effort. Nevertheless, this theorem provides a powerful tool for proving that two sets have the same cardinality, as illustrated in the following example.

Example. $|[0,1]| = |(0,1)|$.
Namely, the intervals $[0,1]$ and $(0,1)$ have the same cardinality.

Proof. Define functions $f\colon (0,1) \to [0,1]$ and $g\colon [0,1] \to (0,1)$ as follows:

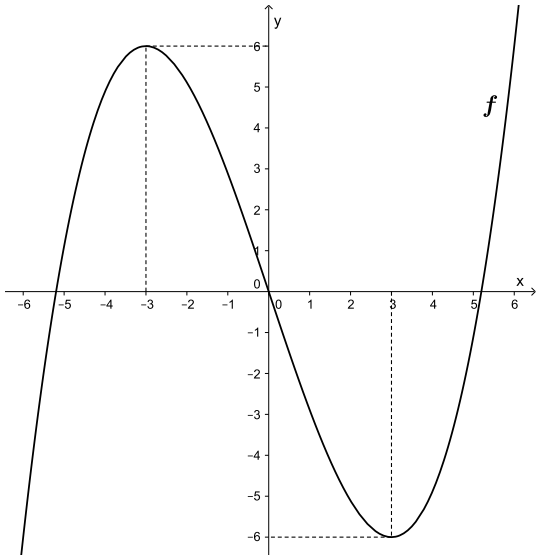
$$f(x) = x \quad \text{for} \quad 0 < x < 1 \quad , \qquad \text{and} \qquad g(x) = \frac{x}{2} + \frac{1}{4} \quad \text{for} \quad 0 \leq x \leq 1 \ .$$

The functions f and g are both injective (why?). f is simply the inclusion of the open interval $(0,1)$ into the closed interval $[0,1]$. The function g shrinks the interval $[0,1]$ by a factor of 2, and translates it a quarter unit to the right (i.e., g sends the interval $[0,1]$ to $[0.25, 0.75]$). Since we have exhibited the two required injections, according to the Schröder-Bernstein's Theorem, $(0,1)$ and $[0,1]$ have the same cardinality.

□

5.6 Exercises for Chapter 5

5.6.1. Here is (part of) the graph of a function $f\colon \mathbb{R} \to \mathbb{R}$.



The domain and the codomain of f can be restricted to various intervals. In each case, decide whether the given restriction is **an injection, a surjection, a bijection, or neither**. Explain your answer briefly.

(a) $f\colon [-3,3] \to [-6,6]$

(b) $f\colon [-4,0] \to [0,6]$

(c) $f\colon [1,4] \to [-7,2]$

(d) $f\colon [-2,2] \to [-7,7]$

(e) $f\colon [-4,-1] \to [-2,7]$

(f) $f\colon [0,4] \to [-6,0]$

5.6.2. For each of the following functions, decide whether it is **an injection, a surjection, a bijection, or neither**. Justify your answer.

(a) $p\colon \mathbb{N}\times\mathbb{N}\to\mathbb{N}$, $p(a,b)=\dfrac{ab(b+1)}{2}$.

(b) $f\colon [0,\infty)\to\mathbb{R}$, $f(x)=\sqrt{x}$.

(c) $g\colon \mathbb{R}^2\to\mathbb{R}$, $g(x,y)=|x+y|$.

(d) $h\colon \mathbb{R}\to\mathbb{R}$, $h(x)=x^3$.

(e) $r\colon \mathbb{R}\to\mathbb{R}$, $r(x)=\dfrac{x}{1+x^2}$.

(f) $q\colon \mathbb{Z}\to\mathbb{Z}$, $q(n)=n+1$.

(g) $t\colon \mathbb{N}\times\mathbb{N}\to\mathbb{N}$, $t(a,b)=a\cdot b$.

5.6.3. For each of the following functions, decide whether it is **an injection, a surjection, a bijection, or neither**. Justify your answer.

(a) $p\colon \mathbb{N}\times\mathbb{N}\to\mathbb{N}$, $p(a,b)=\dfrac{(a+1)b(b+1)}{2}$.

(b) $f\colon \mathbb{R}\to\mathbb{R}$, $f(x)=2x+1$.

(c) $g\colon \mathbb{R}^2\to\mathbb{R}$, $g(x,y)=x+y$.

(d) $h\colon \mathbb{R}\to\mathbb{R}$, $h(x)=\dfrac{x^2}{1+x^2}$.

(e) $r\colon \mathbb{R}\to[0,\infty)$, $r(x)=|x|$.

(f) $q\colon \mathbb{N}\to\mathbb{N}$, $q(n)=n+1$.

(g) $t\colon \mathbb{N}\times\mathbb{N}\to\mathbb{N}$, $t(a,b)=a+b$.

5.6.4. Let $f\colon \mathbb{R}\to\mathbb{R}$ given by $f(x)=\begin{cases} x & \text{if } x\le 0 \\ \dfrac{1}{x} & \text{if } x>0 \end{cases}$.

Is f an **injection**? Is it a **surjection**? Explain.

5.6.5. Which function is **one-to-one** and **NOT onto**? Explain.

- $f\colon \mathbb{Z}\to\mathbb{Z}$, $f(x)=x^3$
- $g\colon \mathbb{R}\to\mathbb{R}$, $g(x)=x^3$
- $h\colon \mathbb{R}\to\mathbb{R}$, $h(x)=x^2$
- $r\colon \mathbb{R}\times\mathbb{R}\to\mathbb{R}$, $r(x,y)=x^2+y$

5.6.6. Which function is an **injection**? Explain.

- $f\colon \mathbb{R}\to\mathbb{R}$, $f(x)=x^2$
- $g\colon \mathbb{Q}\to\mathbb{Q}$, $g(x)=x^2$
- $h\colon \mathbb{Z}\to\mathbb{Z}$, $h(x)=x^2$
- $p\colon \mathbb{N}\to\mathbb{N}$, $p(x)=x^2$

5.6.7. Which function is a **bijection** ? Explain.

- $f\colon \mathbb{R} \to \mathbb{R}$, $f(x) = x^6$
- $g\colon [0,\infty) \to [0,\infty)$, $g(x) = x^6$
- $h\colon \mathbb{Z} \to \mathbb{Z}$, $h(x) = x^6$
- $p\colon \mathbb{Q} \to [0,\infty)$, $p(x) = x^6$

5.6.8. Let $f\colon \mathbb{R} \to \mathbb{R}$ be an arbitrary function.

Each of the following statements, written using the logic symbols, describes a property of f (e.g., injectivity, monotonicity, boundedness, etc.). Identify the property described by each statement.

(a) $(\forall y \in \mathbb{R})(\exists x \in \mathbb{R})(f(x) = y)$

(b) $(\exists M \in \mathbb{R})(\forall x \in \mathbb{R})(|f(x)| \leq M)$

(c) $(\forall x_1 \in \mathbb{R})(\forall x_2 \in \mathbb{R})[(x_1 \neq x_2) \Rightarrow (f(x_1) \neq f(x_2))]$

5.6.9. Prove that the function $f\colon \mathbb{R} \to \mathbb{R}$, $f(x) = x \cdot |x|$ is a bijection.

5.6.10. Consider the function $f\colon \mathbb{N} \to \mathbb{N}$, $f(n) = n + (-1)^{n+1}$.

(a) Compute $f(1), f(2), f(3)$ and $f(4)$.

(b) Is f an **injection**?

(c) Is f a **surjection**?

5.6.11. Consider the sets

$$A = \{1, 2, 3, \ldots, 365\} \quad , \quad B = \{\text{Jan, Feb, } \ldots, \text{ Dec}\} \quad , \quad C = \{1, 2, 3, \ldots, 31\}$$

and the function $f : A \to B \times C$, defined by

$$f(x) = (\text{month for day } x, \text{ day of the month for day } x),$$

in a regular year (not leap year).

For example: $f(365) = (\text{Dec}, 31)$, since the date of the last day of the year is December 31st.

(a) What are $f(1)$, $f(32)$, and $f(359)$?

(b) Is f **injective**? Explain.

(c) Is f **surjective**? Explain.

5.6.12. Let $f\colon [0, 2] \to [5, 6]$ be a function.

(a) If the restrictions of f to $[0,1]$ and to $[1,2]$ are injective functions, must f be an injection? Explain.

(b) If the restrictions of f to $[0,1]$ and to $[1,2]$ are surjective functions, must f be a surjection? Explain.

5.6.13. Let $f : \mathbb{R} \to \mathbb{R}$ be a function, such that $\big|f(x) - f(y)\big| \geq 5\big|x - y\big|$ for all $x, y \in \mathbb{R}$. Show that f is injective.

5.6.14. (Harder!) Show that the function $f\colon \mathbb{N} \to \mathbb{R}$, $f(n) = \sin n$ is one-to-one.
(Hint: You may use, without proof, the fact that π is an irrational number.)

5.6.15. Let $f\colon A \to B$ be an arbitrary function.

(a) Prove that if f is a bijection (and hence invertible), then $f^{-1}(f(x)) = x$ for all $x \in A$, and $f(f^{-1}(x)) = x$ for all $x \in B$.

(b) Conversely, show that if there is a function $g\colon B \to A$, satisfying $g(f(x)) = x$ for all $x \in A$, and $f(g(x)) = x$ for all $x \in B$, then f is a bijection, and $f^{-1} = g$.

5.6.16. (a) Is it true that $\sin^{-1}(\sin x) = x$ for all $x \in \mathbb{R}$? Explain.

(b) The **inverse tangent function** is defined as the inverse of the function

$$f\colon \left(-\frac{\pi}{2}, \frac{\pi}{2}\right) \to \mathbb{R} \qquad , \qquad f(x) = \tan x.$$

We write $f^{-1}(x) = \tan^{-1} x$.
Is it true that $\tan(\tan^{-1} x) = x$ for all $x \in \mathbb{R}$? Explain.

(c) What is the inverse function of $g\colon (-\infty, 0] \to [0, \infty)$, $g(x) = x^2$? Explain.

5.6.17. Consider the functions $\begin{cases} f\colon \mathbb{N} \times \mathbb{N} \to \mathbb{R} \\ f(m,n) = \dfrac{m+1}{n+1} \end{cases}$ and $\begin{cases} g\colon \mathbb{R} \to \mathbb{R} \\ g(x) = \sqrt{x^2+5} \end{cases}$.

Compute $g \circ f(9,4)$ and $f(g(2), g(\sqrt{20}))$.

5.6.18. Consider the following functions: $\begin{cases} f\colon \mathbb{Z} \to \mathbb{Z} \times \mathbb{Z} \\ f(m) = (2m, m-1) \end{cases}$ $\begin{cases} g\colon \mathbb{Z} \times \mathbb{Z} \to \mathbb{Z} \\ g(m,n) = |m \cdot n| \end{cases}$.

(a) Is f **injective** ? Explain.

(b) Is g **surjective**? Explain.

(c) State the **domain** and the **codomain** of $g \circ f$, and write a **formula** for this composition.

(d) State the **domain** and the **codomain** of $f \circ g$, and write a **formula** for this composition.

5.6.19. Let $g\colon \mathbb{Z} \to \mathbb{Z}$ be given by $g(x) = 2x + 1$.

(a) Does g have an **inverse**? Explain.

(b) **Find an expression** for the function $g \circ g \circ g$. Simplify your answer.

5.6.20. Let $f\colon A \to \mathbb{R}$ be a function (where A is a set). Define a new function $g\colon A \to \mathbb{R}$ by $g(x) = 3 \cdot [f(x)]^2 + 1$. Prove that if g is injective, then f is injective.

5.6.21. Consider the function $f\colon [0,\infty) \to [0,\infty) \quad , \quad f(x) = \dfrac{x}{x+1}$.

Prove, by induction, that for any $n \in \mathbb{N}$, $f^n(x) = \dfrac{x}{1 + n \cdot x}$.

Note: f^n is the function obtained by composing n copies of f: $f^n = \underbrace{f \circ f \circ \cdots \circ f}_{n \text{ times}}$.

5.6.22. Consider the function $g\colon \mathbb{R} \to \mathbb{R} \quad , \quad g(x) = 2x + 1$.

Prove, by induction, that for any $n \in \mathbb{N}$, $g^n(x) = 2^n \cdot x + 2^n - 1$.

Note: g^n is the function obtained by composing n copies of g: $g^n = \underbrace{g \circ g \circ \cdots \circ g}_{n \text{ times}}$.

5.6.23. For each of the following statements, decide whether it is **true** or **false**. Justify your answer with a proof or a counterexample.

(a) Any surjective function $f\colon \mathbb{R} \to \mathbb{R}$ is unbounded (i.e., not bounded).

(b) Every unbounded function $f\colon \mathbb{R} \to \mathbb{R}$ is surjective.

(c) Any injective function $f\colon \mathbb{R} \to \mathbb{R}$ is strictly monotone.

(d) The composition of two strictly monotone functions $f, g\colon \mathbb{R} \to \mathbb{R}$ is also strictly monotone.

5.6.24. Is the following statement necessarily true? Provide a proof or a counterexample.

"If $h\colon A \to B$, $g\colon B \to C$ and $f\colon B \to C$ are three functions, and $g \circ h = f \circ h$, then $g = f$."

5.6.25. Let $f\colon A \to B$ be a bijection, where A and B are subsets of $\mathbb{R}$.

Prove that if f is strictly increasing (respectively decreasing) on A, then f^{-1} is strictly increasing (respectively decreasing) on B.

5.6.26. Prove part (a) of Proposition 5.2.2.

5.6.27. Find examples of functions f, g from $\mathbb{R}$ to $\mathbb{R}$, satisfying the following conditions, or prove that such examples do not exist.

(a) $g \circ f$ is injective, but g is not injective.

(b) $g \circ f$ is surjective, but g is not surjective.

(c) $g \circ f$ is surjective, but f is not surjective.

(d) f and g are not injective, but $g \circ f$ is injective.

5.6.28. (**Harder!**) Find two functions $f, g\colon \mathbb{R} \to \mathbb{R}$, such that $f \circ g$ is bijective, while $g \circ f$ is not a bijection.[11]

5.6.29. Show that the following pairs of sets have the same cardinality.

(a) Integers divisible by 3, and the **even** positive integers.

