PAOLO ALUFFI

MODICUM MATHEMATICUM

A SWATH THROUGH THE BASIC LANGUAGE OF ABSTRACT MATH

do not distribute

Contents

0.1 *Introduction* vii

1 *Basic notation, logic, and proofs* 1

1.1 *Basic notation, quantifiers* 1

1.1.1 *Sets: basic notation* 1

1.1.2 *Famous sets* 3

1.1.3 *Indexed sets* 4

1.1.4 *Quantifiers* 4

1.1.5 *Examples* 5

1.1.6 *Homework* 7

1.2 *Logic* 8

1.2.1 *Logical connectives* 8

1.2.2 *Truth tables* 9

1.2.3 *The connective* $\Longrightarrow$ 11

1.2.4 *Caveat* 13

1.2.5 *How many connectives are really necessary?* 14

1.2.6 *Homework* 15

1.3 *More about logic. Techniques of proofs.* 16

1.3.1 *The behavior of quantifiers under negation.* 16

1.3.2 *Do examples prove theorems? Do counterexamples disprove them?* 17

1.3.3 *Techniques of proof* 19

1.3.4 *Homework* 22

do not distribute

2 *Sets and relations* 25
2.1 *Naive set theory.* 25
2.1.1 *Sets: Basic definitions* 25
2.1.2 *Sets: Basic operations* 27
2.1.3 *Properties of ∩, ∪, ∖* 28
2.1.4 *The power set* 30
2.1.5 *The cartesian product* 31
2.1.6 *Homework* 33
2.2 *Relations and induction* 34
2.2.1 *Definition of 'relation'* 34
2.2.2 *Order relations* 36
2.2.3 *The well-ordering principle and induction* 37
2.2.4 *Two examples* 39
2.2.5 *Homework* 44
2.3 *Equivalence relations and partitions.* 46
2.3.1 *Equivalence relations* 46
2.3.2 *Definition of equivalence classes* 47
2.3.3 *Partitions and equivalence relations* 49
2.3.4 *Homework* 51
2.4 *Quotients modulo an equivalence relation.* 52
2.4.1 *Quotient of a set by an equivalence relation* 52
2.4.2 *One more example: ℚ as a quotient* 55
2.4.3 *Homework* 56

3 *Functions and canonical decomposition* 59
3.1 *Functions.* 59
3.1.1 *Functions: basic definitions* 59
3.1.2 *Examples (and more definitions)* 61
3.1.3 *Homework* 64
3.2 *Functions: Composition, inverses, isomorphisms.* 65
3.2.1 *Functions, again* 65
3.2.2 *Composition of functions* 66
3.2.3 *Inverses* 67
3.2.4 *Homework* 69

do not distribute

3.3 *Injectivity and surjectivity.* 69
3.3.1 *Injections and surjections* 69
3.3.2 *Injective vs. left inverses, surjective vs. right inverses* 71
3.3.3 *Proof of Theorem 3.3.7* 72
3.3.4 *Homework* 73
3.4 *Isomorphisms of sets, a.k.a. bijections* 74
3.4.1 *Isomorphisms of sets* 74
3.4.2 *Examples of isomorphisms* 75
3.4.3 *The set of bijections from a set to itself* 78
3.4.4 *Homework* 80
3.5 *Canonical decomposition* 82
3.5.1 *Canonical decomposition* 82
3.5.2 *Applications* 86
3.5.3 *Homework* 88

4 *Constructing* $\mathbb{R}$ 91
4.1 *Incompleteness of* $\mathbb{Q}$. 91
4.1.1 *Real numbers: general considerations* 91
4.1.2 $\sqrt{2}$ *is not rational* 92
4.1.3 *Completeness* 93
4.1.4 *Homework* 95
4.2 *Dedekind cuts* 96
4.2.1 *Dedekind cuts: definition* 96
4.2.2 $\mathbb{Q} \subsetneq \mathbb{D}$ 97
4.2.3 *Order relation on* $\mathbb{D}$ 97
4.2.4 *Least upper bounds in* $\mathbb{D}$ 98
4.2.5 *Homework* 99
4.3 *Cauchy sequences* 100
4.3.1 *Decimal expansions and Cauchy sequences* 100
4.3.2 $\mathbb{Q} \subsetneq \mathbb{S}$ 102
4.3.3 *Order relation on* $\mathbb{S}$ 102
4.3.4 *Least upper bounds in* $\mathbb{S}$ 105
4.3.5 *Homework* 106

do not distribute

4.4 *Decimal expansions, again* 107
4.4.1 *Back to decimal expansions* 107
4.4.2 *Decimal expansions of rational/irrational numbers* 110
4.4.3 *Homework* 111

5 *Cardinality* 113
5.1 *What is a 'number'?* 113
5.1.1 *The concept of 'number'* 113
5.1.2 *A simple-minded lemma on bijections* 114
5.1.3 *Another (not so simple-minded) lemma on bijections* 116
5.1.4 *Homework* 118
5.2 *Shades of infinity* 118
5.2.1 *Comparing cardinalities* 118
5.2.2 *The 'diagonal' argument* 119
5.2.3 *Countable and uncountable sets* 121
5.2.4 *Cardinality of $\mathbb{Z}$ and $\mathbb{Q}$* 123
5.2.5 *Homework* 125
5.3 *Je le vois, mais je ne le crois pas!* 126
5.3.1 *Cardinality and set-theoretic operations* 126
5.3.2 *Algebraic and transcendental numbers* 127
5.3.3 *The power of the continuum* 128
5.3.4 $\mathbb{R} \cong \mathbb{R} \times \mathbb{R}$ 129
5.3.5 *Homework* 130
5.4 *Cantor-Bernstein-Schröder* 130
5.4.1 *The Cantor-Bernstein-Schröder theorem* 130
5.4.2 *Proof of the Cantor-Bernstein-Schröder theorem* 131
5.4.3 *An example* 133
5.4.4 *Homework* 134

6 *Very basic topology* 135
6.1 *Continuity, XIX-century version.* 135
6.1.1 *Expressions depending on quantifiers, again* 135
6.1.2 *Limits and continuity, calculus-style* 136
6.1.3 *Homework* 139

do not distribute

6.2 *Continuity, XX-century version.* 139
6.2.1 *Topological spaces* 139
6.2.2 *Homeomorphisms* 141
6.2.3 *Continuous functions* 142
6.2.4 *Homework* 143
6.3 *Continuity for functions* $\mathbb{R} \to \mathbb{R}$ 144
6.3.1 *The standard topology on* $\mathbb{R}$ 144
6.3.2 *Continuity, reprise* 146
6.3.3 *Homework* 148
6.4 *Topological invariants.* 148
6.4.1 *Induced topology* 148
6.4.2 *Features that are invariant by homeomorphisms* 150
6.4.3 *Dessert: the Euler characteristic* 152
6.4.4 *Homework* 153

Appendix — Solutions to selected problems 155

Index 193

do not distribute

do not distribute

0.1 Introduction

In the standard undergraduate curriculum at US universities, lower-level mathematics courses such as Calculus or Linear Algebra do not expose students to the abstract language of mathematics or to rigorous proofs. Perhaps unfortunately, a student may go through these basic subjects without ever encountering an abstract 'set' or the need to produce (or even seriously understand) logical arguments. These students typically find the impact of proof-based material in upper-level courses very challenging: how does one actually go about 'proving' something, or even communicating in an intelligible way the result of a mathematical investigation? While some students have an intuitive understanding of these matters, and can adapt quickly to the different mindset required in these more advanced courses, many others find the going exceedingly difficult and face a seemingly unreasonable obstacle in these 'proof-based' classes.

The standard way to address this problem is to require students to take a 'bridge' course whose main content is the discussion of essential notions in naive set theory, such as basic terminology, set operations, order and equivalence relations, etc. The discussion of these notions also creates the opportunity for a gentle introduction to proof-based inquiry, helping the students transition from the recipe-oriented style of lower-level courses to the standard expected in courses on abstract algebra, analysis, topology.

These notes were written for this purpose, and the author has used them numerous times when teaching the bridge course at Florida State University. A glance at the table of contents will suffice to gauge their scope. A discussion of naive set theory culminates in what is called here the 'canonical decomposition', that is, the decomposition of an arbitrary set-function as a surjective function, followed by a bijective function, followed by an injective function. This fact is easy to prove and appreciate at the level of sets, provided that the student has really acquired important notions such as equivalence relations and partitions. It also reproduces in this simpler context the essence of various manifestations of the 'first isomorphism theorem' in elementary abstract algebra. This material therefore connects these notes directly to abstract algebra: a student mastering this toy model of the first isomorphism theorem should find the corresponding statements in group or ring theory straightforward when the time comes.