(b) $\mathbb{R}$, and the interval $(0, \infty)$.

(c) The interval $[0, 2)$, and the set $[5, 6) \cup [7, 8)$.

(d) The intervals $(-\infty, -1)$ and $(-1, 0)$.

5.6.30. Understanding the Proof of Theorem 5.4.1.

Read the proof of Theorem 5.4.1, and answer the following questions.

(a) Write down explicitly the sets A_1, A_5 and A_{10}, defined in Step 1.

(b) Which elements of A_{10} appeared in one of the previous A_k's?

(c) Find three A_k's that contain the number 0.28.

(d) Why was it necessary to remove, in Step 3, repeated terms from our initial sequence?

(e) Referring to the sequence (a_n) from Step 3, what are a_{19}, a_{22} and a_{27}?

(f) The bijection $f\colon \mathbb{N} \to \mathbb{Q}$, from Step 5, can be described using algebraic expressions (instead of a diagram). Complete the following alternate definition of f:

$$f(n) = \begin{cases} \_\_\_\_\_ & \text{if} \quad n = 1 \\ \_\_\_\_\_ & \text{if } n \text{ is even} \\ \_\_\_\_\_ & \text{if } n \text{ is odd and greater than 1} \end{cases}$$

[11] There are no such functions if we assume, in addition, that f and g are continuous.

5.6.31. An Alternate Proof of Theorem 5.4.1 The heart of the proof of Theorem 5.4.1 is to arrange all the rational numbers in one infinite sequence. However, there is more than one way to do so. The following pattern can be used to arrange all the rational numbers **in the open interval** $(0,1)$ in an infinite sequence (with repetitions).

$$\frac{1}{2}, \frac{1}{3}, \frac{2}{3}, \frac{1}{4}, \frac{2}{4}, \frac{3}{4}, \frac{1}{5}, \frac{2}{5}, \frac{3}{5}, \frac{4}{5}, \frac{1}{6}, \dots$$

We can then add reciprocals and negatives to include all rational numbers. Use these ideas to construct an alternate proof of Theorem 5.4.1.

5.6.32. (a) Let A and B be disjoint sets, which are both countable. Prove that $A \cup B$ is also countable.

(b) Use part (a) to show that the set of all **irrational** real numbers is uncountable.

5.6.33. Prove that $\mathbb{N} \times \mathbb{N}$ (the set of all pairs of natural numbers) is a countable set.
To do so, define, for each $k \in \mathbb{N}$, the set

$$A_k = \{(m,n) \colon m,n \in \mathbb{N} \text{ and } m+n=k\},$$

and follow the outline of the proof of Theorem 5.4.1.

5.6.34. (a) Find an example of set A and a **nonempty** subset $B \subseteq A$, such that A and $A \setminus B$ have the same cardinality.

(b) Explain why if A is a **finite** set, and $B \subseteq A$ is a subset for which A and $A \setminus B$ have the same cardinality, then $B = \phi$.

5.6.35. Prove that the set of all **finite** subsets of $\mathbb{N}$ is a countable set.
Hint: For each $k \in \mathbb{N} \cup \{0\}$, define A_k to be the collection of all finite subsets of $\mathbb{N}$ whose elements add up to k. For example:

$$A_5 = \{\{1,4\},\{2,3\},\{5\}\} \ , \ A_6 = \{\{1,2,3\},\{1,5\},\{2,4\},\{6\}\} \ .$$

5.6.36. Let A be the set of all infinite sequences consisting of 0's and 1's (i.e., sequences such as $010101010\dots$, $1010010001000\dots$, etc.). Prove that A is **uncountable**.
Hint: Assume that A is countable (i.e., its elements can be arranged in a list), and construct a sequence of zeros and ones which is not on that list. Use Cantor's diagonalization argument.

5.6.37. For each set, decide whether it is finite, countable, or uncountable. Explain your answer briefly.

$$P(\mathbb{N}) \quad , \quad \left\{1, \frac{1}{2}, \frac{1}{3}, \frac{1}{4}, \dots\right\} \cap [0.03, 1] \quad , \quad \mathbb{Q} \quad , \quad \mathbb{Z} \quad , \quad (0,\infty)$$

5.6.38. For each set, decide whether it is finite, countable, or uncountable. Explain your answer briefly.

$$P(\mathbb{Z}) \quad , \quad (2,3) \quad , \quad \text{all the prime numbers} \quad , \quad \mathbb{Q} \cap [0,1] \quad , \quad \mathbb{N} \cap (-\infty, 1000)$$

5.6.39. Let $A = \left\{2, \frac{1}{2}, 3, \frac{1}{3}, 4, \frac{1}{4}, 5, \frac{1}{5}, \dots\right\}$. Which set is **infinite** and **uncountable**? Explain.

- A
- $A \cap \mathbb{N}$
- $P(A \cap [0,1])$
- $P(A \cap [1,2])$

5.6.40. Let A and B be two **infinite** sets that **do not** have the same cardinality.

(a) If A is countable, must B be uncountable? Explain.

(b) If A is uncountable, must B be countable? Explain.

5.6.41. Let $A = \{4,5,6,7\}$. Fill in the blanks with either $\in$ or $\subseteq$.

$\{4\}$ ____ A $\quad$ $\{5,6\}$ ____ $P(A)$ $\quad$ $\{\phi\}$ ____ $P(A)$ $\quad$ 5 ____ A

A ____ $P(A)$ $\quad$ $\{A, \phi\}$ ____ $P(A)$

5.6.42. Fill in the blanks with either $\in$ or $\subseteq$.

ϕ ____ $\mathbb{Z}$ $\quad$ $\mathbb{N}$ ____ $P(\mathbb{Z})$ $\quad$ $P(\mathbb{N})$ ____ $P(\mathbb{Q})$

$\{\sqrt{2}\,,\ 4.5\}$ ____ $\mathbb{R}$ $\quad$ $\frac{3}{4}$ ____ $\mathbb{Q}$ $\quad$ $\mathbb{R}$ ____ $P(\mathbb{R})$

5.6.43. Let $A = \{1\,, 2\,, \{1\,, 2\}\}$.

(a) Fill in the blanks with either $\in$ or $\subseteq$.

$\{2\}$ ____ A $\quad$ $\{\{1\}\}$ ____ $P(A)$ $\quad$ $\{1\,, 2\}$ ____ $P(A)$

$\{\{1\,, 2\}\}$ ____ A $\quad$ $\{1\,, \{1\,, 2\}\}$ ____ $P(A)$

(b) How many elements are in the set $P(A) \setminus A$?

5.6.44. (a) Let $X = \{0,1\}$. List all the elements in $P(P(X))$.

(b) Let $A = \{1,2\}$ and $B = \{2,3\}$. Find the sets $P(A \cap B)$ and $P(A) \cup P(B)$.

5.6.45. For each statement, decide whether it is **true** or **false**. Justify with a proof or a counterexample.

(a) For any two sets A, B we have $P(A \cap B) = P(A) \cap P(B)$.

(b) For any two sets A, B we have $P(A \cup B) = P(A) \cup P(B)$.

(c) For any two sets A, B, if $A \cap B = \phi$, then $P(A) \cap P(B) = \{\phi\}$.

5.6.46. Consider the function $f\colon P(\mathbb{Z}) \to P(\mathbb{N})$, $f(A) = A \cap \mathbb{N}$.

(a) What are $f(\{-2,-1,0,1,2\})$, $f(\{-1,-2,-3,\dots\})$ and $f(\mathbb{N})$?

(b) Is f **surjective**? Explain.

(c) Is f **injective**? Explain.

5.6.47. Let $f\colon \mathbb{N} \to P(\mathbb{N})$ be given by $f(n) = \{n+1, n+2, n+3, \dots\}$.

(a) Find the **set** $f(3) \cap [-8\,, 8]$.

(b) Is f an **injection**? Explain.

(c) Is f a **surjection**? Explain.

5.6.48. Let X be a nonempty set, and define B to be the set of all functions from X to the two-element set $\{0,1\}$:

$$B = \{f\colon X \to \{0,1\}\}.$$

Construct a **bijection** $h\colon P(X) \to B$.

5.6.49. Let A and B be two sets. Prove that if $P(A \cup B) = P(A) \cup P(B)$, then $A \subseteq B$ or $B \subseteq A$.

5.6.50. Let A and B be two sets. Prove that if $|A| \leq |B|$ and $|B| \leq |C|$, then $|A| \leq |C|$.

5.6.51. Use the Schröder-Bernstein Theorem to prove that:

(a) $|[0,1]| = |[0,1)|$

(b) $|[0,\infty)| = |(0,\infty)|$

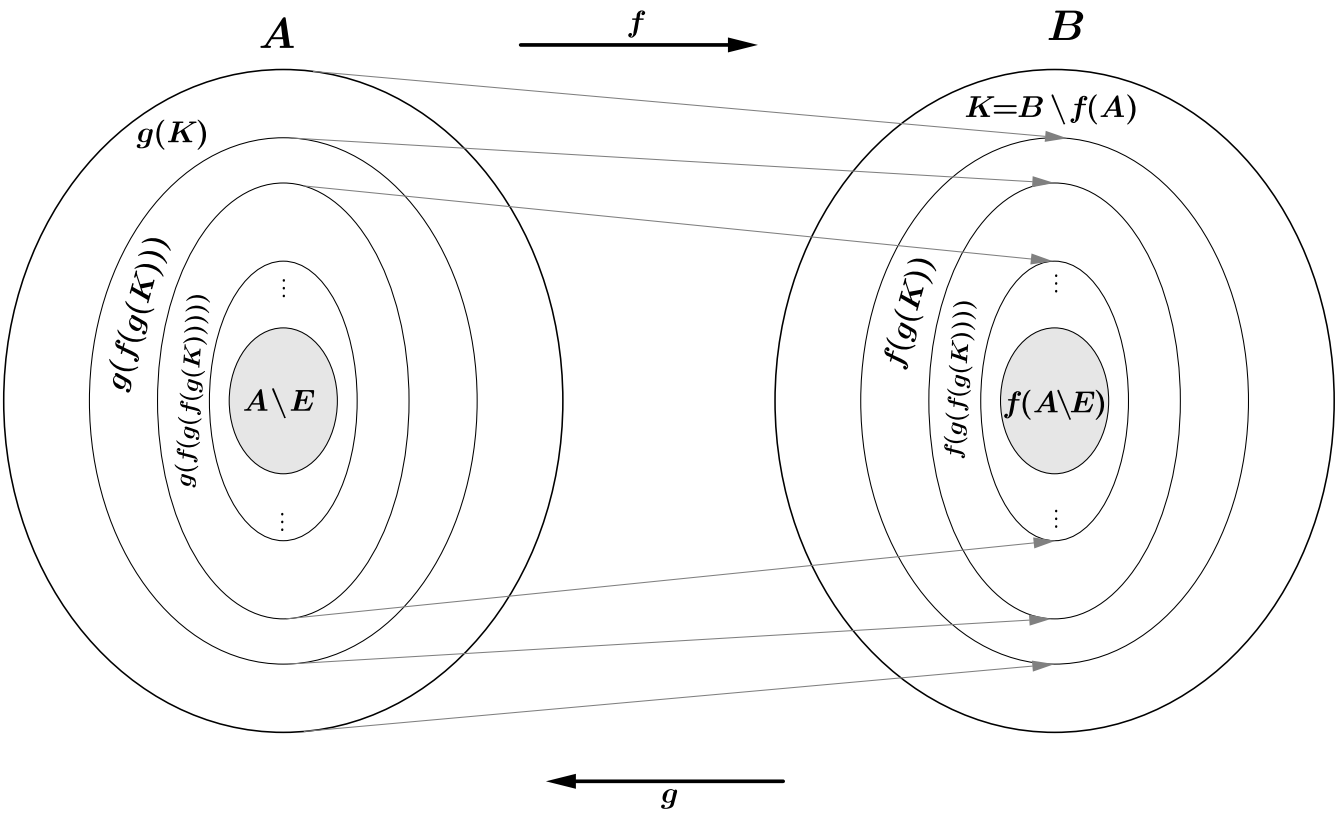
(c) $|[0,1]| = |\mathbb{R}|$

(d) $|\mathbb{R} \setminus \mathbb{Z}| = |\mathbb{R}|$

5.6.52. Proof of the Schröder-Bernstein Theorem.
Recall that the Schröder-Bernstein Theorem says that if $f\colon A \to B$ and $g\colon B \to A$ are two **injections**, then there is a **bijection** $h\colon A \to B$. To prove the theorem, we need to construct the function h from the functions f and g.

It is important to realize, that since we have no information whatsoever about the sets A, B, and the functions f, g (other than the fact that they are one-to-one), we have little choice in constructing h. The function h must send every $a \in A$ to some $b \in B$ via either f or g.

The key to understanding the proof is making sense of the following diagram.



Here is what is going on. Denote by K the set of all the elements in B which are not in the image of f (i.e., $K = B \setminus f(A)$). As h is supposed to be a surjection, it must send some elements of A to K. The function f will not be able to do that, and so g has to be used. In other words, h will send elements in $g(K)$ to B using g:

$$h(a) = g^{-1}(a) \quad \text{if} \quad a \in g(K) \ .$$

Now, that h has been defined on $g(K)$, we must take care of the rest of the elements of A, namely – $A \setminus g(K)$. In other words, we need to define a bijection from $A \setminus g(K)$ to $B \setminus K = f(A)$. But we now face the exact same problem: The set $f(A \setminus g(K))$ is a subset of $f(A)$. Elements in $f(A)$ which are not in $f(A \setminus g(K))$ cannot be reached through f, so g must be used again. That is,

$$h(a) = g^{-1}(a) \quad \text{if} \quad a \in g(f(g(K))) \ .$$

This process can be repeated indefinitely. To formalize it, we denote, for every $n = 0, 1, 2, \ldots$ the composition of n copies of $f \circ g$ by $(f \circ g)^n$:

$$(f \circ g)^n = \underbrace{(f \circ g) \circ (f \circ g) \circ \cdots \circ (f \circ g)}_{n \text{ times}}$$

(if $n = 0$, we interpret $(f \circ g)^n$ as the identity function, sending each element of B to itself). We define

$$E = \{a \in A : \ a = g \circ (f \circ g)^n(b) \ \text{ for some } \ b \in K \ \text{ and some } \ n = 0, 1, 2 \dots\}$$

(note that E is just the union of $g(K), g(f(g(K))), g(f(g(f(g(K))))), \dots$).