The notes include a small detour into constructions of the set of real numbers, by Dedekind cuts and by equivalence classes of Cauchy sequences. This material is essentially optional, but the author feels that it is helpful for the students at this stage to be aware that 'real numbers' involve a level of abstraction beyond our familiarity with

do not distribute

integers or rational numbers. Dealing with Cauchy sequences requires a comfort level with comparatively complex arguments that may exceed what is expected of the intended audience; instructors may well decide that the leaner discussion of Dedekind cuts suffices, or even that the more general considerations about completeness given in §4.1 and §4.4 already carry the relevant message.

Similar considerations may apply to the discussion of cardinality which follows, leading to a proof of the Cantor-Bernstein-Schröder theorem. Some instructors may well feel that this material is *not* necessary for a bridge course; the author feels that an early awareness that there are many levels of infinity is useful, and this is a good opportunity for students to acquire this awareness.

In any case, these topics reflect the author's taste, and illustrate well the fact that the level of abstraction gained in the first part of the course can lead to nontrivial results. Also reflecting the author's taste is the choice to cover the absolute basics of point-set topology, including a proof that 'continuity' in the abstract topological setting reproduces the metric notion, with which the students are familiar to some extent from their exposure to calculus. This serves both as a further opportunity to exercise the use of the abstract language acquired in the first part and as a way to connect this course to topology and advanced calculus, again facilitating the students' future exposure to these subjects.

The notes include many exercises. They are typically straightforward, and it would be fair to expect that students complete *all* of them in a run of the course. Instructors are advised to spend a substantial amount of class time discussing them. To facilitate such discussions, and to help a reader going through these notes outside of a classroom environment, many solutions are included in an appendix. (Problems whose solution is included are marked with a ▷ sign.) Students are encouraged to try their best at solving the problems and *then* compare their work with the suggested solutions. Dear reader: If you do this the other way around—first look at the solution, then try the problem—then you will miss the most straightforward opportunity the author is giving you to learn this material.

Thanks are due to the many students who have taken courses based on these notes, for their comments and their patience. Somehow, the author also feels that he should apologize to his logician and set-theorist friends for the sloppiness and imprecision with which many important issues are treated in these notes. It is clear (and comments to this effect are frequently included in the text) that a reader seriously intending to study logic or set theory should look elsewhere. Hopefully these notes may at least serve to whet some readers' appetite for more.

Tallahassee, Florida, and Toronto, Ontario, 2020

do not distribute

do not distribute

do not distribute

1
Basic notation, logic, and proofs

'Mathematics is ridden through and through with the pernicious idioms of set theory.'
(L. Wittgenstein, Philosophische Bermerkungen)

1.1 Basic notation, quantifiers

1.1.1 Sets: basic notation

In these notes, *Sets* will simply be 'collections of elements'. Dear reader: please realize that this is not a precise definition; for example, we are not saying what an 'element' should be. Our goal is not to develop set theory, which is a deep and subtle mathematical field. Our goal is simply to become familiar with several standard operations and with a certain language, and practice the skill of using this language properly and unambiguously. What we are dealing with is often called *naive* set theory, and for our purposes we can rely on the intuitive understanding of what a 'collection' is.

The 'elements' of our sets will be anything whatsoever: they do not need to be numbers, or particular mathematical concepts. In other contexts, for example in calculus, you may have mostly (secretly) dealt with sets ultimately having to do with *real numbers.* In these notes, we are taking a more noncommittal standpoint.

Our mock definition of 'set' amounts to the (important!) claim that if A and B are sets, then

$$A = B \text{ if and only if } A \text{ and } B \text{ consist of the same elements.}$$

Here is how one usually specifies a set:

$$S = \{\underbrace{\cdots\cdots\cdots}_{\text{(the type of elements we are talking about)}} \underbrace{\ |\ }_{\text{"such that"}} \underbrace{\cdots\cdots\cdots}_{\text{(some property defining the elements of } S\text{)}}\}$$

For example, we can write:

$$S = \{n \text{ even integer} \,|\, 0 \leq n \leq 8\}$$

to mean: *S is the set of all even integers between* 0 *and* 8*, inclusive.* This sentence is the English equivalent of the 'formula' written above.

The | in the middle of the notation may also be written as a colon, ":", or explicitly as "such that". On the other hand, there is no leeway in the use of parentheses:

$$S = (n \text{ even integer} \,|\, 0 \leq n \leq 8)$$

or

$$S = [n \text{ even integer} \,|\, 0 \leq n \leq 8]$$

are simply not used in standard notation, and you should avoid improvising such alternative ways of denoting sets, if you care about being understood by the mathematical community. Doing away with the parentheses,

$$S = n \text{ even integer} \,|\, 0 \leq n \leq 8$$

do not distribute

is even worse, and hopefully you see why: by the time you have read '$S = n$' you should think that S and n are the same thing, while they are not supposed to be.

This type of attention to detail may seem excessive, but it is not—we must simply agree on the basic orthography of mathematics, and adhere to it just as we adhere to the basic orthography of English. Yu kan prbbly mk yself undrstud evenf yo mngle evry sinl wrd af o sentnc, but it is usually not acceptable to put such burden on the reader, is it?

An alternative way to denote a set is by listing all its elements, again within {—}:

$$S = \{0, 2, 4, 6, 8\}$$

again denotes the set of even integers between 0 and 8, inclusive. Since two sets are equal precisely when they have the same elements, we can write

$$\{n \text{ even integer} \mid 0 \leq n \leq 8\} = \{0, 2, 4, 6, 8\} \quad .$$

This is a true (if mind-boggingly simple) mathematical statement, that is, a *theorem.*

The 'order' in which the elements are listed in a set is immaterial:

$$\{0, 2, 4, 6, 8\} = \{2, 0, 8, 4, 6\}$$

since these two sets consists of the same elements. This is another reason why we have to be careful with the use of parentheses:

$$(0, 2, 4, 6, 8)$$

means something, and that something is *not* the set $\{0, 2, 4, 6, 8\}$. It is the *ordered* set of even integers between 0 and 8, for which the *order* in which the elements are listed is part of the information conveyed by the notation. Thus,

$$(0, 2, 4, 6, 8) \neq (2, 0, 8, 4, 6) \quad :$$

the elements are the same, but the orders differ.

Repetitions are also immaterial:

$$\{0, 0, 0, 0, 2, 2, 6, 8, 4, 4, 6, 2, 0, 6, 8\} = \{0, 2, 4, 6, 8\} \quad .$$

Denoting a set by listing its elements (as in {0,2,4,6,8}) runs into an obvious problem if the set we want to specify is *infinite.* This is not a problem in the first style of notation reviewed above. For example,

$$\{n \text{ even integer} \mid n \geq 0\}$$

is a fine way to denote the set of nonnegative even integers. The second style may often be adopted for such sets, by using $\cdots$ to imply a pattern that is hopefully apparent from the rest: most people would recognize

$$\{0, 2, 4, 6, 8, \cdots\}$$

as a good description of the same set of nonnegative even integers. My personal preference is to *not* use this alternative if possible, because it makes the reader's life a little harder, and it is inevitably ambiguous. Does

$$\{3, 5, 7, \ldots\}$$

denote the set of odd integers strictly larger than 1, or does it denote the set of positive odd prime integers? We are not given enough information to decide. It is an excellent idea to avoid any such ambiguity. In practice,

do not distribute

this means that the 'first style' of notation is more useful: sets are defined precisely by the properties satisfied by their elements[1]. Our first goal in this course is to discuss ways to manipulate 'logically' such properties.

1.1.2 *Famous sets*

The sentence *"x is an element of the set S"* is abbreviated by the shorthand

$$x \in S \quad .$$

(The symbol $\in$ evolved from the Greek letter 'epsilon'.) To denote that x is *not* an element of S, one may write $x \notin S$. Thus,

$$1 \notin \{0,2,4,6,8\} \quad , \quad 4 \in \{0,2,4,6,8\} \quad .$$

Many famous sets are denoted by standard symbols. The *empty set* is the set containing no elements whatsoever, and it is denoted by $\emptyset$. Note that the two sets

$$\emptyset \quad , \quad \{\emptyset\}$$

are very different! The first is the empty set, and has no elements. The second does contain one element, and that element happens to be the empty set.