Finally, we define:

$$h\colon A \to B \qquad , \qquad h(a) = \begin{cases} g^{-1}(a) & \text{if } \ a \in E \\ f(a) & \text{if } \ a \notin E \end{cases}.$$

The rest of the work is left for you. Answer the following questions.

(a) How come we use the notation $g^{-1}(a)$, when g is not necessarily invertible (as a function from A to B)?

(b) The identity $f(A) \setminus f(A \setminus g(K)) = f(g(K))$ was used in the proof (where?). Prove that it is indeed a valid identity.

(c) Prove that h is one-to-one. Namely, show that if $h(a_1) = h(a_2)$ for some $a_1, a_2 \in A$, then $a_1 = a_2$.
Hint: If a_1, a_2 are both in E or both not in E, then clearly $a_1 = a_2$.

(d) Prove that h in onto.
Hint: If $b \in B$, then either $b \in f(A \setminus E)$ or $b \notin f(A \setminus E)$.

Chapter 6

Integers and Divisibility

This chapter is devoted to studying, in more depth, the set of integers $\mathbb{Z}$, its structure and properties. The integers play a fundamental role in many areas of mathematics, science, and beyond. The integers are closely related to the set of natural numbers, and thus are often used in problems involving counting, sequences, and sets with finitely many elements (such as finite fields).

You might wonder – why can we not just study the real numbers instead? After all, the integers are a subset of the real numbers, so once we understand the reals, we ought to have a solid understanding of the integers as well, right? Well, unfortunately, in mathematics, studying a certain object does not always reveal properties of its sub-objects.

Consider, for instance, the notion of **divisibility**. Given two integers a and b, with $b \neq 0$, either a is or is not divisible by b. However, it makes no sense to apply this notion to the set of real numbers: Is 72.5 divisible by $\sqrt{5}$? Yes it is (any real number is divisible by any other nonzero real number). Divisibility is a notion that becomes interesting and relevant when dealing with integers, and gives rise to other important notions (such as **an even** or **a prime** number).

6.1 Divisibility and the Division Algorithm

We begin by recalling Definitions 1.4.1 and 1.4.2 from Chapter 1, and introduce a common and useful notation.

Definition 6.1.1. Let a be an integer, and b a **nonzero** integer.

(a) We say that a is **divisible** by b (or that b **divides** a), and write $b|a$, if there exists an integer m, for which $a = m \cdot b$. We write $b \not| a$ to indicate that a is **not** divisible by b.

(b) A natural number $p > 1$ is called a **prime number**, if the only natural numbers that divide p are 1 and p.

Examples.

(a) 40 is divisible by 8, which can be written as $8|40$.

(b) 13 divides 52, which we can write as $13|52$.

(c) 72 is not a multiple of 7, and hence $7 \not| 72$.

(d) Any nonzero integer is divisible by itself and by 1. Thus, for any $k \in \mathbb{Z}$, with $k \neq 0$, we have $1|k$ and $k|k$.

(e) Similarly, for any nonzero $k \in \mathbb{Z}$, we have

$$k|5k \quad , \quad k|k^2 \quad , \quad k|0 \quad , \quad k|(k^3 - 7k) \quad , \quad \text{etc.}$$

When dividing an integer by another, say 35 by 8, the result may not be an integer. However, we know from our elementary school years, that we can perform **division with remainder**: 35 divided by 8 gives 4, with a reminder of 3. In other words, we can break 35 into four groups of 8 (and not more), after which we are left with three unused units:

$$35 = 4 \cdot 8 + 3.$$

Division with remainder allows us to perform division without leaving the world of integers (i.e., without referring to fractions or decimals). The following theorem gives a general description of this operation.

Theorem 6.1.2. *(The Division Algorithm)*
If $a, b \in \mathbb{N}$, then there is a unique pair of integers, q and r, with $q \geq 0$ and $0 \leq r < b$, such that

$$a = q \cdot b + r$$

*(q is called the **quotient**, and r the **remainder**).*

Examples.

(a) Dividing $a = 20$ by $b = 3$ gives a quotient of $q = 6$ and a remainder of $r = 2$, as $20 = 6 \cdot 3 + 2$.

(b) If $a = 47$ and $b = 10$, then the quotient is $q = 4$, and the remainder is 7, as $47 = 4 \cdot 10 + 7$.

(c) 28 is divisible by 7. In this case, $q = 4$ and $r = 0$: $28 = 4 \cdot 7 + 0$.

(d) What if $a < b$? For instance, say $a = 12$ and $b = 34$? Then we cannot squeeze even one group of 34 into 12, and so the quotient must be zero. The remainder, in this case, is equal to a:

$$12 = 0 \cdot 34 + 12.$$

In general, whenever $a < b$, the quotient q is zero, and $r = a$.

Proof. (of Theorem 6.1.2) The theorem states that **there are** numbers q, r, and that they are **unique**. Therefore, our proof has two parts: **The existence part** (showing that there exist such q and r), and **the uniqueness part** (showing that q and r are unique).

Proof of Existence.

We prove existence by **strong induction** on a. That is, we treat b, throughout the proof, as a fixed unknown integer.

For $a = 1$, we distinguish between two cases:

- If $b = 1$, we can write $1 = 1 \cdot b + 0$, and so $q = 1$ and $r = 0$ satisfy the conclusion of the theorem (dividing 1 by 1 gives a quotient of 1 and a remainder of 0).
- If $b > 1$, we write $1 = 0 \cdot b + 1$, and hence $q = 0$ and $r = 1$ are appropriate choices.

We proved the theorem for $a = 1$ (regardless of the value of b), which confirms the base case.

Now assume that the theorem holds true for $a = 1, 2, \ldots, k$, for some $k \in \mathbb{N}$, and consider $a = k + 1$. We need to show that $k + 1 = q \cdot b + r$, for some integers q, r with $q \geq 0$ and $0 \leq r < b$. Again, we proceed by cases:

- $\mathbf{k + 1 < b}$.

  In this case we attempt to divide a natural number by a large number, which means that the quotient must be zero. Indeed, we can write

  $$k + 1 = 0 \cdot b + (k + 1),$$

  proving the theorem (in this case) with $q = 0$ and $r = k + 1$.

- $\mathbf{k + 1 = b}$.

  Here our a and b are the same, and thus the quotient is 1 and the remainder is 0:

  $$k + 1 = 1 \cdot b + 0.$$

- $\mathbf{k+1 > b}$.

  In this case, we finally use our induction hypothesis. From the fact that $k+1 > b$ (and $b \in \mathbb{N}$), we conclude that $k+1-b$ is a natural number, smaller than $k+1$, and hence covered by our (strong) induction hypothesis. We have

  $$k+1-b = q \cdot b + r$$

  for some integers q, r, with $q \geq 0$ and $0 \leq r < b$. Rewriting this equality as

  $$k+1 = (q+1) \cdot b + r$$

  shows that the theorem holds for $a = k+1$ as well (with quotient $q+1$ and remainder r).

In conclusion, we proved (the existence part of) the theorem for $a = k+1$ (and any $b \in \mathbb{N}$), and hence for all a, b.

Proof of Uniqueness.

A common way to prove uniqueness in mathematics, is to assume that there are two elements satisfying the conclusion of a theorem, and then show that these two elements must be, in fact, equal to each other.

To apply this strategy in our case, we assume that there are two pairs of q and r as required by the theorem. Namely, suppose that

$$a = q_1 \cdot b + r_1 \qquad \text{and} \qquad a = q_2 \cdot b + r_2$$

for some integers q_1, q_2, r_1, r_2, with $q_1, q_2 \geq 0$ and $0 \leq r_1, r_2 < b$. We show uniqueness by proving that $q_1 = q_2$ and $r_1 = r_2$.

By equating the two right-hand sides, we get

$$q_1 \cdot b + r_1 = q_2 \cdot b + r_2 \qquad \Rightarrow \qquad (q_1 - q_2) \cdot b = r_2 - r_1.$$

Note that since both r_1 and r_2 are between 0 and b, their difference must be between $-b$ and b :

$$-b < r_2 - r_1 < b.$$

However, from $(q_1 - q_2) \cdot b = r_2 - r_1$ we see that $r_2 - r_1$ must be an integer multiple of b. Well, the only multiple of b that is strictly between $-b$ and b is 0, and thus

$$r_2 - r_1 = 0 \qquad \Rightarrow \qquad r_1 = r_2.$$

Also, $(q_1 - q_2) \cdot b = r_2 - r_1 = 0$, which implies $q_1 = q_2$ (as $b > 0$). This completes the proof of the uniqueness part of the theorem. □

The Division Algorithm is a fundamental theorem in mathematics, with many applications. We will use it soon to prove other powerful theorems and methods related to divisibility, but before doing so, here is an example.

Example 6.1.3. Let n be an integer. Show that if n is **not divisible** by 5, then $n^4 - 1$ **is divisible** by 5.

We can easily confirm the claim for a few special cases:

- If $n = 2$ (not divisible by 5), then $2^4 - 1 = 15$ is divisible by 5.
- if $n = 9$, then $9^4 - 1 = 6560$ is divisible by 5. Etc.

Our task is to prove the claim for any $n \in \mathbb{N}$, and we can do so by using the division algorithm.

Proof. According to the division algorithm, there exist integers q and r, with $q \geq 0$ and $0 \leq r < 5$, such that

$$n = q \cdot 5 + r.$$

We know that n is not divisible by 5, and hence r cannot be zero (i.e., $r = 1, 2, 3$ or 4).

Note that $n^4 - 1 = (n^2 - 1)(n^2 + 1)$. To prove that $n^4 - 1$ is divisible by 5, it is enough to show that either $n^2 - 1$ or $n^2 + 1$ is a multiple of 5.

If $r = 1$ or 4, then $n = 5q + 1$ or $n = 5q + 4$, and hence

$$n^2{-}1 = (5q{+}1)^2{-}1 = 25q^2{+}10q = 5{\cdot}(5q^2{+}2q) \qquad \text{or} \qquad n^2{-}1 = (5q{+}4)^2{-}1 = 25q^2{+}40q{+}15 = 5{\cdot}(5q^2{+}8q{+}3).$$

If $r = 2$ or 3, then $n = 5q + 2$ or $n = 5q + 3$, and we get

$$n^2 + 1 = (5q + 2)^2 + 1 = 5 \cdot (5q^2 + 4q + 1) \qquad \text{or} \qquad n^2 + 1 = (5q + 3)^2 + 1 = 5 \cdot (5q^2 + 6q + 2).$$

This shows that for $r = 1, 2, 3, 4$ either $n^2 - 1$ or $n^2 + 1$ is divisible by 5, which implies that $n^4 - 1$ is divisible by 5, as needed. □

6.2 Greatest Common Divisors and the Euclidean Algorithm

The greatest common divisor (gcd) of two integers is a useful notion in number theory, with many applications in mathematics and science. It is defined as follows.

Definition 6.2.1. Let a and b be two integers, not both zero.
The **greatest common divisor** (or **GCD**) of a and b, denoted as $\gcd(a, b)$, is the largest integer that divides both numbers.
If $\gcd(a, b) = 1$, we say that a and b are **relatively prime**[1].

Note that as 1 divides every integer, the GCD of two integers must be at least 1 (and in particular – a natural number). Also note that the GCD of two numbers is a **symmetric operation**: $\gcd(a,b) = \gcd(b,a)$.

Examples. (a) Take $a = 36$ and $b = 42$. The (positive) divisors of a are $1, 2, 3, 4, 6, 9, 12, 18$ and 36, and the divisors of b are $1, 2, 3, 6, 7, 14, 21$ and 42. By comparing these lists, we can see that the largest number dividing both 36 and 42 is 6, and so $\gcd(36, 42) = 6$.

(b) The (positive) divisors of $a = 4$ are $1, 2$ and 4, and the divisors of $b = 9$ are $1, 3, 9$. Therefore, $\gcd(4, 9) = 1$. Namely, 4 and 9 are **relatively prime**.

(c) What if a and b are larger? Listing all their divisors, and making sure we have not missed any of them, can be a long and tedious process. For instance, take $a = 23814$ and $b = 8232$. To find their greatest common divisor more efficiently, we can decompose the numbers as products of prime numbers (which is guaranteed to be possible, by Theorem 4.5.1):

$$a = 11907 \cdot 2 = 1323 \cdot 9 \cdot 2 = 147 \cdot 9 \cdot 9 \cdot 2 = 49 \cdot 3 \cdot 9 \cdot 9 \cdot 2 = 7 \cdot 7 \cdot 3 \cdot 9 \cdot 9 \cdot 2 = 7^2 \cdot 3^5 \cdot 2$$
$$b = 2058 \cdot 4 = 1029 \cdot 2 \cdot 4 = 343 \cdot 3 \cdot 2 \cdot 4 = 7 \cdot 49 \cdot 3 \cdot 2 \cdot 4 = 7^3 \cdot 3 \cdot 2^3$$

Numbers, other than 1, which divide a and b, must be products of 2's, 3's and 7's. The 'longest' such product, that divides both a and b, is $7 \cdot 7 \cdot 3 \cdot 2 = 294$, and so $\gcd(23814, 8232) = 294$.

(d) If $a, b \in \mathbb{N}$, and $a|b$, then $\gcd(a, b) = a$, as a clearly divides both a and b, and any number larger than a cannot divide a.

(e) For any nonzero integer n, we have $\gcd(n, 0) = |n|$ (why?).

For really large numbers (with dozens or even hundreds of digits), prime factorization becomes extremely difficult, and in practice – impossible, **even with the aid of computers**. For instance, in 2009, a number known as RSA-768, with 232 decimal digits was factored, after **two years** of utilizing **hundreds of powerful computers**.

There is, however, an alternate method, known as **the Euclidean Algorithm**, for computing the greatest common divisor of two numbers, that does not involve prime factorization, nor the listing of all divisors. It relies on the division algorithm, which makes it much more efficient (computationally).