Here are a few 'number-based' examples of famous sets:

- $\mathbb{N}$: the set of all 'natural numbers', that is, nonnegative integers;
- $\mathbb{Z}$: the set of all integers;
- $\mathbb{Q}$: the set of rational numbers, that is, quotients of integers;
- $\mathbb{R}$: the set of real numbers;
- $\mathbb{C}$: the set of complex numbers.

Defining some of these sets precisely is actually not an easy task. Taking $\mathbb{Z}$ for granted (as we will do in the first several chapters), we can define[2] as

$$\mathbb{N} = \{n \in \mathbb{Z} \,|\, n \geq 0\} = \{0,1,2,3,4,\dots\} \quad .$$

However, defining $\mathbb{Q}$ requires thinking in terms of 'equivalence classes of fractions'; defining $\mathbb{R}$ requires (for example) preliminary considerations on the lack of completeness of $\mathbb{Q}$. Luckily, the expected readership of these notes has already developed a working knowledge of these sets, and at least for now I will rely on this familiarity. One of my goals in this notes is to help the reader form a more accurate understanding of these notions; but that's for later chapters. Your current knowledge will already allow you to parse something like

$$S = \{x \in \mathbb{R} \,|\, x > 25\}$$

without difficulties: translating into English, this reads *"S is the set consisting of all real numbers strictly greater than* 25*"*. Is it true that $27 \in S$? Yes. Is it true that $25 \in S$? No.

[1] Actually Bertrand Russel's paradox shows that this statement cannot be taken too literally, but let's not worry about this for now.

[2] Later in this text we will define $\mathbb{N}$ from scratch, and use that to define $\mathbb{Z}$. But at this stage it is simpler to just rely on your familiarity with $\mathbb{Z}$.

do not distribute

It is also occasionally convenient to decorate the sets listed above with specifications limiting their range. For example,

$$\mathbb{R}^{\geq 0} := \{r \in \mathbb{R} \,|\, r \geq 0\}$$

may be used to denote the set of *nonnegative* real numbers. With this style of notation, $\mathbb{N}$ is another name for $\mathbb{Z}^{\geq 0}$. (And be warned of the fact that some people prefer to use $\mathbb{N}$ for $\mathbb{Z}^{>0}$. In this text, 0 is viewed as a natural number.)

1.1.3 *Indexed sets*

Sometime it is useful to 'label' the elements of a set by using another set: for example, we can write

$$\{s_n \,|\, n \in \mathbb{N}\} \quad \text{or also} \quad \{s_n\}_{n\in\mathbb{N}}$$

to mean that the elements of this set are called $s_0, s_1, s_2, \ldots$: one index for each natural number. You will recall dealing with *sequences* in calculus, and those are a particular case of this notion. For example, with $s_n = n^2$ you get the set

$$\{n^2 \,|\, n \in \mathbb{N}\} = \{0, 1, 4, 9, 16, 25, \ldots\}$$

of squares of integers. For $s_n = \frac{1}{n+1}$:

$$\left\{\frac{1}{n+1} \,|\, n \in \mathbb{N}\right\} = \left\{1, \frac{1}{2}, \frac{1}{3}, \frac{1}{4}, \frac{1}{5}, \ldots\right\} \quad .$$

An explicit recipe such as '$s_n = n^2$' is not needed in general. Also, keep in mind that our sets can be sets of any type of elements—not just numbers. In principle, even the 'indexing set' ($\mathbb{N}$ in these examples) does not need to consist of numbers.

In a sense, all we are doing here is choosing a particular way to call the elements in the set. But note that this operation is not completely innocent: for example, it allows for repetitions (what if $s_n = s_m$ for two distinct m and n?), and in the examples I just gave we seem to be able to keep track of the order in which the elements are listed. If you feel uneasy about calling these things 'sets', after what we agreed upon in §1.1.1, you are right: we shouldn't. These things are called *indexed* sets, and defining them more precisely than I have done here would require a good understanding of the notion of function between sets. We will get there in due time (Example 3.1.8).

1.1.4 *Quantifiers*

One of our first goals is to develop a language allowing us to manipulate 'logical expressions', for example in order to deal effectively with properties defining sets. One important issue is the scope to which a given statement applies. For example, we cannot decide on the truth of a statement such as $x > 5$ unless we are given some information on the 'range' of x, even after agreeing that x denotes (for example) an integer. The statement:

$$x > 5 \text{ for } \textit{some} \text{ integer } x$$

is *true:* for example, the integer 10 is indeed larger than 5. On the other hand, the statement

$$x > 5 \text{ for } \textit{every} \text{ integer } x$$

is *false:* for example, the integer 1 is not larger than 5. The distinction between 'some' and 'every' is expressed by employing a *quantifier:*

do not distribute

- $\exists$ stands for *'for some...'*
- $\forall$ stands for *'for every...'*

'There exists...' ($\exists$) and 'for all...' ($\forall$) are popular alternative translations. $\exists$ is the *existential* quantifier; $\forall$ is the *universal* quantifier.

It is common (and good practice) to start off a statement by specifying the appropriate quantifier or quantifiers. Thus, the two statements written above can be expressed more concisely by

$$\begin{aligned} \exists x \in \mathbb{Z} : &\quad x > 5 \\ \forall x \in \mathbb{Z} : &\quad x > 5 \quad . \end{aligned}$$

Again, the first statement is *true* and the second is *false.* This tells us that quantifiers make a difference: the statement $x > 5$ can be true and can be false, according to the quantifier governing x. Without a quantifier (maybe given implicitly, but preferably explicitly written out) we cannot tell whether it is one or the other.

One potentially confusing observation is that the order in which quantifiers appear in a statement also makes a difference. Compare the two statements

$$\begin{aligned} \forall x \, \exists y : &\quad y > x \\ \exists y \, \forall x : &\quad y > x \end{aligned}$$

where x, y are taken to be (for example) integers. The first statement says that no matter how you choose an integer x (that is: $\forall x$), there is an integer y (that is: $\exists y$) that beats it. This is *true:* for example, $y = x + 1$ would do the job. On the other hand, the second statement says that there is a magic integer y ($\exists y$) with the amazing property that it is larger than every integer x ($\forall x$). This is not true, there is no such integer! So the second statement is *false.*

The two statements only differ in the order in which the quantifiers are considered. Therefore, we have to pay attention to this order. If you write something like

$$\begin{array}{cc} & \forall x \\ y > x & \\ \exists y & \end{array}$$

then your reader (think: grader) will have no clue as to what you mean.

The order does not matter if the quantifiers are the same: the statements

$$\begin{aligned} \forall x \in \mathbb{Z} \, , \, \forall y \in \mathbb{Q} : &\quad x + y \in \mathbb{R} \\ \forall y \in \mathbb{Q} \, , \, \forall x \in \mathbb{Z} : &\quad x + y \in \mathbb{R} \end{aligned}$$

are equivalent, and true (and not very interesting).

1.1.5 Examples

At this point we have already acquired a bit of language that allows us to state concisely and precisely facts about sets and their elements. Statements as above, involving quantifiers, sets, and properties satisfied by the quantified variables, may be used to describe features of a set. As a very simple example, the statement

$$\forall x \in S : \quad x \in T$$

do not distribute

tells us that 'for every element x of S, x is an element of T'. In other words, the set T contains all the elements of the set S. This situation is denoted $S \subseteq T$, and we say that S is a *subset* of T. We will come back to this and other basic terminology concerning (naive) sets in Chapter 2.

It is a good idea to practice using this language, and you can do it all by yourself: just string a few quantified variables before a property involving those variables, in some comfortable environment such as $\mathbb{R}$, and figure out the 'meaning' of the statement you have written. To make sure that you write meaningful mathematics, it should suffice to spell out what each symbol means, and you should get a sensible, grammatically correct English sentence. For example, let S be a set whose elements are real numbers (notation: $S \subseteq \mathbb{R}$). The statement

$$\forall x \in \mathbb{R}\,,\, \exists y \in S : \quad y > x$$

reads out:

$\forall x \in \mathbb{R}$	$\exists y \in S$	:	$y > x$
for every real number x	there exists a y in S	such that	y is larger than x

This makes sense as a sentence in English, giving you confidence that the math is also grammatically correct[3]. Spend a moment thinking about what this statement says about the set S.

Ready?

It says that S contains 'arbitrarily large' real numbers: given any real number, you can find some element of S that beats it. This may or may not be true for a given set S. For example, it is true for the set $S = \mathbb{Z}$ of integers—given any real number, you can find a larger integer. It is not true for the set S of negative rational numbers: the real number 0 is larger than every negative rational number, so there are some real numbers that are not beaten by any element of S.

Try another one:

$$\forall x \in \mathbb{R}\,,\, \exists y \in S : \quad |y| \leq |x| \quad .$$

Translation: *For every real number x, there is a y in S whose absolute value is no larger than the absolute value of x.*

What does this say about S? Hmmm... If *for every* real number x there is a y in S such that $|y| \leq |x|$, then this must be the case for $x = 0$: there must be a y in S such that $|y| \leq |0| = 0$. On the other hand, $|y| \geq 0$ by definition of absolute value. Thus for $x = 0$ we would simultaneously have $|y| \leq 0$ and $|y| \geq 0$, and this says $y = 0$: we have discovered that $0 \in S$. Our statement is a complicated way of saying that 0 is an element of S.

What about

$$\forall x > 0\,,\, \exists y \in S : \quad |y| < x \quad ?$$

(Here I am understanding $x \in \mathbb{R}$: we are talking about real numbers in these examples.) This says: *'For all positive real numbers x, there exists a y in S whose absolute value is smaller than x.* Does this again say that 0 is necessarily an element of S? Think about it.

Answer: No. For example, the set S of positive rational numbers does not contain 0 and yet satisfies this condition, since there are arbitrarily small rational numbers (you can make $1/n$ as small as you please, by choosing n large enough). This new statement is saying that S contains 'arbitrarily small' nonnegative real numbers.

[3] Something like

$$\forall x\,, \exists y \in y > x$$

would read as 'for every x there exists a y in y is larger than x', which makes no sense in either English or math. Just reading your math aloud in English should allow you to tell if you are writing 'grammatically correct' formulas.

do not distribute

What about

$$\forall \epsilon > 0\,,\, \exists y \in S\,:\quad |y - L| < \epsilon \tag{1.1}$$

where L is some fixed real number? What is this telling us about S? In parsing (1.1), it may be helpful to keep in mind that the inequality $|y - L| < \epsilon$ is equivalent to $-\epsilon < y - L < \epsilon$, and therefore to $L - \epsilon < y < L + \epsilon$. Incidentally, the Greek letter ϵ is just a popular choice of notation for these expressions. You could use 'x' without changing the meaning of the expression in the least.

And what about

$$\forall \epsilon > 0\,,\, \exists N\,,\, \forall n\,:\quad n > N \implies |s_n| < \epsilon \tag{1.2}$$

where $\epsilon \in \mathbb{R}$, $n, N \in \mathbb{N}$, and $\{s_n\}_{n\in\mathbb{N}}$ is some indexed set of real numbers? The symbol $\implies$ stands for *if…, then…*, so that the condition expressed here is: *"if $n > N$, then $|s_n| < \epsilon$"*. We will look into this notation more carefully in the next section.

Do give this a try. We will actually come back to this very statement in Chapter 6, and it would be great if you had figured out 'what it means' before then. You will also have the opportunity to practice manipulating more complicated expressions of this type in §4.3.

1.1.6 Homework

A ▷ next to a problem indicates that a solution is included in the appendix.

▷ **1.1.1:** Assume that $P(x)$ is some statement whose truth depends on x. Which of the following is 'grammatically correct'?

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

▷ **1.1.2:** Let S be an indexed set of real numbers:

$$S = \{s_n \,|\, n \in \mathbb{N}\} \quad .$$

For example, we could have $s_n = n^2$, or $s_n = 1/(n+1)$, as in the examples in §1.1.3. Consider the statement

$$\forall x \in \mathbb{R}\,, \exists n \in \mathbb{N}\,:\quad s_n > x \quad .$$

- Find three examples of indexed sets S for which this statement is true, and three examples for which it is false.
- Find an adequate 'translation in English' of what this statement is saying[4].

1.1.3: Do the same, for the following statement (where implicitly $\epsilon \in \mathbb{R}$):

$$\forall \epsilon > 0\,,\, \exists n \in \mathbb{N}\,:\quad |s_n - 1| < \epsilon \quad .$$

1.1.4: The same, for the last statement shown in the text:

$$\forall \epsilon > 0\,,\, \exists N\,,\, \forall n\,:\quad n > N \implies |s_n| < \epsilon$$

where $\epsilon \in \mathbb{R}$, $n, N \in \mathbb{N}$, and again $\{s_n\}_{n\in\mathbb{N}}$ is some indexed set of real numbers.

[4] This is a rather open-ended request. Of course, *"For all real numbers x there exists a natural number n such that s_n is larger than x"* is a fine translation, but you should look for something more colloquial and yet precise. Try not to use symbols such as x or n in your answer.

do not distribute

1.2 Logic

1.2.1 Logical connectives

Sets are specified by spelling out 'properties' satisfied by their elements. My examples have been a little sloppy, actually: when I wrote

$$S = \{n \text{ even integer} \,|\, 0 \leq n \leq 8\}$$

you could have objected that I am already prescribing some property (that is, 'being even') in the declaration of the type of elements we are dealing with in this set. Once we have agreed that $\mathbb{Z}$ denotes the set of integers, maybe a preferable way to describe the same set would be

$$S = \{n \in \mathbb{Z} \,|\, n \text{ is even AND } 0 \leq n \leq 8\} \quad .$$

This would be better form: the first declaration simply states that n belongs to a set that has already been defined, and the second lists all the properties required of n in order to be an element of S.

This will serve as a motivation to understand how to combine different properties together by connectives such as 'AND'. I will use letters such as p, q, r, ... to denote statements such as 'n is even', '$0 \leq n \leq 8$', or anything that may be *true* (T) or *false* (F). I want to be able to combine such statements together to form more complex statements, keeping track of when these more complex statements are T or F.

'AND' is a good example, because its meaning is very transparent. If p denotes the property '$x > 0$' for a real number x, and q denotes '$x \leq 5$', then p AND q should express both requirements simultaneously: $0 < x \leq 5$. This new statement will be true if both statements p and q are true, while it will be false if one or both fail. The symbol $\wedge$ is used as a replacement for the conjunction AND in formulas. Thus,

$$p \wedge q \text{ is } \begin{cases} \text{T when both } p \text{ and } q \text{ are T} \\ \text{F when at least one of } p \text{ and } q \text{ is F} \end{cases} .$$

The basic negation 'NOT' is equally transparent: if p is a statement, then 'NOT p' should be the statement with opposite truth value as p: F when p is T, and T when p is F. In formulas, 'NOT p' is usually written $\neg p$. Thus,

$$\neg p \text{ is } \begin{cases} \text{T when } p \text{ is F} \\ \text{F when } p \text{ is T} \end{cases} .$$

What 'OR' should be is perhaps less clear. If I say, *I'll take one or the other,* I am excluding the possibility of taking both. On the other hand, if I say *I don't like the weather when it is windy or rainy,* I am certainly entertaining the possibility that it may be windy *and* rainy. The 'or' in the first sentence is not really the same as the 'or' in the second, because the first excludes that both alternatives apply, while the second does not exclude this possibility. In math, such ambiguities are forbidden: we should decide whether OR should be 'exclusive' or 'inclusive'[5]. The convention is that it is *in*clusive. So 'p OR q' will be true if p is T, if q is T, and if both p and q are T. It will be F precisely when both p and q are F. In formulas, we write $\vee$ for OR:

$$p \vee q \text{ is } \begin{cases} \text{T when at least one of } p \text{ and } q \text{ is T} \\ \text{F when both } p \text{ and } q \text{ are F} \end{cases} .$$

[5] Which is it, in this last sentence?

do not distribute

You should practice the use of these 'logical connectives' a little, and again this is something you can do by yourself: just invent a couple of statements, and contemplate the meaning of the statement obtained by ANDing or ORing them, perhaps after NOTting one or the other or both, and so on.