We first prove a Proposition, on which the algorithm is based, and then describe the algorithm in detail.

[1] The term 'relatively prime' may seem strange at first, but it will become clearer once we present the Fundamental Theorem of Arithmetic, in the next section.

Proposition 6.2.2. Let a, b and k be integers, with a and b not both zero. Then $\gcd(a, b) = \gcd(a-kb, b)$.

In other words, subtracting from a an integer multiple of b does not change the GCD.

Proof. To prove the proposition, we show that the pairs (a, b) and $(a-kb, b)$ have the same set of (positive) divisors.

- If $d \in \mathbb{N}$ divides both a and b, then both are multiples of d. That is, $a = md$ and $b = ld$ for some $m, l \in \mathbb{Z}$. But then $a - kb = md - kld = (m - kl)d$, and hence $d|a - kb$.
- Conversely, if $d \in \mathbb{N}$ divides both $a - kb$ and b, then $a - kb = td$ and $b = ld$ for some $t, l \in \mathbb{Z}$. Then we have
$$a = kb + td = kld + td = (kl + t)d,$$
which shows that $d|a$.

We conclude that the pairs (a, b) and $(a - kb, b)$ have the same set of positive divisors, and hence the same greatest common divisor, as needed. □

The Euclidean Algorithm.

Suppose that a and b are two natural numbers, with $a \geq b$. To compute the GCD of a and b, we apply the division algorithm repeatedly, as follows.

- Divide a by b, obtaining a quotient q_1 and a remainder r_1:
$$\boldsymbol{a} = q_1 \cdot \boldsymbol{b} + r_1.$$
- If $r_1 = 0$ (that is, if a is divisible by b), then stop. Otherwise, divide b by r_1. Denote the new quotient and remainder by q_2 and r_2:
$$\boldsymbol{b} = q_2 \cdot \boldsymbol{r_1} + r_2.$$
- If $r_2 = 0$, then stop. Otherwise, divide r_1 by r_2. Denote the new quotient and remainder by q_3 and r_3:
$$\boldsymbol{r_1} = q_3 \cdot \boldsymbol{r_2} + r_3.$$
- Keep repeating this process, as long as the remainder is not zero:
$$\boldsymbol{r_2} = q_4 \cdot \boldsymbol{r_3} + r_4 \quad , \quad \boldsymbol{r_3} = q_5 \cdot \boldsymbol{r_4} + r_5 \quad , \quad \boldsymbol{r_4} = q_6 \cdot \boldsymbol{r_5} + r_6 \quad , \quad \text{etc.}$$

The key observation here, is that the remainders form a **strictly decreasing sequence of nonnegative integers**, as the remainder is always smaller than the divisor (Theorem 6.1.2):

$$r_1 > r_2 > r_3 > r_4 > \dots$$

On the other hand, any decreasing sequence of nonnegative integers must terminate (why?), and so at some point, a remainder of zero will show up. Using Proposition 6.2.2, we show that **the last positive remainder must be the greatest common divisor of a and b.**

Claim 6.2.3. Let $a, b \in \mathbb{N}$ with $a \geq b$, and consider the remainders $r_1, r_2, r_3, \dots$ obtained from the process described above. If $r_n = 0$, then $r_{n-1} = \gcd(a, b)$.

Proof. According to Proposition 6.2.2, subtracting a multiple of one number from another does not change the GCD. We can therefore apply the proposition repeatedly, and get:

$$\begin{aligned}
\gcd(a,b) = \gcd(a - q_1 b, b) &= \gcd(r_1, b) &&= \\
&= \gcd(r_1, b - q_2 r_1) &&= \gcd(r_1, r_2) &&= \\
&= \gcd(r_1 - q_3 r_2, r_2) &&= \gcd(r_3, r_2) &&= \\
&= \gcd(r_3, r_2 - q_4 r_3) &&= \gcd(r_3, r_4) &&= \dots
\end{aligned}$$

This process terminates once we get to the n-th remainder, $r_n = 0$, and we conclude that

$$\gcd(a,b) = \gcd(r_1, r_2) = \gcd(r_2, r_3) = \dots = \gcd(r_{n-1}, r_n) = \gcd(r_{n-1}, 0) = r_{n-1},$$

as needed (recall that $\gcd(n, 0) = n$ for any natural number n). □

The procedure of repeatedly applying the division algorithm, as described above, and eventually finding the GCD of two given numbers, is called **the Euclidean Algorithm**. This is one of the oldest algorithms in mathematics which is in common use. The starting point of the algorithm is a pair of natural numbers (the original a and b), and in each step, one of the numbers in the pair is replaced by a smaller number, without affecting the GCD. Thus, as we saw, the pairs

$$(a,b)\ ,\ (b,r_1)\ ,\ (r_1,r_2)\ ,\ (r_2,r_3)\ ,\ (r_3,r_4)\ ,\dots$$

have all the same greatest common divisor. Once we get a pair with zero as one of the arguments, the other (positive) argument is the GCD of all the pairs (and in particular, of a and b).

Note that the algorithm can be used even when a and b are not both positive, since changing the sign of either a or b does not change their GCD. In other words, the following pairs have all the same GCD as the pair $(|a|, |b|)$:

$$(a,b) \quad , \quad (-a,b) \quad , \quad (a,-b) \quad , \quad (-a,-b).$$

Examples. (a) Suppose we want to find the GCD of 154 and 35. That is, $a = 154$ and $b = 35$. We start by dividing a by b, with remainder:

$$\mathbf{154} = 4 \cdot \mathbf{35} + 14.$$

As the remainder is **not** zero, we proceed with dividing 35 by 14, with remainder:

$$\mathbf{35} = 2 \cdot \mathbf{14} + 7.$$

The remainder is still not zero, so we continue, and divide 14 by 7:

$$\mathbf{14} = 2 \cdot \mathbf{7} + 0.$$

Now, the remainder is zero (as 14 is divisible by 7), and so the GCD of the two given numbers, is equal to our last nonzero remainder, which is 7:

$$\gcd(154, 35) = 7.$$

Another way to describe this process is by writing the pairs (typed above in **boldface** font) obtained in each step:

$$(154, 35) \quad \to \quad (35, 14) \quad \to \quad (14, 7) \quad \to \quad (7, 0)\ .$$

(b) It is quite easy to execute the Euclidean by hand (without any calculator), even when the initial numbers are larger. For instance, here we perform the algorithm on the numbers 1533 and 150.

$$\mathbf{1533} = 10 \cdot \mathbf{150} + 33$$
$$\mathbf{150} = 4 \cdot \mathbf{33} + 18$$
$$\mathbf{33} = 1 \cdot \mathbf{18} + 15$$
$$\mathbf{18} = 1 \cdot \mathbf{15} + 3$$
$$\mathbf{15} = 5 \cdot \mathbf{3} + 0$$

The last nonzero remainder is 3, and so $\gcd(1533, 150) = 3$.
As the GCD is not affected by changing the signs of our initial numbers, we also conclude that

$$\gcd(1533, -150) = \gcd(-1533, 150) = \gcd(-1533, -150) = 3.$$

(c) Suppose that x is a natural number **greater than** 1. What is the GCD of $x^7 - 1$ and $x^5 - 1$?
We can go ahead and apply the Euclidean algorithm, but note that our quotients and remainders

must be expressed in terms of the unknown number x. This is not too hard to do, as long as we remember how to divide polynomials (with remainder).

$$\boldsymbol{x^7 - 1} = x^2 \cdot (\boldsymbol{x^5 - 1}) + (x^2 - 1)$$
$$\boldsymbol{x^5 - 1} = (x^3 + x) \cdot (\boldsymbol{x^2 - 1}) + (x - 1)$$
$$\boldsymbol{x^2 - 1} = (x + 1) \cdot (\boldsymbol{x - 1}) + 0$$

Therefore, $\gcd(x^7 - 1, x^5 - 1) = x - 1$.

A fundamental and extremely useful property of the GCD, is that it can be always expressed as an **integer linear combination** of the two given numbers. This is the content of Bézout's Identity.

Theorem 6.2.4. *(Bézout's Identity)*
Let a and b be two integers, not both zero. Then there are $m, n \in \mathbb{Z}$, such that

$$a \cdot m + b \cdot n = \gcd(a, b).$$

One of the proofs of Bézout's Identity is based on the **back-substitutions** strategy, demonstrated below. This strategy provides us with a procedure for finding appropriate values for m and n. We discuss a few examples before presenting the proof of the theorem.

- We have previously found that $\gcd(154, 35) = 7$. We rewrite the steps produced by the Euclidean Algorithm, keeping, in each step, the remainder on the right-hand-side, and moving the 'quotient term' to the left. We also omit the last equality (where the remainder is zero).

$$\boldsymbol{154} - 4 \cdot \boldsymbol{35} = 14$$
$$\boldsymbol{35} - 2 \cdot \boldsymbol{14} = 7$$

Now, using the first equality, we replace the 14, in the second equation, with $154 - 4 \cdot 35$, and rearrange:

$$35 - 2 \cdot (154 - 4 \cdot 35) = 7 \quad \Rightarrow \quad 35 - 2 \cdot 154 + 8 \cdot 35 = 7 \quad \Rightarrow \quad 154 \cdot (-2) + 35 \cdot 9 = 7.$$

We have expressed the GCD of $a = 154$ and $b = 35$ as an integer linear combination:

$$\underbrace{154}_{a} \cdot \underbrace{(-2)}_{m} + \underbrace{35}_{b} \cdot \underbrace{9}_{n} = \underbrace{7}_{\gcd(154,35)}.$$

- We also computed the GCD of $a = 1533$ and $b = 150$ to be 3. Rewriting the steps obtained by the Euclidean Algorithm (and omitting the last step), we obtain the following.

$$\mathbf{1533} - 10 \cdot \mathbf{150} = 33$$
$$\mathbf{150} - 4 \cdot \mathbf{33} = 18$$
$$\mathbf{33} - 1 \cdot \mathbf{18} = 15$$
$$\mathbf{18} - 1 \cdot \mathbf{15} = 3$$

  The back-substitution procedure leads to the following.

$$\begin{aligned} 18 - 1 \cdot (33 - 1 \cdot 18) = 3 \quad &\Rightarrow \quad 2 \cdot 18 - 33 = 3 \\ 2 \cdot (150 - 4 \cdot 33) - 33 = 3 \quad &\Rightarrow \quad 2 \cdot 150 - 9 \cdot 33 = 3 \\ 2 \cdot 150 - 9 \cdot (1533 - 10 \cdot 150) = 3 \quad &\Rightarrow \quad 1533 \cdot (-9) + 150 \cdot 92 = 3 \end{aligned}$$

  And we obtained $1533 \cdot (-9) + 150 \cdot 92 = 3$, as needed.

Proof. (of Theorem 6.2.4) The back-substitution idea can be turned into an elegant proof using strong induction.

We assume, for now, that a and b are natural numbers, and perform induction **on the sum $a + b$**. The other cases (where a or b are zero or negative) are left as an exercise (see Exercise 6.4.18).

As the smallest sum of two natural numbers is 2, this is going to be our base case. If $a + b = 2$, then $a = b = 1$, and $\gcd(a, b) = \gcd(1, 1) = 1$. We can then pick $m = 1$ and $n = 0$ to satisfy the identity:

$$\underbrace{1}_{a} \cdot \underbrace{1}_{m} + \underbrace{1}_{b} \cdot \underbrace{0}_{n} = \underbrace{1}_{\gcd(a,b)} .$$

Now assume that the theorem is true for any $a, b \in \mathbb{N}$ with $2 \leq a + b \leq k$ (where $k \in \mathbb{N}$), and suppose that $a, b \in \mathbb{N}$ with $a+b = k+1$. We prove the $(k+1)$-st case by considering the following three possibilities.

- If $a = b$, then $\gcd(a, b) = a$, and we can write $a \cdot 1 + b \cdot 0 = \gcd(a, b)$. That is, the theorem holds true with $m = 1$ and $n = 0$.

- If $a < b$, then a and $b - a$ are natural numbers, whose sum is $a + (b - a) = b < a + b$, and thus covered by the induction hypotheses. That is, we have

$$a \cdot m + (b - a) \cdot n = \gcd(a, b - a).$$

  Remember though, that $\gcd(a, b - a) = \gcd(a, b)$ by Claim 6.2.3, and so, after rearranging the equation, we can write

$$a \cdot (m - n) + b \cdot n = \gcd(a, b),$$

which proves the $(k+1)$-st case.

- The last case, where $a > b$, is done similarly, by applying the induction hypothesis on the numbers $a - b$ and b.

We proved the $(k+1)$-st case, which completes the proof of the theorem. □

We end this section with an application of Bézout's Identity.

Example. Prove that there are integers x, y, that solve the equation $1533x + 150y = 27$.

We computed the GCD of 1533 and 150 to be 3 (see Example (b) on Page 149). According to Bézout's Identity, there are $m, n \in \mathbb{Z}$, for which $1533m + 150n = 3$ (we even found such m and n, but the actual values are not needed here). The numbers $x = 9m$ and $y = 9n$ solve the required equation, as

$$1533x + 150y = 1533 \cdot 9m + 150 \cdot 9n = 9 \cdot (1533m + 150n) = 9 \cdot 3 = 27.$$

6.3 The Fundamental Theorem of Arithmetic

We start with the following claim.

Claim 6.3.1 (Euclid's Lemma[2])**.**
Let $p > 1$ be a prime number, and $a, b \in \mathbb{Z}$. If $p|ab$, then $p|a$ or $p|b$.

Proof. To prove the claim, we show that $p \nmid a$ implies $p|b$ (why is that sufficient?).

As p is a prime number, its only positive divisors are 1 and p. Since $p \nmid a$, the greatest common divisor of a and p must be 1, i.e., $\gcd(a, p) = 1$. In other words, a and p are **relatively prime**. By Bézout's Identity, there are integers m, n for which

$$m \cdot a + n \cdot p = 1.$$

Multiplying both sides by b gives

$$m \cdot ab + n \cdot pb = b.$$

We know that $p|ab$, and so $p|(m \cdot ab)$. Also, $p|(n \cdot pb)$, and so p divides the whole left-hand side, $m \cdot ab + n \cdot pb$. Consequently, $p|b$, as needed. □

[2] In mathematics, a Lemma is a proposition (or a 'helping theorem') used as a stepping stone to prove a larger result, of possibly more interest.