For example, if p is '$x > 0$', then $\neg p$ is '$x \leq 0$' (watch out: this is not the same as '$x < 0$'). If q is $x \leq 5$, then $p \wedge q$ is $0 < x \leq 5$. What is $(\neg p) \wedge q$? Does any real number x satisfy this new requirement? What does $\neg(p \wedge q)$ say about x?

What about $p \vee q$? (Answer: $p \vee q$ says *nothing* about a real number x, since every x is > 0, or ≤ 5, or both. The statement $p \vee q$ is T no matter what x is.)

Next, let r be the statement '$x \in \mathbb{Q}$'. What is the meaning of $\neg(p \wedge q \wedge r)$?

1.2.2 *Truth tables*

The bit of practice at the end of the previous subsection should illustrate the fact that figuring out the meaning of a composite statement can be a bit tricky, as soon as two or three statements are involved. *Truth tables* are a systematic way to analyze such situations, by just going through all possible alternatives. The point is that the result of connecting logical statements only depends on whether these statements are true or false, not on any deeper 'meaning' of these statements. All we need to know about $\wedge$, $\vee$, $\neg$ are the definitions given above, which may be represented by 'truth tables' as follows:

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

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

p	$\neg p$
T	F
F	T

The first few columns simply list all possibilities for the different variables. (There will be 2^n possibilities if there are n variables.) The other columns display the results of applying the corresponding expression to the variables. Larger tables can be compiled by successively looking up the corresponding entries in these tables. Here are a few examples:

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

Oh! The columns $p \wedge (q \vee r)$ and $(p \wedge q) \vee (p \wedge r)$ agree! This says that the two expressions $p \wedge (q \vee r)$ and $(p \wedge q) \vee (p \wedge r)$ are *logically equivalent.* No matter what p, q, r mean, these two expressions would turn out to mean the same thing, in the sense that they would be simultaneously both true or both false, in all possible situations. This is where truth tables help. Maybe one could, in any given circumstance, translate a given logical expression into what it 'means', and compare it with the 'meaning' of other logical expressions. But this can be difficult; a truth table is easier to handle and infallible, precisely because it is not concerned with the meaning of the expressions, but exclusively with their possible truth values.

do not distribute

To express that $p \wedge (q \vee r)$ and $(p \wedge q) \vee (p \wedge r)$ are logically equivalent, I will write[6]

$$p \wedge (q \vee r) \iff (p \wedge q) \vee (p \wedge r) \quad . \tag{1.3}$$

The new symbol $\iff$ is another logical connective, which returns T if its arguments agree, and F if they do not; in writing (1.3) on its own, I am implicitly telling the reader that this statement is T in all possible cases. In English, $\iff$ is translated to *'if and only if'*, or *'is equivalent to'*. ('If and only if' is sometimes contracted to 'iff'.) In terms of truth tables, $\iff$ is defined by

p	q	$p \iff q$
T	T	T
T	F	F
F	T	F
F	F	T

Thus, we have verified above that

$$p \wedge (q \vee r) \iff (p \wedge q) \vee (p \wedge r)$$

has value T *no matter* what its arguments p, q, r stand for. We call such a statement a *tautology.* The simplest tautology is probably $p \vee \neg p$:

p	$\neg p$	$p \vee \neg p$
T	F	T
F	T	T

This is also called the 'principle of excluded middle': in common language, it states that so long as a statement p must be definitely true or false, then it (p) or its negation ($\neg p$) *must* be true[7].

The 'opposite' of a tautology is called a *contradiction:* this is a statement that is F no matter what. The simplest example is $p \wedge \neg p$. (The tautology $\neg(p \wedge \neg p)$ also has a name: it is called the 'law of noncontradiction'.)

There are several useful tautologies, showing that certain collections of connectives can be replaced with other collections. For example, you can verify that both $\wedge$ and $\vee$ are 'associative' operations, in the sense that

$$\begin{aligned} (p \wedge q) \wedge r &\iff p \wedge (q \wedge r) \\ (p \vee q) \vee r &\iff p \vee (q \vee r) \quad ; \end{aligned}$$

it follows that we can simply write $p \wedge q \wedge r$, $p \vee q \vee r$, without ambiguity as to the meaning of these expressions. Another example: *De Morgan's laws* state that

$$\begin{aligned} \neg(p \vee q) &\iff (\neg p) \wedge (\neg q) \\ \neg(p \wedge q) &\iff (\neg p) \vee (\neg q) \end{aligned}$$

So: distributing NOT switches AND with OR. It is hopefully clear how you would go about proving De Morgan's laws: it would suffice to compile the corresponding truth tables and check that you get columns of T. Please do this. And by the way, the 'principle of the excluded middle' and the 'law of noncontradiction' are equivalent, thanks to De Morgan's laws. Check this too.

[6] Purist will object here: It would be better form to use the symbol $\iff$ for the connective, and $\longleftrightarrow$ for 'logical equivalence'. See the *caveat* that will follow this discussion, §1.2.4.
[7] There is no other alternative, no 'middle' ground.

do not distribute

Here is the truth table proving a curious equivalence:

p	q	$\neg p$	$\neg q$	$(\neg p) \vee q$	$p \vee (\neg q)$	$((\neg p) \vee q) \wedge (p \vee (\neg q))$	$p \iff q$
T	T	F	F	T	T	T	T
T	F	F	T	F	T	F	F
F	T	T	F	T	F	F	F
F	F	T	T	T	T	T	T

This shows that the 'if and only if' connective, $\iff$, is actually logically equivalent to a strange combination of $\neg$, $\wedge$, $\vee$: if you say that

$$((\neg p) \vee q) \wedge (p \vee (\neg q)) \tag{1.4}$$

then you are really saying that $p \iff q$. If we were interested in economizing our logical connectives, then we would not really need to introduce $\iff$, since we can produce precisely the same effect with a clever combination of $\neg$, $\wedge$, $\vee$. However, I challenge you to make a mental model of $((\neg p) \vee q) \wedge (p \vee (\neg q))$ and 'understand' that that is the same as $p \iff q$. (I am not sure I can do this myself.) Again, the point is that we do not need the cleverness necessary to parse such complicated statements in order to understand them: compiling the simple truth table shown above reduces the problem to a completely straightforward verification, which requires no cleverness whatsoever. We could even delegate checking the equivalence of these two statements to a computer program, if we needed to.

1.2.3 *The connective* $\implies$

Enunciating $\iff$ as 'if and only if' underscores the fact that $\iff$ can be decomposed further into interesting connectives: one expressing the 'if' part, and the other expressing the 'only if' part.

The notation[8] for *If p, then q* is $\implies$. What '$\implies$' should really stand for is actually a little subtle. Is the statement

If all lions are red, then there exists a green cow

T or *F*? You may be tempted to think that such a statement is simply meaningless, but on the face of it, something like this must be *T* or *F*, right? And then, which is it? How do we assign a truth value to such nonsense in some consistent fashion?

I like to think about the statement *If p, then q* as a 'law' that links the truth of q to that of p; the value of $p \implies q$ should be *F* if the law is broken, and *T* otherwise. If the law says, *'If you are driving on a highway, then your speed must be* 65 *mph or less'*, then you will get a fine if you *are* driving on a highway (p is *T*), and your speed *exceeds* 65 mph (q is *F*). In every other case (for example, if you are *not* driving on a highway) this particular law will not get you a fine. The expression $p \implies q$ is *F* when you get a *F*ine, and *T* o*T*herwise. In this sense, it is not unreasonable that $p \implies q$ should *only* be *F* if p is *T* and q is *F*; for example, it should be *T* if p is *F*. If you are not driving, no one should complain about your breaking speed limits. So here is the official truth table defining $\implies$:

p	q	$p \implies q$
T	T	T
T	F	F
F	T	T
F	F	T

[8] Again, see the *caveat* subsection §1.2.4 that follows.

do not distribute

Thus, if you happen to know that $p \implies q$ is T, and that p is T, then you can conclude that q is T. If $p \implies q$ is T, and p is F, then the value of q is irrelevant: it can be F or T. In particular, you get no information about q in this situation. The silly statement about lions and cows is *true* (!) according to this convention, because it just *isn't* true that all lions are red. (So we are in the situation in which p is F, and it is immaterial what q is.)

In English, I read $p \implies q$ as '*p implies q*', making $\implies$ a verb. Popular alternatives are *'If p, then q'*, *'q is true if p is true'*, *'p is a sufficient condition for q'*, and variations thereof. I prefer 'implies' because it is short and simple. In any case, please keep in mind that this symbol has this one precise meaning, codified by the truth table given above. It may be tempting to use it in a loose sense, to mean 'the preceding thoughts lead me to the following thoughts'. Please don't do that. Every mathematician on the planet interprets $\implies$ in the precise sense explained above, and any other use leads to confusion.

Common pitfall: It is also unfortunately common, and mistaken, to take the symbol $\implies$ to be a replacement for the 'then' part of *If... then...*. It is not. If you catch yourself writing something like *'If $p \implies q \ldots$'*, think twice. Do you really mean *If it is true that p implies q...*? This is what you are writing. It is *not* the same as *If p, then q*. If *this* is what you mean, then you should simply write $p \implies q$, without the *'If'* in front. To reiterate,

$$\text{'}p \implies q\text{'} \quad \text{and} \quad \text{'If } p \implies q\text{'}$$

mean very different things. The first one is a statement, which may be T or F according to the truth values of p and q. The second one is the beginning of a sentence, and by itself it is neither T or F. What you write should match what you mean. ⌟

The converse of *'q is true* if *p is true'* (that is, $p \implies q$) is *'q is true* only if *p is true*, written $p \impliedby q$. (It may take a moment to realize that this is really nothing but $q \implies p$, but so it is.) The truth table for this connective is

p	q	$p \impliedby q$
T	T	T
T	F	T
F	T	F
F	F	T

with the lone F moved to account for the switch between p and q. If I have to, I read $p \impliedby q$ as *'p is implied by q'*; again, there are common alternatives, such as *'p is a necessary condition for q.'* Again, just keep in mind that $\impliedby$ is a loaded symbol, it means exactly what is represented by the above truth table, no more and no less. Use it appropriately.

Next, let's check something that should be clear at this point: it is indeed consistent to use *'if and only if'* as a translation of $\iff$. The tautology underlying this piece of language is captured by the truth table

p	q	$p \implies q$	$p \impliedby q$	$(p \implies q) \wedge (p \impliedby q)$
T	T	T	T	T
T	F	F	T	F
F	T	T	F	F
F	F	T	T	T

The last column matches the truth values of $p \iff q$, so indeed *'if and only if'* is the same as, well, *'if AND only if'*.

Finally, look back at the complicated expression (1.4) equivalent to $\iff$. The first part of that expression,

do not distribute

$(\neg p) \vee q$, is just a restatement of $p \implies q$, as a small truth table confirms:

p	q	$\neg p$	$(\neg p) \vee q$
T	T	F	T
T	F	F	F
F	T	T	T
F	F	T	T

In words: 'p implies q' is true when either p is true, and then we require q to also be true, or p is false (in which case we don't care what q is); 'p implies q' is false precisely when p is true and q is false. This deserves to be highlighted:

Lemma 1.2.1. *$p \implies q$ is equivalent to $(\neg p) \vee q$.*

Of course, the other part of (1.4), is just a restatement of $p \impliedby q$, as you can verify with another truth table:

p	q	$\neg q$	$p \vee (\neg q)$
T	T	F	T
T	F	T	T
F	T	F	F
F	F	T	T

or by just switching the roles of p and q if you prefer.

Remark 1.2.2 (The meaning of '*p does not imply q*'). The expression $\neg(p \implies q)$ has truth table

p	q	$\neg(p \implies q)$
T	T	F
T	F	T
F	T	F
F	F	F

Using the equivalent formulation of $\implies$ in terms of $\neg$ and $\vee$, and De Morgan's laws, we see that $\neg(p \implies q)$, that is, $\neg((\neg p) \vee q)$, is equivalent to $p \wedge (\neg q)$. In other words, (it is true that) p does not imply q precisely when p is true *and yet* q is false: that is, in situations in which the truth of p is not sufficient to trigger the truth of q. This makes some sense, does't it? ⌟

1.2.4 Caveat

You should be warned that, in spite of all my propaganda, a lot of what I have written above is woefully imprecise. The main culprit is a common abuse of language. If I write $7 > 5$, I usually just mean that, well, 7 is larger than 5. That is, I really mean that the *statement* $7 > 5$ is true. In principle, 'p' and 'p is true' are different gadgets: the first is a statement (say, $7 > 5$), and the second (say, $7 > 5$ is T) is an assertion about this statement. If we needed to be very precise, we should make a distinction between the two, maybe use a different symbol. In practice, the context usually indicates which of the two we mean.

Similarly: '$p \implies q$' is a statement, but we often write $p \implies q$ to mean that p does imply q, that is, that the corresponding statement is true. To make the distinction clear, we should adopt another symbol instead of $\implies$. One such symbol is $\vdash$; or we could agree to use something like $\to$. Luckily the context usually takes

do not distribute

care of this, so at the risk of angering my logician friends I am going to just gloss over the distinction. My feeling is that when you really perceive the nature of these distinctions, then you will also be savvy enough to be able to deal with them as precisely as may be needed.

1.2.5 How many connectives are really necessary?

I have observed that one can formulate $p \iff q$ and $p \implies q$ by just using $\neg$, $\wedge$, $\vee$: for example, $p \implies q$ is equivalent to $(\neg p) \vee q$. I will now claim that *every* logical manipulation can be expressed in terms of $\neg$, $\wedge$, $\vee$: indeed, every truth table can be put together with these connectives.

Why?

A truth table is an assignment of T or F to every possible truth value of the variables. Now if I have a way to produce a truth table with T in exactly one place, and F otherwise, then by OR-ing many such tables I would get any desired truth table. For example, suppose I am interested in the following (randomly chosen) table:

p	q	r	??
T	T	T	F
T	T	F	T
T	F	T	T
T	F	F	F
F	T	T	F
F	T	F	T
F	F	T	F
F	F	F	F

I have three Ts in the table, corresponding to TTF, TFT, and FTF. I can produce these Ts individually, by negating some of p, q, and r and AND-ing the results. Here is how I can do it:

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

The point is that $\wedge$ has exactly one T in its table, so I just need to arrange for that T to show up in the right spot. And then, OR-ing the three resulting columns reproduces the table I was given, so my mystery operation ?? can be written as

$$\big(p \wedge q \wedge (\neg r)\big) \vee \big(p \wedge (\neg q) \wedge r\big) \vee \big((\neg p) \wedge q \wedge (\neg r)\big) \quad .$$

There may be much more economical ways to reproduce the same table, but this works. As any table can be approached in the same way, we see that $\neg$, $\wedge$, $\vee$ suffice to reproduce any logical operation.

do not distribute

In fact, you can do even better. Define a connective NAND ('NOT AND') as follows: p NAND q will be the same as $\neg(p \wedge q)$. Truth table:

p	q	p NAND q
T	T	F
T	F	T
F	T	T
F	F	T

Then I claim that *every logical operation may be obtained by exclusively stringing NANDs.* You will check this in your homework. Thus, a *single* operation would suffice; using three ($\neg$, $\wedge$, $\vee$) or more ($\implies$, $\iff$) is a convenient luxury. It is (to me) philosophically satisfying to realize that one basic logic operation suffices in order to reproduce the most complex logical expression one could imagine; but again, our goal in communicating mathematics should be to make the reader's life as easy as possible, and writing our logic by exclusively stringing together elaborate combinations of NANDs would definitely not help the reader. Please show some restraint in your writing.

1.2.6 Homework

1.2.1: Assume that $L(x)$ means *'x is a lion'*; $C(x)$ means *'x is a cow'*; $R(x)$ means *'x is red'*; $G(x)$ means: *'x is green'*; and S is the set of all animals that can fly.

- Translate into plain English:
  - $\exists x: \quad L(x)$
  - $\exists x: \quad L(x) \wedge \neg C(x)$
  - $(\exists x:\ L(x) \wedge \neg C(x)) \implies (\forall x:\ L(x) \implies x \in S)$
- Translate into symbols:
  - Some cows can fly.
  - There are red lions.
  - There are no red lions.
  - Every red cow is a green lion.
  - If red cows cannot fly, then green lions can fly.

▷ **1.2.2:** Translate into a logical expression the adage *'All that glitters is not gold'*, assuming that $P(x)$ means *'x glitters'*, and $Q(x)$ means *'x is gold'*. (Watch out! this is trickier than it looks. Make sure that your translation matches the *meaning* of the sentence, not just its apparent structure.)