Example. The product of 30 and 259 is 7770, and hence divisible by 7. As 7 is prime, we are guaranteed, by Euclid's Lemma, that 7 divides either 30 or 259. Indeed, 259 is divisible by 7, as $259 = 7 \cdot 37$.

Remarks.

- Euclid's Lemma can be easily extended to products of more than two integers:

  If p divides a product of n integers, it must divide one of the factors.

  The proof of this statement (by induction) is left as an exercise (see Exercise 6.4.22).

- The requirement that p is a prime number is essential. If we take, for instance, $a = 4$ and $b = 9$, then their product, $ab = 36$, is divisible by 6. However, neither 4 nor 9 is divisible by 6.

Previously, we mentioned without proof, that the number $\sqrt{7}$ is irrational (i.e., it cannot be expressed as a quotient of two integers). Now, that we have Euclid's Lemma at our disposal, we are ready to prove this fact. The strategy we use can be applied to other radical numbers, such as $\sqrt{2}$, $\sqrt[5]{9}$, etc.

Claim 6.3.2. The number $\sqrt{7}$ is irrational. That is, $\sqrt{7} \notin \mathbb{Q}$.

Proof. We prove the claim by contradiction. Assume that $\sqrt{7}$ is a rational number. Then $\sqrt{7} = \frac{a}{b}$ for some nonzero integers a, b. As $\sqrt{7} > 0$, we may assume that $a, b > 0$. Moreover, we assume that $\gcd(a, b) = 1$ (namely, a and b are relatively prime), in which case the fraction $\frac{a}{b}$ is said to be in **lowest terms**, or **completely reduced**.

From the equality $\sqrt{7} = \frac{a}{b}$, we get

$$7 = \frac{a^2}{b^2} \qquad \Rightarrow \qquad 7b^2 = a^2,$$

and hence $7|a^2$ (or $7|a \cdot a$). By Euclid's Lemma, $7|a$, and hence $a = 7n$ for some integer n. Replacing a by $7n$ gives

$$7b^2 = (7n)^2 \qquad \Rightarrow \qquad 7b^2 = 49n^2 \qquad \Rightarrow \qquad b^2 = 7n^2,$$

from which we conclude that $7|b^2$. Again, by Euclid's Lemma, we see that $7|b$, leading to a contradiction. The fraction $\frac{a}{b}$ was assumed to be in lowest terms, which is inconsistent with our conclusion, that both a and b are divisible by 7.

Consequently, our initial assumption must be false, and thus $\sqrt{7} \notin \mathbb{Q}$. □

We have seen, in Chapter 4, that every natural number, greater than 1, can be expressed as a product of prime numbers (see Theorem 4.5.1). In other words, prime numbers are, in some sense, the **building**

blocks of all the natural numbers. If a product of prime numbers includes repetitions, we can use exponents to shorten the product. For instance:

$$3969 = 3 \cdot 3 \cdot 3 \cdot 3 \cdot 7 \cdot 7 = 3^4 \cdot 7^2$$
$$169400 = 2 \cdot 2 \cdot 2 \cdot 5 \cdot 5 \cdot 7 \cdot 11 \cdot 11 = 2^3 \cdot 5^2 \cdot 7 \cdot 11^2 \ .$$

The Fundamental Theorem of Arithmetic, stated below, extends Theorem 4.5.1, adding that factoring a number as a product of primes can be done, essentially, in one way only. We can certainly reorder the exponents, and write 3969 as $7^2 \cdot 3^4$ (instead of $3^4 \cdot 7^2$), but that is the only change we can make. The prime numbers appearing in the product, and the corresponding exponents, are determined uniquely.

Theorem 6.3.3 (The Fundamental Theorem of Arithmetic)**.** *Every natural number $n \geq 2$ is either a prime, or can be expressed as a product of powers of distinct primes, in a unique way (except for reordering of the factors).*

Proof. We have already proved the existence part of the theorem (Theorem 4.5.1). We thus proceed by proving **the uniqueness part**, using strong induction.

For the base case, $n = 2$, the theorem holds true, as 2 is a prime number. Assume that the uniqueness part of the theorem is true for $n = 2, 3, 4, \ldots, k$, and consider the number $k+1$. We already know that $k+1$ is a prime number, or can be expressed as a product of prime numbers. Our task is to prove that, if $k+1$ is composite, then its factorization is unique.

We use a similar strategy we used in the proof of the Division Algorithm (Theorem 6.1.2). Suppose that we can factor $k+1$, as a product of distinct powers of primes, in two ways. That is

$$k+1 = p_1^{a_1} \cdot p_2^{a_2} \cdot \cdots \cdot p_\ell^{a_\ell} \ = \ q_1^{b_1} \cdot q_2^{b_2} \cdot \cdots \cdot q_m^{b_m} \ , \qquad (*)$$

where $p_1, \ldots, p_\ell, q_1, \ldots, q_m$ are prime numbers, and $a_1, \ldots, a_\ell, b_1, \ldots, b_m$ are natural numbers.

Clearly, $p_1 | k+1$, as p_1 appears in the product $p_1^{a_1} \cdot p_2^{a_2} \cdot \cdots \cdot p_\ell^{a_\ell}$. Therefore, p_1 must divide the other representation of $k+1$:

$$p_1 | q_1^{b_1} \cdot q_2^{b_2} \cdot \cdots \cdot q_m^{b_m} .$$

Now, as p_1 is prime, we conclude, from Euclid's Lemma, that it must divide one of the numbers $q_1, \ldots, q_m$. Assume, for simplicity, that $p_1 | q_1$ (if p_1 divides one of the other q's, we can re-assign indices, so that $p_1 | q_1$).

Remember that q_1 is also a prime number, and so its only positive divisors are 1 and q_1. As $p_1 \neq 1$, we conclude that $p_1 = q_1$. We now divide the equalities $(*)$ by p_1 (or, equivalently, by q_1), and obtain

$$\frac{k+1}{p_1} = p_1^{a_1 - 1} \cdot p_2^{a_2} \cdot \cdots \cdot p_\ell^{a_\ell} \ = \ q_1^{b_1 - 1} \cdot q_2^{b_2} \cdot \cdots \cdot q_m^{b_m} \ .$$

Finally, we can use the induction hypothesis to complete the proof. As $\frac{k+1}{p_1}$ is smaller than $k+1$, it is covered by our hypothesis, and thus satisfies the uniqueness part of the theorem. This means that the p's, the q's, and the corresponding exponents must be the same. More explicitly,

- The number of factors is the same. That is, $\ell = m$.
- The prime factors themselves have to be the same, though perhaps in some other arrangement. Thus, possibly after some re-indexing, we have $p_1 = q_1$, $p_2 = q_2$, $\ldots$, $p_\ell = q_\ell$.
- The exponents on the factors have to be the same, i.e., $a_1 - 1 = b_1 - 1$, $a_2 = b_2$, $\ldots$, $a_\ell = b_\ell$.

We conclude, from the above observations, that the two initial factorizations for $k+1$ (in $(*)$) were identical, as needed. We have thus proved the uniqueness part for $k+1$, which concludes the proof of the theorem. □

As a quick application of the Fundamental Theorem of Arithmetic, we prove that a certain logarithm is irrational.

Example 6.3.4. The number $\log_{48}(72)$ is irrational.

Proof. Assume, by contradiction, that $\log_{48}(72)$ is rational. That is,

$$\log_{48}(72) = \frac{m}{n}$$

for some $m, n \in \mathbb{Z}$, with $n \neq 0$. In fact, we may assume that $m, n \in \mathbb{N}$, as $\log_{48}(72) > 0$.

Recall that the equality $\log_a(b) = c$ is equivalent to $b = a^c$, and so

$$\log_{48}(72) = \frac{m}{n} \quad \Rightarrow \quad 72 = 48^{m/n} \quad \Rightarrow \quad 72^n = 48^m \ .$$

We proceed by factoring 48 and 72 and a product of powers of distinct prime numbers:

$$(2^3 \cdot 3^2)^n = (2^4 \cdot 3)^m \quad \Rightarrow \quad 2^{3n} \cdot 3^{2n} = 2^{4m} \cdot 3^m \ .$$

Now we apply the uniqueness part of Theorem 6.3.3. As the last equality represents a number as a product of powers of distinct primes, the exponents must match:

$$\begin{aligned} 3n &= 4m \\ 2n &= m \end{aligned} \quad \Rightarrow \quad 3n = 4 \cdot 2n \quad \Rightarrow \quad n = 0 \ .$$

This is a contradiction, as $n \in \mathbb{N}$, and so we conclude that $\log_{48}(72)$ is indeed an irrational number, as needed. □

6.4 Exercises for Chapter 6

6.4.1. Some tend to confuse the divisibility symbol, |, with symbols such as — or /, used for fractions and division. Choose the correct notation in each of the phrases below. Explain your choice briefly.

(a) As | 2\|58 | 58\|2 | 2/58 | 58/2 | , we conclude that 58 is an even number.

(b) Since | 60\|15 | 15\|60 | 60/15 | 15/60 | , 60 is divisible by 15.

(c) As | 7\|84 | 84\|7 | 7/84 | 84/7 | is an integer, 84 is divisible by 7.

(d) If | $b\|a$ | $a\|b$ | b/a | a/b | (where $b \neq 0$), then | $b\|a$ | $a\|b$ | b/a | a/b | is an integer.

6.4.2. Which of the following are **true** statements? Explain.

(a) $6|54$

(b) $33|3$

(c) $-1|1$

(d) $4 \not| 2$

(e) $4 \not| 8$

(f) $-11 \not| -111$

6.4.3. Which number is divisible by 7^5 ? Explain.

- $210^3 \cdot 98^2$
- $7^2 \cdot 17^9$
- $77^4 \cdot 5^7$
- $27^5 \cdot 35^4$

6.4.4. Let a, b, d be three integers, with $d \neq 0$. For each statement, decide if it is **true** or **false**. Provide a short proof or a counterexample.

(a) If $d|a$ and $d|b$, then $d|(a+b)$.

(b) If $d|(a+b)$, then $d|a$ and $d|b$.

(c) If $d|(a+b)$, then $d|a$ or $d|b$.

(d) If $d|(a+b)$, then either d divides both a and b, or d does not divide a nor b.

6.4.5. Let a, b, d be three integers, with $d \neq 0$. For each statement, decide if it is **true** or **false**. Provide a short proof or a counterexample.

(a) If $d|a$ and $d|b$, then $d|a \cdot b$.

(b) If $d|a \cdot b$, then $d|a$ or $d|b$.

(c) If $d|a$ or $d|b$, then $d|a \cdot b$.

(d) **(Harder!)** If $d^2|a^2$, then $d|a$.

6.4.6. In Example 6.1.3, we proved that for any $n \in \mathbb{N}$, if $5 \nmid n$, then $5|n^4 - 1$. Could we have used induction to prove this claim?

6.4.7. There are ways to generalize the Division Algorithm (Theorem 6.1.2), so that it can be applied to all integers (both positive and negative). Here is one possible generalization.

> **If $a, b \in \mathbb{Z}$, with $b \neq 0$, then there is a unique pair of integers, q and r, with $0 \leq r < |b|$, such that $a = q \cdot b + r$.**

Note that the remainder is still required to be nonnegative, so for instance, if we divide -21 by -4, the quotient is 6 and the remainder is 3, as $-21 = 6 \cdot (-4) + 3$.

(a) Find the quotient and the remainder, obtained when dividing a by b.

- $a = 27$ and $b = -8$.
- $a = -15$ and $b = 2$.
- $a = -36$ and $b = -9$.
- $a = 5$ and $b = -7$.
- $a = -4$ and $b = 9$.

(b) Prove the generalized version of the Division Algorithm given above.
(Hint: Instead of using induction, proceed by cases, and refer to Theorem 6.1.2.)

6.4.8. Let a, b, c be three integers, such that $a^2 + b^2 = c^2$. Show that if c is even, then both a and b are even.

6.4.9. Choose a three-digit number (for instance, 273). Form a six-digit number, by repeating the digits of the original number twice (i.e., 273273). Prove that the resulting number is not prime. In fact, prove that it has at least three distinct prime factors.

6.4.10. In Definition 6.2.1, why do we require that a and b are not both zero? Why can we allow one of them to be zero?

6.4.11. Use prime factorization to compute the GCD of the following pairs of numbers (without a calculator).

(a) 210 and 405

(b) $10^6 \cdot 6^2 \cdot 5^{11}$ and $6 \cdot 15 \cdot 3^7$

(c) $18^2 \cdot 21^3$ and $6 \cdot 10^3 \cdot 5^3$

(d) $300 \cdot 35 \cdot 7^5$ and $33^5 \cdot 3 \cdot 64$

6.4.12. Let $a, b \in \mathbb{N}$ and $d = \gcd(a, b)$. Prove that $\gcd\left(\frac{a}{d}, \frac{b}{d}\right) = 1$.

6.4.13. (a) Let n be a natural number. What is the GCD of $7n + 1$ and $8n + 1$?

(b) Let $a, b \in \mathbb{Z}$, not both zero. Prove that $\gcd(a+b, a+2b) = \gcd(a, b)$.

6.4.14. Let n be a natural number.

(a) What are all the possible values of $\gcd(n+1, 2-n)$? Explain.

(b) What are all possible values of $\gcd(7^n, 7^n+4)$? Explain.

6.4.15. Find, with proof, all the integers a that satisfy the equation $\gcd(a, 10) = a$.

6.4.16. Use the Euclidean Algorithm to compute the GCD of the following pairs of numbers (without a calculator). Also, express the GCD as an **integer linear combination** of the two numbers.

(a) 1872 and 300

(b) 15477 and 15477154

(c) 270028 and 27

(d) 35530 and 355

(e) 325299 and 325

(f) 24 and $54 + 24^7$

6.4.17. Let a, b be two positive integers. We define the **Least Common Multiple** (LCM) of a and b, denoted as $\text{lcm}(a, b)$, to be the smallest positive integer, that is divisible by both a and b. For instance, $\text{lcm}(12, 18) = 36$.

(a) Find $\text{lcm}(6, 15)$, $\text{lcm}(4, 22)$ and $\text{lcm}(9, 5)$ (do not use any computing device).

(b) Let $a \in \mathbb{N}$. What are $\text{lcm}(1, a)$, $\text{lcm}(a, a)$ and $\text{lcm}(a, a+1)$? Explain.

(c) Prove that if $a, b \in \mathbb{N}$, then $a \cdot b = \gcd(a, b) \cdot \text{lcm}(a, b)$.

(d) When would the LCM of two natural numbers be their product? Explain.

6.4.18. Complete the proof of Bézout's Identity, in the cases where $a \leq 0$ or $b \leq 0$.

6.4.19. (a) Let a, b, c be integers.

Prove that the equation $ax + by = c$ has an integer solution (i.e., there are $x, y \in \mathbb{Z}$ solving the equation) if and only if $\gcd(a, b) | c$.

(Hint: One implication is quite simple. For the other, use Bézout's identity.)

(b) For which of the following values of m will the equation $12x + my = 30$ have integer solutions?

- $m = 12$
- $m = 16$
- $m = 18$
- $m = 24$

6.4.20. Let a, b, c be integers.

Prove that if a pair of integers (x_0, y_0) solves the equation $ax + by = c$, then the pair $(x_0 + k \cdot b, y_0 - k \cdot a)$ is also an integer solution of $ax + by = c$ (where k is an arbitrary integer).

This shows that if $ax + by = c$ has an integer solution, then it has **infinitely many** solutions.

6.4.21. Let a, b, d be nonzero integers. Show that if $d|a$ and $d|b$, then $d| \gcd(a, b)$.

6.4.22. Prove, by induction, the following generalization of Euclid's Lemma (Claim 6.3.1):

If a prime p divides a product of n integers, it must divide one of the factors.

6.4.23. Find three integers a, b, c, such that their product, abc, is divisible by 33, but none of them is divisible by 33. Does that contradict Euclid's Lemma (Claim 6.3.1)? Explain.

6.4.24. Let a, b, c be three integers. Prove that if a and c relatively prime, and $c|ab$, then $c|b$.

(Hint: Read carefully the proof of Euclid's Lemma.)

6.4.25. If j and k are natural numbers, and $37j = 12k$, prove that $j + k$ is divisible by 7.

6.4.26. Prove that $\sqrt[3]{20}$ is an irrational number.

6.4.27. Prove, that for any **odd** integer k, the number $\sqrt{2k}$ is irrational.

6.4.28. (a) Prove that $\sqrt{11}$ is an irrational number.

(b) Consider the function $f : \mathbb{N} \times \mathbb{N} \to \mathbb{R} \quad , \quad f(a, b) = a + b \cdot \sqrt{11}$.

Is f **injective**? Is it **surjective**? Justify your arguments.

6.4.29. If $n, m \in \mathbb{N}$ such that $6^{2m+2} \cdot 3^n = 4^n \cdot 9^{m+3}$, what are n and m? Explain.

6.4.30. Is the following statement **true** or **false**? Provide a proof or a counterexample.

"If $2^a \cdot 4^b = 2^c \cdot 4^d$, then $a = c$ and $b = d$."

6.4.31. Consider the function $f : \mathbb{N} \times \mathbb{N} \to \mathbb{N}$ given by $f(n, m) = 5^n \cdot 7^m$. Is f **injective**? Is f **surjective**? **Justify** your answer.

6.4.32. Consider the function $f : \mathbb{N} \times \mathbb{N} \to \mathbb{N}$ given by $f(a, b) = 12^a \cdot 18^b$. Is f **injective**? Is f **surjective**? **Justify** your answer.

6.4.33. Prove that if p and q are two distinct (positive) prime numbers, then $\log_p(q)$ is irrational.

6.4.34. (a) Show that $\log_{216}(36)$ is a **rational** number.

(b) Obviously, if we try to prove that $\log_{216}(36)$ is irrational, by applying the strategy from Example 6.3.4, we will fail. Where exactly does the argument break? Explain.

Chapter 7

Relations

In this chapter, we discuss **relations**, a central notion in mathematics. As we will shortly see, we have already encountered and used many mathematical relations, without being aware of it. We begin by formally defining what a relation is, and then introduce a special type of relation, called an **equivalence relation**, and its associated notion of an **equivalence class**. In Section 7.4, we study an important and useful equivalence relation: **congruence modulo n**.

7.1 The Definition of a Relation

The word 'relation' (or 'relationship') is used in everyday language to describe a certain connection between two (or more) people or objects. For instance, some people are Canadian citizens, and some are not. Some are U.S. citizens, and some are not. "Being a citizen" is a relation between, or a way of connecting, countries and people. Given a person x and a country y, x either is or is not a citizen of country y. In other words, given a pair (x, y) of a person and a country, the pair either satisfies the relation of "the first being a citizen of the second", or it does not.

Marriage is another example that comes to mind when thinking about relations between people. If two people are put in front of us, then either they are married, or they are not. Namely, any given pair of people either satisfies or does not satisfy the relation of "being married".

The relation "being the capital" can be applied to cities and countries. For instance, Ottawa is the capital of Canada, Paris is the capital of France, and Tel-Aviv is **not** the capital of Israel. In other words, the pairs (Ottawa, Canada) and (Paris, France) satisfy the relation of "being the capital of", while (Tel-Aviv, Israel) does not.

As we can see, a relation is (informally), a condition, connection, or property, that is satisfied by some pairs of objects (or people), drawn from certain collections such as countries, people, cities, etc.

In this course, we focus on **mathematical relations**, and we have seen plenty of them already:

- Given two real numbers, x and y, either $x < y$ or $x \not< y$. The condition of "being less than" is an example of a relation in mathematics.
- "Being an element of a set" is a relation. Given a set A and an object a, either $a \in A$ or $a \notin A$.
- In Chapter 6 we encountered the notion of divisibility, which is another example of a relation. Given two integers a and b (with $b \neq 0$), either a is or is not divisible by b (i.e., either $b|a$ or $b \not| a$).

To be able to work with relations in mathematics, and prove things about them, we need a formal definition. It is perfectly fine to think of relations as conditions or properties satisfied by certain objects, but in a formal definition, we must stay away from informal notions such as "condition" or "connection".

To formalize the idea of a relation in mathematics, let us take a closer look at the example of cities, countries, and the relation of "being the capital". The two relevant sets in this example are cities and countries, so we denote:

$$A = \{\text{all cities}\} \qquad \text{and} \qquad B = \{\text{all countries}\} \ .$$

Whenever we form a pair of elements (x, y), where x is a city and y is a country, that pair either satisfies or does not satisfy the relation. In other words, a relation **separates** the collection of all pairs in $A \times B$ into two categories – pairs that satisfy our relation, and pairs that do not. This observation leads to a formal and elegant definition of a relation in mathematics. Instead of referring to a "condition" or a "property", we can simply **declare** which pairs satisfy the relevant condition, and **define a relation as the collection of these pairs**.

Definition 7.1.1. Let A and B two sets. A **relation** R between A and B is a subset of $A \times B$.
That is, a set R so that $R \subseteq A \times B$.

Remarks.

- In mathematics, we often refer to a subset of $A \times B$ as a **binary** relation, as it involves only two sets, A and B. More generally, we can define a relation on any number of sets (say, $A_1, A_2, \ldots, A_n$), as a **subset** $R \subseteq A_1 \times A_2 \times \cdots \times A_n$. However, in these notes, we focus on binary relations only.
- When the sets A and B are equal, we simply say that R **is a relation on** A.

Examples.

(a) Let $A = \{1, 2, 3, 4, 5\}$ and $B = \{6, 7, 8, 9, 10\}$. We define a relation R between A and B as follows:

$$R = \{(1, 6), (1, 7), (1, 8), (2, 6), (2, 7), (3, 6)\}.$$

Note that R is a subset of $A \times B$. The pair $(2, 6)$ satisfies the relation, while $(4, 8)$ does not.
We can describe the relation R in a "condition-type" fashion, as

$$R = \{(x, y) : \ x + y \leq 9\},$$

and think of R as the relation of "having a sum less than or equal to 9". Nevertheless, simply listing the pairs is a perfectly valid definition.

(b) The **divisibility relation**, on $\mathbb{N}$, can be described as a set of pairs:

$$\{(a, b) : \ a \ \text{ is divisible by } \ b\}.$$

This is a subset of $\mathbb{N} \times \mathbb{N}$. The pair $(36, 4)$, for instance, satisfies the relation (namely, $(36, 4) \in R$), as 36 is divisible by 4. However, $(53, 8) \notin R$, as 53 is **not** divisible by 8.

(c) Let A be the set of all the words in the English language. Define the relation R between A and $\mathbb{N}$ as

$$R = \{(X, n) : \text{ the number of letters in } X \text{ is } n\} \subseteq A \times \mathbb{N}.$$

The pair $(\text{wisdom}, 6)$ satisfies the relation, while $(\text{computer}, 9)$ does not.

Symbols for Commonly Used Relations.

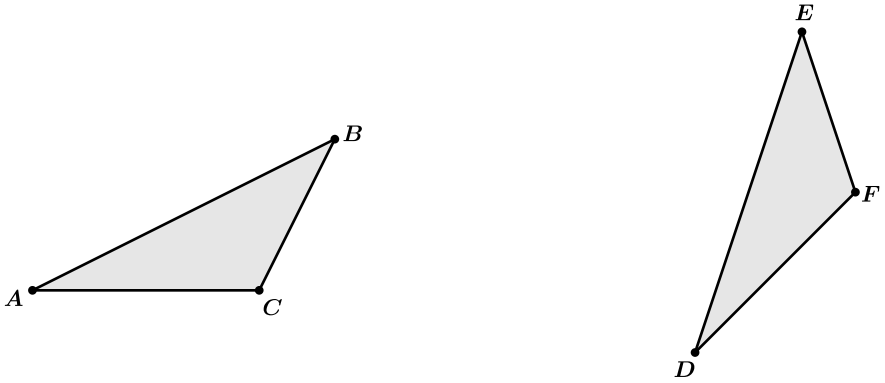
The 'pair-notation' is useful in many cases. However, there are many relations in mathematics, for which we have special symbols, and we tend to prefer them over pairs. For instance:

- The symbol $\mid$ is used to denote **divisibility**, and we usually write $3|15$ and $6 \nmid 14$, rather than $(3, 15) \in R$ and $(6, 14) \notin R$.
- The symbols $\in$ and $\subseteq$ are used to denote membership in a set, and inclusion of sets.
- Symbols such as $<$, $\leq$ and $>$ are used for **order relations** ("greater than", "less than or equal to", etc.).

7.2 Equivalence Relations

In mathematics, we often encounter objects that represent the same abstract entity. For instance, the two simple fractions $\frac{2}{3}$ and $\frac{8}{12}$, though they look different, represent the same abstract number **two-thirds**. In fact, we even say that $\frac{2}{3}$ and $\frac{8}{12}$ are equal, and write $\frac{2}{3} = \frac{8}{12}$.

The triangles $\triangle ABC$ and $\triangle DEF$ below are congruent, and so can be thought of as being two copies of the same abstract triangle. In some sense, these two triangles are equal (or equivalent).



The notion of an equivalence relation formalizes this idea. Here is the definition.

Definition 7.2.1. An **equivalence relation** R on a set S is a relation (that is, $R \subseteq S \times S$), such that:

(a) For any $x \in S$, $(x, x) \in R$ (**reflexive** property).

(b) For any $x, y \in S$, if $(x, y) \in R$, then $(y, x) \in R$ (**symmetric** property).

(c) For any $x, y, z \in S$, if $(x, y) \in R$ and $(y, z) \in R$, then $(x, z) \in R$ (**transitive** property).

In words, a relation that is **reflexive**, **symmetric** and **transitive**, is an equivalence relation. At the moment, it might be hard to see how these three properties capture the idea of equivalence. We will clarify this point later. For now, let us look at a few examples.

Examples.

(a) The **order relation** $<$ ('less than') on the set of **real numbers** $\mathbb{R}$ is ...

- **not reflexive**, as for instance, $2 \not< 2$ (and so $(2, 2)$ is not part of the relation).
- **not symmetric**, since $5 < 7$ but $7 \not< 5$ (i.e., $(5, 7)$ is in the relation, while $(7, 5)$ is not).
- **transitive**, as if $x < y$ and $y < z$, then $x < z$ (namely, if (x, y) and (y, z) satisfy the relation, then so does (x, z)).

Therefore, the relation $<$ is **not** an equivalence relation.

Note how all three requirements in Definition 7.2.1 involve universal quantifiers ('for all...'). For this reason, we could use specific numbers to disprove reflexivity and symmetry, while we had to use variables (such as x, y, z) to prove transitivity.

(b) The **inclusion relation** $\subseteq$ on the collection of sets is ...

- **reflexive**, as for any set B, we have $B \subseteq B$ (see remark on page 26).
- **not symmetric**, as we can easily find an example of a set that is properly contained in another. For instance, $\{1,2,3\} \subseteq \{1,2,3,4,5\}$, but $\{1,2,3,4,5\} \not\subseteq \{1,2,3\}$.
- **transitive**, since if A, B, C are sets, with $A \subseteq B$ and $B \subseteq C$, then $A \subseteq C$.

This relation is not symmetric, and hence **not** an equivalence relation.

(c) Define the following relation R on the set of real numbers:

$$R = \{(x,y)\colon x^2 + y^2 = 1\} \subseteq \mathbb{R} \times \mathbb{R}.$$

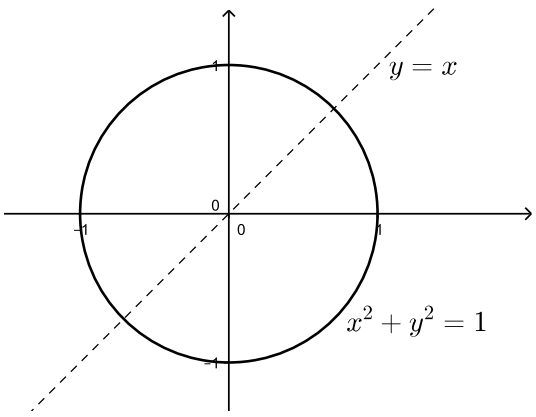
A number x satisfies this relation, with respect of another number y, if and only if $x^2 + y^2 = 1$. For instance $\left(\frac{3}{5}, \frac{4}{5}\right) \in R$, while $\left(\frac{1}{2}, \frac{1}{2}\right) \notin R$. Is R reflexive? Symmetric? Transitive? Let us check.