▷ **1.2.3:** Consolidate the statement

$$(\exists a : p(a)) \vee (\exists b : q(b))$$

so as to use only one quantifier on only one variable. Can the statement $(\exists a : p(a)) \wedge (\exists b : q(b))$ be expressed with just one quantifier? What about $(\forall a : p(a)) \vee (\forall b : q(b))$? And $(\forall a : p(a)) \wedge (\forall b : q(b))$?

1.2.4: Prove De Morgan's laws.

1.2.5: The 'distributive' law for multiplication w.r.t. addition states that $(a+b)c = ac + bc$ for all a, b, c. Prove 'distributive laws' for the $\wedge$ and $\vee$ logical connectives:

$$p \wedge (q \vee r) \iff (p \wedge q) \vee (p \wedge r)$$
$$p \vee (q \wedge r) \iff (p \vee q) \wedge (p \vee r)$$

do not distribute

1.2.6: Prove that $((p \vee q) \implies r) \iff ((p \implies r) \wedge (q \implies r))$.

1.2.7: Prove that the statement

$$(p \wedge q) \vee ((\neg p) \wedge (\neg q))$$

is also equivalent to $p \iff q$.

▷ **1.2.8:** In §1.2.5 we have seen how to get any truth table by OR-ing suitable ANDs. Explain how to get any truth table by AND-ing suitable ORs.

▷ **1.2.9:** Play with the NAND operation:

- Verify that $p \implies q$ is equivalent to p NAND (q NAND q).
- Show[9] that every truth table can be obtained by exclusively using NAND operations. (Hint: By what we have seen, it suffices to write $\neg$, $\wedge$, $\vee$ by just using NANDs. Why?)
- Write $p \iff q$ by just using NANDs.

1.3 *More about logic. Techniques of proofs.*

1.3.1 *The behavior of quantifiers under negation.*

Suppose a statement p depends on a variable and a quantifier:

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

as in §1.1.4. What is the negation

$$\neg(\forall x\,,\ p(x)) \quad ?$$

Think about it a second before reading ahead.

Ready?

Suppose you have a deck of cards, and I tell you: *every card in your deck is an ace of spades.* You turn the first one, and sure enough it is an ace of spades. Then you turn the second, and it is the queen of hearts! Do you need to turn any other card to know that my statement was false? Of course not. The *existence* of a card that is not the ace of spades suffices to show that my claim was not true. In fact:

$$\neg(\forall x\,,\ p(x)) \quad \iff \quad (\exists x\,,\ \neg p(x)) \quad .$$

Note that the $\forall$ quantifier magically turns into an $\exists$ quantifier as $\neg$ goes through it. And of course

$$\neg(\exists x\,,\ p(x)) \quad \iff \quad (\forall x\,,\ \neg p(x)) \quad :$$

if you want to convince me that *none* of your cards is the ace of clubs, then you have to turn *all* of them and show that each and every one of them 'fails the property of being the ace of clubs' (that is, it is not the ace of clubs).

If all this sounds too common-sensical to be mathematics, so be it: at this level, we are doing nothing more than formalizing common sense into a (more) precise language.

[9] 'Show', 'Prove', etc. really mean the same thing. Here I am not looking for something very formal; just a precise explanation of why the statement that follows is true.

do not distribute

1.3.2 *Do examples prove theorems? Do counterexamples disprove them?*

A *theorem* (or 'proposition', if we do not want to emphasize its importance) is a true (mathematical) statement. The words *lemma* and *corollary* are also used to denote theorems: a 'lemma' is a preliminary result whose main use is as an aid in the proof of a more important theorem, and a 'corollary' is a true statement which follows easily from some known theorem. Sometime I use the word 'claim' to highlight a small but important observation in a larger discussion. But these distinctions are just meant to help a reader perceive a hierarchy in the statements as meant by the writer. Lemmas, corollaries, claims, and propositions are just theorems.

Typically, a mathematical theorem may state that a certain statement q (the 'thesis') is true subject to the truth of another statement p (the 'hypothesis'). Both p and q may depend on some variable, say x, and then the theorem could be stating that

$$\forall x, \quad p(x) \implies q(x) \quad : \tag{1.5}$$

in words, *'For all x, if $p(x)$ is true, then $q(x)$ is true'*; or *'Property q holds for all the x satisfying property p'*. Many theorems fit this general template. For example, *'if n is an odd integer, then n^2 is also odd'* is a statement of this type:

$$\forall n \in \mathbb{Z}, \quad p(n) \implies q(n) \quad , \tag{1.6}$$

where $p(n)$ stands for *'n is odd'* and $q(n)$ stands for 'n^2 is odd'. Would the example $3^2 = 9$ prove (or 'demonstrate', or 'show'...) this theorem? That's an odd integer whose square is an odd integer. But: *no*, this example does not prove the theorem. The *existence* of this example proves that

$$\exists n \in \mathbb{Z}, \quad p(n) \implies q(n) \quad ,$$

and that is not the same statement as (1.6); it is 'weaker'. By and large, proving a statement of the general form (1.5), such as (1.6), will require providing a string of logical arguments deriving the truth of q from the truth of p. Looking at examples may be extremely beneficial in order to come up with such a general argument, or to illustrate the argument[10], but the examples themselves will not prove a statement of the type (1.5) (unless of course one can list and verify *all* possible cases covered by $\forall x$).

How would you prove that my theorem is *not* true? That is the same as convincing me that its negation

$$\neg(\forall x, \quad p(x) \implies q(x))$$

is true. As we have seen in §1.3.1, this is equivalent to

$$\exists x, \quad \neg(p(x) \implies q(x))$$

and as we have seen in Remark 1.2.2, this is equivalent to

$$\exists x, \quad p(x) \wedge (\neg q)(x) \quad .$$

Thus, to 'disprove' (1.5) it is enough to produce one example for which the hypothesis p holds but the thesis q does not hold. Show me a single *odd* integer whose square is *not* an odd integer[11], and that will suffice

[10] I have myself used this technique to try to convince you that $\neg$, $\wedge$, $\vee$ suffice to reproduce any logical operation: I have illustrated this general fact by an example. What I have written was *not* a proof, even if it was (I hope) a convincing discussion. If it looks like a proof, it is because the example illustrates a general algorithm, and that algorithm could be the basis of a 'real' proof.

[11] Good luck with that.

do not distribute

to prove that my statement (1.6) was not correct. A *counterexample* suffices to *disprove* a general statement depending on a $\forall$ quantifier.

On the other hand, there are interesting theorems that do not fit the above template. For example, a theorem may state that there exist integers[12] $n > 1$ with the property that n equals the sum of its proper positive divisors. Such numbers are said to be 'perfect'. For example, the proper divisors of 12 are 1, 2, 3, 4, 6, whose sum is $1 + 2 + 3 + 4 + 6 = 16$, so 12 is *not* perfect. I could propose the theorem

$$\exists n > 1 : \quad n \text{ is perfect.} \tag{1.7}$$

(where $n \in \mathbb{Z}$ implicitly). In this case, an *example* of a perfect number would be plenty to prove the theorem, because the theorem is asking for just that: the existence of a perfect number. *Disproving* the theorem, however, would amount to showing that $\neg(\exists n > 1 : n \text{ is perfect.})$, that is (§1.3.1 again):

$$\forall n > 1 : \quad n \text{ is not perfect.}$$

One example, or a million examples, will not suffice to show this. Thus, in this case you could not *dis*prove my theorem by finding a counterexample.

I am stressing these points because I have met many students who seem to be confused about whether one can 'prove a theorem by giving an example'. This depends on how the theorem is formulated. If the theorem depends on a *universal* quantifier, then an example won't do. If on the other hand the theorem states the existence of an entity with a certain property, then an example will do. Everything becomes clear if one writes out the quantifiers explicitly, and processes a bit of simple logic as need be. With a bit of practice, such logic manipulations will become second-nature.

What would be a 'string of logical arguments' proving our first theorem (1.6), stating that if n is an odd integer, then n^2 is an odd integer? This sounds fancier than it is. An integer is odd if it is one plus an even number: if n is odd, then $n = 2k + 1$ for some integer k. Then

$$n^2 = (2k + 1)^2 = 4k^2 + 4k + 1 = 2(2k^2 + 2k) + 1 \quad :$$

this shows that if n is odd, then n^2 also is one plus an even number, so n^2 is indeed necessarily odd.