- We have just mentioned that $\left(\frac{1}{2}, \frac{1}{2}\right)$ is not part of the relation R, and so R is **not reflexive**. Note that the pair $\left(\frac{1}{\sqrt{2}}, \frac{1}{\sqrt{2}}\right)$ does satisfy the relation, but reflexivity requires that **any** pair of the form (x,x) be in R, which is not the case here.
- The relation R **is symmetric**. If $(x,y) \in R$, then $x^2 + y^2 = 1$, or equivalently, $y^2 + x^2 = 1$. This shows that the pair (y,x) is in R as well, which proves the symmetry of the relation.
- Transitivity, in many cases, is harder to detect. Here, we can experiment with a couple of pairs, and conclude that R is **not transitive**.

  Take, for instance, $x = -1$, $y = 0$ and $z = 1$. Then $(x,y) \in R$ (as $(-1)^2 + 0^2 = 1$), and $(y,z) \in R$ (as $0^2 + 1^2 = 1$). However, $(x,z) \notin R$, since $(-1)^2 + 1^2 \neq 1$, and therefore transitivity fails.

In this example, R is **not** an equivalence relation, as it is not reflexive nor transitive.

A relation on a set of real numbers is a subset of the Cartesian product $\mathbb{R}^2 = \mathbb{R} \times \mathbb{R}$, and hence can be visualized as a curve or a region in the two-dimensional plane. Here, the relation R is nothing but the unit circle in $\mathbb{R}^2$.



A reflexive relation (on $\mathbb{R}$) must contain all the pairs of the form (x, x), and hence the line $y = x$. Our unit circle has only two points in common with the line $y = x$, which implies that R is **not reflexive**.

On the other hand, the unit circle is symmetric with respect to $y = x$ (or to switching x and y), and therefore represents a **symmetric relation**.

There is no simple way to detect transitivity from the graph of a relation on $\mathbb{R}$.

(d) Define a relation D on the set of **integers**, as follows:

$$D = \{(a, b) \colon a + b \text{ is an even number}\}.$$

We now prove that D is an equivalence relation, by showing that it is reflexive, symmetric and transitive.

- For any $a \in \mathbb{Z}$, we have $a + a = 2a$, which is an even number. Therefore, $(a, a) \in D$ and so D is reflexive.
- For any $a, b \in \mathbb{Z}$, if $a + b$ is an even number, then so is $b + a$. Consequently, $(a, b) \in D$ implies $(b, a) \in D$, which proves the symmetry of D.
- Finally, we show that D is transitive. Let $a, b, c \in \mathbb{Z}$ such that $(a, b) \in D$ and $(b, c) \in D$ (i.e., $a + b$ and $b + c$ are both even numbers). Our task is to prove that $(a, c) \in D$, namely – that $a + c$ is even, and we do so as follows. Write

$$a + c = (a + b) + (b + c) - 2b.$$

We know that $a + b$ and $b + c$ are even, and clearly $2b$ is even as well. Sums and differences of even numbers are also even, and hence $a + c$ is an even number as needed.

Remarks.

(a) On **any** set S, there is always an obvious equivalence relation – **the equality relation**! In other words, the pair (x, y) satisfies the relation if and only if $x = y$:

$$R = \{(x, x) \colon x \in S\}.$$

Equality is an equivalence relation, as it is reflexive, symmetric and transitive (check!). We often think of equivalence relations as a **generalization**, or **extension**, of the notion of equality.

(b) We have mentioned, in the beginning of the section, equivalence relations that we have already encountered in previous studies. Congruence and similarity of triangles, and equivalence of fractions are examples of equivalence relations. They are all reflexive, symmetric and transitive.
If two line segments, say AB and CD, have the same length, we often write $AB = CD$, even though strictly speaking, AB and CD are not equal as sets of points in the plane. In fact, we are using the relation of 'having the same length' to identify two line segments of the same length. This is another example of an equivalence relation.

(c) For equivalence relations, we often use symbols such as $\equiv, \sim, \approx, \simeq, \cong, \cong$, etc. instead of the pair notation. These symbols resemble the equality symbol, and further emphasize the fact that elements satisfying an equivalence relation are often though of as being **equal** or **equivalent**.

For instance, we may write $a \cong b$ or $a \equiv b$, instead of $(a, b) \in R$.

7.3 Equivalence Classes

We have already mentioned, that given an equivalence relation on a set, pairs of elements satisfying the relation are often thought of as representing the same abstract idea. For instance, the fractions $\frac{4}{10}$ and $\frac{14}{35}$ represent the same number **two-fifths**. Congruent triangles represent copies of the same **abstract triangle**, and so on.

We can even find the notion of equivalence outside mathematics. For instance, young children, who learn about colors, are taught that a tomato's color is red, and that a strawberry is red as well. They are

then able to identify objects as being red, as **they have the same color** as a tomato (or a strawberry). For instance, a car or a shirt is red, because it has the same color as a tomato.

Children learn about colors without being given precise definitions. How would you define the color 'red' without referring to physical objects? Proper definitions can be given using frequency or wavelengths of electromagnetic waves, but that is not how we first learn about colors. An understanding of what it means for two objects to have the same color, and knowing that a tomato is red, is enough in order to identify other red objects. If we define

$$S = \{\text{All objects, which have the same color as a tomato}\},$$

we may treat S as a set, representing the abstract notion of 'the color red'. That is, the notion 'red' is defined as a property of a collection of objects, grouped together.

Note that '**having the same color**' is an equivalence relation on a collection of objects.

Similarly, the abstract number **two-fifths**, can be described as the set of all fractions $\frac{a}{b}$, in which $5a = 2b$.

Given an arbitrary equivalence relation on a set, we may group equivalent elements together, to form **equivalence classes**. Each equivalence class may be thought of as one abstract entity. Every member of the class is **a representative** of that class. Here is the definition.

Definition 7.3.1. Let R be an equivalence relation on a set S, and $x \in S$.
The **equivalence class of** x is the set of all elements $y \in S$, which are equivalent to x:

$$\{y \in S \colon (x, y) \in R\}.$$

We denote the equivalence class of x by $[x]$.

Examples.

(a) Define an equivalence relation $\sim$ on the set of real numbers, as follows:

$$x \sim y \qquad \text{if and only if} \qquad x \text{ and } y \text{ have } \textbf{the same truncation} \text{ (or } \textbf{integer part}\text{)}.$$

By '**truncation**' or '**integer part**', we mean the integer left after removing all digits to the right of the decimal point. For instance:

- The truncation of 4.35 is 4, and the truncation of $\pi = 3.1415\ldots$ is 3.
- The truncation of -6.01 is -6, and the truncation of -9.99 is -9.

- The truncation of both 0.123 and -0.734 is 0.

The numbers 5.11 and 5.9999 satisfy the relation, as they have the same truncation. We can then write $5.11 \sim 5.9999$. Similarly, we have $-0.5 \sim 0.5$.

On the other hand, 4.1 and -4.1 do not have the same truncation, and hence do not satisfy the relation: $4.1 \not\sim -4.1$.

Having the same truncation is an equivalence relation (check that it is indeed reflexive, symmetric and transitive). How do equivalence classes look like for this relation?

For example, what is the equivalence class of the number 2.35? Well, by definition, we need to find all real numbers that have the same truncation as 2.35 (which is 2). These are all the numbers between 2 and 3, including 2, but not including 3. In other words:

The equivalence class of 2.35 is the half-open and half-closed interval $[2, 3)$.

Similarly, the equivalence class of -6.5, is the set of all real numbers with truncation -6, and so:

The equivalence class of -6.5 is the interval $(-7, -6]$.

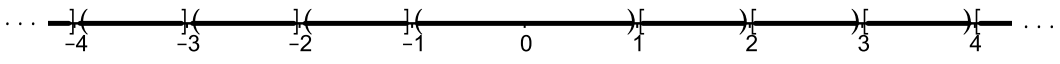
What is the equivalence class of 0? There are both positive and negative numbers, whose truncation is 0. In this case, the equivalence class is an open interval, containing all the numbers between -1 and 1:

The equivalence class of 0 is the open interval $(-1, 1)$.

By looking at a few more cases, we see that there are three types of equivalence classes:

- Classes with positive numbers: $[1, 2), [2, 3), [3, 4), [4, 5), \ldots$.
- Classes with negative numbers: $\ldots, (-4, -3], (-3, -2], (-2, -1]$.
- One class with both positive and negative numbers: $(-1, 1)$.

The following diagram visualizes the equivalence classes on the number line.



(b) On the set of integers, we define the following equivalence relation:

$$a \equiv b \qquad \text{if and only if} \qquad a - b \text{ is divisible by 4.}$$

For instance, $15 \equiv 7$, as $15 - 7 = 8$ is divisible by 4, but $1 \not\equiv 8$, as $1 - 8 = -7$ is not divisible by 4. In words, we may describe this relation by saying that two integers are equivalent, if they have the same remainder when divided by 4 .

What is [7] (that is, the equivalence class of the number 7)? By definition, we have

$$[7] = \{b \in \mathbb{Z} : \ 7 - b \text{ is divisible by } 4\}.$$

Which integers b satisfy the condition that $7 - b$ is divisible by 4? There are many, of course. If $b = 3$, then $7 - 3 = 4$ is divisible by 4. Also, if $b = 7$ or $b = -1$, then $7 - b$ is divisible by 4. Looking at a few more examples, we easily conclude that

$$[7] = \{\dots, -13, -9, -5, -1, 3, 7, 11, 15, 19, \dots\}$$

(these are all the numbers which have a remainder of 3 when divided by 4). Note that the equivalence class of 3 or -9 (or any number which is equivalent to 7) will be identical to [7]. That is,

$$[7] = [3] = [-9].$$

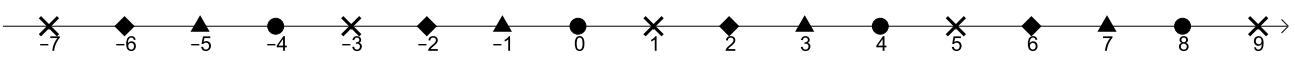
What about the equivalence class of 4? Well, the number $4 - b$ (for an integer b) is divisible by 4 if and only if b itself is divisible by 4, and so the equivalence class of 4 is simply the set of all multiples of 4:

$$[4] = \{\dots, -12, -8, -4, 0, 4, 8, 12, 16, 20, \dots\}.$$

Having covered all multiples of 4, and all numbers leaving a remainder of 3 when divided by 4, we realize that there are two more equivalence classes that have not been mentioned: The ones containing integers with remainder 1, and with remainder 2 (when divided by 4):

$$[1] = \{\dots, -11, -7, -3, 1, 5, 9, 13, 17, \dots\}$$
$$[2] = \{\dots, -10, -6, -2, 2, 6, 10, 14, 18, \dots\}$$

Overall, there are **four** equivalence classes for this relation. To visualize them, we use four different symbols (•, ×, ♦ and ▲) to mark integers belonging to the same equivalence class.



Looking at the two examples above, we make two important observations:

- The equivalence classes **cover** our original set. In the first example, the equivalence classes are intervals, covering all the real numbers. In the second example, the four equivalence classes (or more precisely, their union) contain all the integer numbers.
- In both examples, the equivalence classes **do not overlap**. Namely, the intersection of any two different classes is empty.

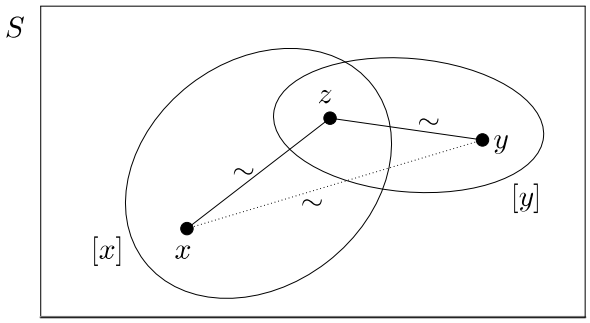
These two observations are completely general, and apply to any equivalence relation.

Theorem 7.3.2. *Let $\sim$ be an equivalence relation on a set S.*
Then any $x \in S$ belongs to some equivalence class, and any two different equivalence classes are disjoint.

Proof. Given an element $x \in S$, we have $x \sim x$ (as an equivalence relation is reflexive), and so $x \in [x]$. This shows that each element in x belongs to some equivalence class (namely, its own class).

The second part of the theorem claims that any two different equivalence classes are disjoint. We prove the contrapositive. That is, we show that two equivalence classes which are **not disjoint**, must be **equal to each other**. Let $[x]$ and $[y]$ be two non-disjoint equivalence classes (i.e., $[x] \cap [y] \neq \phi$). As their intersection is not empty, there is an element $z \in [x] \cap [y]$.

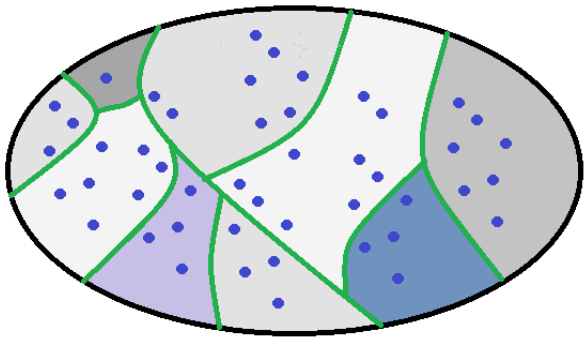
From the fact that $z \in [x]$ and $z \in [y]$, we conclude that $x \sim z$ and $y \sim z$. Symmetry of the relation implies that $z \sim y$, and transitivity implies that $x \sim y$ (see diagram).



We will now prove that $[x] = [y]$ by mutual inclusion (after all, $[x]$ and $[y]$ are **sets**). If $u \in [y]$, then $y \sim u$. But we also know that $x \sim y$. From transitivity of the relation, we conclude that $x \sim u$, and hence $u \in [x]$ (can you construct a diagram representing the argument in the last three sentences?). We have thus proved the inclusion $[x] \supseteq [y]$. The inclusion $[x] \subseteq [y]$ is proved in a similar way, from which we conclude that $[x] = [y]$, as needed. □

Note how the three defining properties of equivalence relations (reflexivity, symmetry and transitivity) are used in the proof. Without them, the theorem will not be valid.

Another way to phrase the theorem, is to say that an equivalence relation **induces a partition of** S **into equivalence classes** (i.e., it splits S into non-overlapping regions). Elements in the same class (or region) represent the same abstract entity.



7.4 Congruence Modulo n

In this section, we focus on one fundamental and widely used equivalence relation: Congruence modulo n. It is frequently used within mathematics (in number theory and discrete mathematics, for instance), and has numerous applications in related fields, such as computer science and cryptography. In fact, the relation in Example (b) on page 168 is a special case of congruence. Here is the definition.

Definition 7.4.1. Let $n \in \mathbb{N}$. The relation, on the set of **integers**, defined by:

$$a \equiv b \pmod{n} \qquad \text{if and only if} \qquad a - b \text{ is divisible by } n,$$

is called **congruence modulo** n.
If $a, b \in \mathbb{Z}$ satisfy this relation, we say that they are **congruent modulo** n.

Examples.

- $7 \equiv 10 \pmod{3}$, since 3 divides $7 - 10$ (which we can also write as $3|7 - 10$).
- $35 \equiv 15 \pmod{5}$, since $5|35 - 15$. We say that 35 is **congruent to** 15 **modulo** 5.
- $4 \equiv 4 \pmod{7}$, since 7 divides $4 - 4 = 0$.
- The number 3 is **not** congruent to 8 modulo 2, as $3 - 8$ is **not** divisible by 2. We denote this by writing $3 \not\equiv 8 \pmod{2}$.

It is not hard to see that if integers a and b are congruent modulo n, then they produce the same remainder when divided by n (prove this!).

We now proceed by showing that congruence is an equivalence relation.

Theorem 7.4.2. *For any $n \in \mathbb{N}$, congruence modulo n is an equivalence relation on $\mathbb{Z}$.*

Proof. To prove the theorem, we need to show that congruence modulo n is a reflexive, symmetric and transitive relation on integers.

- **(Reflexivity)** For any $k \in \mathbb{Z}$, $k \equiv k \pmod{n}$, as $k - k = 0$ is divisible by any positive integer. Therefore, this relation is reflexive.
- **(Symmetry)** If $a \equiv b \pmod{n}$, then $a-b$ is divisible by n. But then $b-a = -(a-b)$ is also divisible by n, which implies that $b \equiv a \pmod{n}$, as needed.
- **(Transitivity)** Suppose that $a \equiv b \pmod{n}$ and $b \equiv c \pmod{n}$. By definition, we get that **both** $a - b$ and $b - c$ are divisible by n. Consequently, their sum, $(a - b) + (b - c) = a - c$ is also divisible by n, which shows that $a \equiv c \pmod{n}$, as needed.

□

Other than being an equivalence relation, congruence has other useful properties. The following claim shows that congruence behaves well with respect to addition and multiplication of integers.

Claim 7.4.3. Let $n \in \mathbb{N}$ and $a, b, r, s \in \mathbb{Z}$.
If $a \equiv r \pmod{n}$ and $b \equiv s \pmod{n}$, then $a + b \equiv r + s \pmod{n}$ and $a \cdot b \equiv r \cdot s \pmod{n}$.

Proof. We are given that $a \equiv r \pmod{n}$ and $b \equiv s \pmod{n}$, which means that both $a - r$ and $b - s$ are divisible by n (or, in other words, are multiples of n):

$$a - r = n \cdot m \qquad \text{and} \qquad b - s = n \cdot k \qquad \text{for some } \ m, k \in \mathbb{Z}.$$

By adding these two equalities, we get

$$(a - r) + (b - s) = n \cdot m + n \cdot k \qquad \Rightarrow \qquad (a + b) - (r + s) = n \cdot (m + k).$$

We see that $(a + b) - (r + s)$ is divisible by n, and hence $a + b \equiv r + s \pmod{\ n}$, which proves the first part of the claim.

To prove the multiplication part, we rewrite the equalities $a - r = n \cdot m$ and $b - s = n \cdot k$ as

$$a = r + n \cdot m \qquad \text{and} \qquad b = s + n \cdot k,$$

and then multiply them:

$$a \cdot b = (r + n \cdot m) \cdot (s + n \cdot k) \qquad \Rightarrow \qquad a \cdot b = r \cdot s + r \cdot n \cdot k + n \cdot m \cdot s + n^2 \cdot m \cdot k.$$

Rearranging the last equality, we get

$$a \cdot b - r \cdot s = n \cdot (r \cdot k + m \cdot s + n \cdot m \cdot k),$$

which shows that $a \cdot b - r \cdot s$ is divisible by n, and hence $a \cdot b \equiv r \cdot s \pmod{n}$, as needed. □

By applying the multiplication part of the claim repeatedly, with $a = b$, we obtain the following conclusion.

Conclusion 7.4.4. Let $a, r \in \mathbb{Z}$ and $n \in \mathbb{N}$.
If $a \equiv r \pmod{n}$ and $k \in \mathbb{N}$, then $a^k \equiv r^k \pmod{n}$.

Here are a few examples, demonstrating the use of this equivalence relation, and Claim 7.4.3.

Examples.

(a) What is the unit (i.e., the rightmost) digit of the number 7^{20} ?

Answer: Given any natural number, the unit (or rightmost) digit equals the remainder obtained when dividing the number by 10 (assuming the number is written in the usual base 10 form). For instance, when the number 14523 is divided by 10, the quotient is 1452 and the remainder is 3, since

$$14523 = 1452 \cdot 10 + 3.$$

To find the unit digit of 7^{20}, we therefore use the **congruence modulo 10** equivalence relation, as follows.

	$7 \equiv -3 \pmod{10}$	(as $7 - (-3)$ is divisible by 10)
$\Rightarrow$	$7^4 \equiv (-3)^4 \equiv 81 \equiv 1 \pmod{10}$	(by Conclusion 7.4.4)
$\Rightarrow$	$(7^4)^5 \equiv 1^5 \pmod{10}$	(Conclusion 7.4.4 as well)
$\Rightarrow$	$7^{20} \equiv 1 \pmod{10}$	

Therefore, the remainder obtained when 7^{20} is divided by 10 is 1, and hence 1 is the unit digit of 7^{20}.
(**Remark:** A chain of equivalences of the form

$$A \equiv B \equiv C \equiv \cdots \equiv D \pmod{n}$$

indicates that any pair of terms from the chain are congruent modulo n.)

(b) What is the remainder obtained when the product $7697 \cdot 9154$ is divided by 5?

Answer: An integer is divisible by 5 if and only if it ends with either 5 or 0. Therefore, we have:

$$7697 \equiv 2 \pmod 5 \qquad \text{(as } 7697 - 2 \text{ is divisible by 5)}$$
$$9154 \equiv 4 \pmod 5 \qquad \text{(as } 9154 - 4 \text{ is divisible by 5)}$$

By Claim 7.4.3, we get:

$$7697 \cdot 9154 \equiv 2 \cdot 4 \equiv 8 \equiv 3 \pmod 5$$

Consequently, the required remainder is 3 (note that the remainder must be either $0, 1, 2, 3$ or 4).

(c) Let $a \in \mathbb{Z}$. Show that if $a^2 + 5$ is **not** divisible by 7, then $a - 3$ is also **not** divisible by 7.

Solution: We apply Claim 7.4.3 repeatedly, to prove the contrapositive. Namely, we show that if $a - 3$ is divisible by 7, then so is $a^2 + 5$ (try to justify each of the following steps).

$$\begin{aligned} & a - 3 \equiv \pmod 7 \\ \Rightarrow \quad & a \equiv 3 \pmod 7 \\ \Rightarrow \quad & a^2 \equiv 9 \equiv 2 \pmod 7 \\ \Rightarrow \quad & a^2 + 5 \equiv 7 \equiv 0 \pmod 7 \end{aligned}$$

From which it follows that $a^2 + 5$ is divisible by 7, as needed.

7.5 Exercises for Chapter 7

7.5.1. (a) Prove directly, that the following relation, on the set of integers, is an equivalence relation.

$$a \equiv b \qquad \text{if and only if} \qquad a - b \text{ is divisible by 4.}$$

(b) Show that if two integers satisfy the relation in part (a), then they have the same remainder when divided by 4 (refer to Theorem 6.1.2 and Exercise 6.4.7).

7.5.2. Define, on the set of integers, the following equivalence relation.

$$k \sim \ell \qquad \text{if and only if} \qquad |k| = |\ell|.$$

(a) Prove that the above relation is indeed an equivalence relation.

(b) Describe the equivalence classes for this relation.

7.5.3. For each of the following relations R on **the set of real numbers**, decide whether it is reflexive, symmetric and/or transitive. Justify your arguments. Is the relation an equivalence relation? Explain.

(a) $(x, y) \in R$ if and only if $|x - y| \leq 3$.

(b) $(x, y) \in R$ if and only if $x \cdot y > 0$.

(c) $(x, y) \in R$ if and only if $x^2 - y = y^2 - x$.

(d) $(x, y) \in R$ if and only if $(x - y)(x^2 + y^2 - 1) = 0$.

(e) $(x, y) \in R$ if and only if $|x + y| = |x| + |y|$.

7.5.4. Let $f\colon A \to B$ be an arbitrary function. Prove that the relation

$$x \sim y \quad \text{if and only if} \quad f(x) = f(y),$$

on the set A, is an equivalence relation.

7.5.5. Define a relation R on $\mathbb{Z}$ as follows: $(m, n) \in R$ if and only if $m + n$ is odd.
Is the relation reflexive? symmetric? transitive? Is it an equivalence relation? Explain.

7.5.6. Define a relation R on $\mathbb{Z}$ as follows: $(m, n) \in R$ if and only if $m + n$ is divisible by 3.
Is the relation reflexive? symmetric? transitive? Is it an equivalence relation? Explain.

7.5.7. Define a relation R on the set $\{2, 3, 4, \dots\}$, as follows:

$$(x, y) \in R \quad \text{if and only if} \quad x \text{ and } y \text{ have a common factor greater than 1.}$$

Is this relation reflexive? symmetric? transitive?
Is this an equivalence relation? Justify your arguments.

7.5.8. Is the following relation on $\mathbb{Z}$ reflexive? symmetric? transitive? Is it an equivalence relation? Prove your arguments.

$$a \simeq b \quad \text{if and only if} \quad 3a + 5b \text{ is divisible by 8.}$$

7.5.9. Define a relation R on $\mathbb{R} \setminus \{0\}$ as follows: $(x, y) \in R$ if and only if $x + \dfrac{1}{y} = y + \dfrac{1}{x}$.
Is this relation reflexive? symmetric? transitive? Is it an equivalence relation? Explain.

7.5.10. Let $\simeq$ be a relation on $\mathbb{Z}$ defined as follows:

$$a \simeq b \quad \text{if and only if} \quad 2a + 3b \text{ is divisible by 5.}$$

(a) Show that $\simeq$ is an equivalence relation.

(b) What is the equivalence class of 0 ?

7.5.11. Define a relation on $\mathbb{Q} \setminus \{0\}$ as follows:

$$x \sim y \qquad \Leftrightarrow \qquad \frac{x}{y} = 2^k \quad \text{for some} \quad k \in \mathbb{Z} \quad .$$

Prove that this is an equivalence relation.

7.5.12. Consider the following equivalence relation on $\mathbb{R} \setminus \{0\}$:

$$a \sim b \quad \text{if and only if} \quad \frac{a}{b} \in \mathbb{Q} \qquad .$$

In the following list of equivalence classes, find two classes which are **equal**. Explain your choice briefly.

- $[\sqrt{3}]$
- $[1]$
- $[\sqrt{12}]$
- $[\sqrt{6}]$

7.5.13. On the set $\mathbb{N} \times \mathbb{N}$, define the following relation:

$$(a\,,b) \sim (c\,,d) \quad \text{if and only if} \quad a + d = b + c.$$

(a) Show that this is an equivalence relation.

(b) Describe the **equivalence class** of $(1\,,1)$.

7.5.14. Consider the equivalence relation **congruence mod 5**, on the set of integers.

(a) Describe the equivalence class of 33.

(b) How many difference equivalence classes are there for this relation?

(c) Which $a \in \mathbb{Z}$ satisfy the condition $[a] = [17]$? Explain.

7.5.15. For each statement, decide whether it is **true** or **false**. Justify your answer briefly.

(a) For any $n \in \mathbb{N}$ and $k \in \mathbb{Z}$, we have $k \equiv k \pmod{n}$.

(b) For any $n \in \mathbb{N}$ and $k \in \mathbb{Z}$, we have $k \equiv n \pmod{n}$.

(c) For any $a, b \in \mathbb{Z}$, we have $a \equiv b \pmod{1}$.

(d) For any $m \in \mathbb{Z}$, we have $(m+3)^2 \equiv m^2 + 3^2 \pmod{6}$.

(e) For any $m \in \mathbb{Z}$, we have $(m+3)^2 \equiv m^2 + 3^2 \pmod{9}$.

(f) For any $a, b \in \mathbb{Z}$ and $n \in \mathbb{N}$, if $a^2 \equiv b^2 \pmod{n}$, then $a \equiv b \pmod{n}$.

7.5.16. The following statement if **false**:

"If a relation is both symmetric and transitive, then it is also reflexive."

Find the mistake is the following false proof:

Suppose the R is a relation on a set A, which is both symmetric and transitive, and let $x \in A$. Choose an element $y \in A$, for which $(x, y) \in R$. By symmetry, we also have $(y, x) \in R$, and since the relation R is transitive, we conclude from $(x, y) \in R$ and $(y, x) \in R$, that $(x, x) \in R$. Therefore, the relation R is also reflexive.