This paragraph would be considered a proof of the fact we stated. It does not pick one example for n; rather, it uses the only thing we may assume about n (that it is odd) to trigger a computation whose outcome proves what we need to prove.

What about my second theorem (1.7), 'perfect numbers exist'? Proof: 6 is perfect, since the proper divisors of 6 are 1, 2, 3, and $1 + 2 + 3 = 6$ as required by the definition of perfect number.

This is all that is needed in this case.

In case you find this somewhat anticlimactic, note that *no one knows whether there is any* odd *perfect number.* Thus, the statement

$$\exists n > 1 : \quad n \text{ is odd and perfect.}$$

is unsettled: no one has been able to prove this statement, and no one has been able to disprove it. If you prove it, or disprove it, you will gain some notoriety: maybe the New York Times will run an article on your accomplishment, and you will probably get an instant math Ph.D. and an offer of a posh postdoc in a good university on the strength of this one achievement.

[12] Incidentally, 'there exist integers' may suggest that we are stating there are *at least two* integers. For whatever reason, this is not the case: the plural 'integers' stands for 'one or more integers'. So finding *one* example will suffice here. If we needed to insist that there should be at least two, we would say so explicitly: *there exist two or more. . . .*

do not distribute

1.3.3 *Techniques of proof*

Let's now focus on the standard techniques used to construct proofs of theorems.

To give a *proof* of a theorem means to give a sequence of statements, starting from some statement that is already known to be true, with each statement logically following from the preceding ones, such that the last statement in the list is the needed theorem. In practice one writes an explanation in plain English, maybe enhanced by notation, from which the steps proving the theorem could be written out formally if needed.

The 'logic' needed here goes back to tautologies (as in §1.2) that could be checked explicitly if needed. For example, it is common to string one implication after the other to construct a larger argument. The basic syllogism

> Socrates is a man
> All men are mortal
> Therefore Socrates is mortal

may be interpreted as satisfying this general pattern:

> x is Socrates $\implies$ x is a man
> x is a man $\implies$ x is mortal
> Therefore x is Socrates $\implies$ x is mortal.

This probably makes sense to you without further comment, *but* if you challenged me on its 'mathematical' truth, I would argue that:

- If we know that
  - $p \implies q$ is true, and that
  - $q \implies r$ is true,

  then we can conclude that $p \implies r$ is also true.

If I had to defend this reasoning, I could do it with a truth table. Indeed, from

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

it follows that

$$\big((p \implies q) \wedge (q \implies r)\big) \implies (p \implies r)$$

is a tautology, as claimed.

Thankfully, in real life we almost never have to go all the way down to a truth table, even if in principle this can be done. The point is that only a few basic patterns show up with any frequency in the way proofs are written, so once you get used to them you do not need to verify each time that they work.

In fact, it makes sense to look at these patterns, as templates for our own proofs. The basic strategy underlying most proofs is the following.

do not distribute

- If we know that
  - p is true, and that
  - $p \implies q$ is true,

  then we can conclude that q is also true.

Why does this work? because of the truth table for $\implies$: if p is T, then $p \implies q$ is T if q is also T, and it is F if q is F. That is, once the truth of p is established, then proving q or proving $p \implies q$ are equivalent actions.

This type of argument is sometimes called a *direct proof.* Typically the logic will be so transparent that it is hard to even see it at work.

Example 1.3.1. *Let n be an integer. Assume that n is even; then n^2 is even.*

Here p is 'n is even'; q is 'n^2 is even'. A direct proof could go as follows:

—Assume that n is even; so $n = 2k$ for some integer k.
—If $n = 2k$, then $n^2 = (2k)^2 = 4k^2 = 2(2k^2)$.
—As n^2 is of the form 2·(an integer), we conclude n^2 is indeed even.

The *'if...then'* in the middle realizes the needed implication $\implies$ in this case. □

In general, the interesting part of the basic strategy is the implication $p \implies q$: once you establish the truth of this implication, then as soon as you run into a situation where p is known to be true, then the truth of q follows. Different flavors of proofs arise from statements that are tautologically equivalent to $p \implies q$. For instance:

- If we know that
  - $(\neg q) \implies (\neg p)$ is true,

  then we can conclude that $p \implies q$ is also true.

Indeed, $(\neg q) \implies (\neg p)$ is equivalent to $p \implies q$, as a truth table checks in a moment:

p	q	$\neg q$	$\neg p$	$(\neg q) \implies (\neg p)$	$p \implies q$
T	T	F	F	T	T
T	F	T	F	F	F
F	T	F	T	T	T
F	F	T	T	T	T

The statement $(\neg q) \implies (\neg p)$ is called the *contrapositive* of $p \implies q$. Replacing $p \implies q$ with $(\neg q) \implies (\neg p)$ in the basic strategy leads to a 'proof by contrapositive'.

This may look strange at first, particularly because in the process of proving $(\neg q) \implies (\neg p)$ you are likely to assume that $\neg q$ is *true,* that is, that q is *false.* So we are trying to prove that q is true, and we do that by assuming it is not? What nonsense is that?

Thinking carefully about it now will reconcile you with this 'nonsense' once and for all and realize that it is, on the contrary, perfectly sound reasononing. Again, $p \implies q$ and $(\neg q) \implies (\neg p)$ are certifiably equivalent, so the contrapositive strategy makes exactly as much or as little sense as the direct one.

Example 1.3.2. Here is a simple fact we can contemplate at this stage.

$$(\forall x \in \mathbb{R}\,,\, x^2 \in S) \implies S \text{ contains every nonnegative real number.} \tag{1.8}$$

do not distribute

It is a good exercise for you to build a 'direct' proof of this fact (Exercise 1.3.6). For fun, let's prove it 'by contrapositive'. This amounts to establishing that

$$\neg\big(S \text{ contains every nonnegative real number}\big) \implies \neg\big(\forall x \in \mathbb{R}\,,\, x^2 \in S\big)$$

that is (cf. §1.3.1)

$$\big(S \text{ does not contain every nonnegative real number}\big) \implies \big(\exists x \in \mathbb{R}\,,\, x^2 \notin S\big)$$

Proof. Assume r is a nonnegative real number that is *not* an element of S, and let $x = \sqrt{r}$. Then x is a real number such that $x^2 = r \notin S$. Thus we see that $\exists x \in \mathbb{R}\,,\, x^2 \notin S$, as claimed. □

This would count as a proof of (1.8). Whether you like it better than a direct argument is a matter of taste. There is a general tendency to favor direct arguments when they are available, because their logic is (even) more transparent. ⌟

Here is another 'technique of proof':

- If we know that
  - $p \wedge (\neg q)$ implies a statement that is known to be false,

  then we can conclude that $p \implies q$ is true.

Indeed, if $p \wedge (\neg q)$ implies a false statement, then it is itself false[13]. However, $\neg(p \wedge (\neg q))$ is equivalent to $p \implies q$:

p	q	$\neg q$	$p \wedge (\neg q)$	$\neg(p \wedge (\neg q))$	$p \implies q$
T	T	F	F	T	T
T	F	T	T	F	F
F	T	F	F	T	T
F	F	T	F	T	T

Thus, establishing that $(p \wedge (\neg q))$ is F is the same as establishing that $p \implies q$ is T, and then we are reduced again to our basic strategy underlying proofs.

As *false* statements are called 'contradictions', proofs adopting this strategy are called 'proofs by contradiction'. In practice, in proving a statement by contradiction one assumes that the hypothesis and the negation of the thesis both hold, and proves that some nonsensical statement would follow. It is often the case that minor changes in the wording will turn a contradiction proof into a contrapositive one; these distinctions are not very substantial.

Example 1.3.3. Consider again the statement proved in Example 1.3.1:

Let n be an integer. Assume that n is even; then n^2 is even.

Proof by contradiction. We assume that n is even, so $n = 2k$ for some integer k, *and* that $m = n^2$ is odd. Since $m = n^2 = (2k)^2 = 2(2k^2)$, this integer m would have to be both even and odd. This is a contradiction: an integer cannot be simultaneously even and odd. This contradiction proves our statement. □

You should feel a bit cheated, in the sense that this argument 'by contradiction' is none other than our original 'direct' argument, just dressed up a little differently. Again, these distinctions are not very substantial,

[13] This makes sense, but can you formalize it in terms of truth tables? Exercise 1.3.5.

do not distribute

and the way you decide to write down a proof is often a matter or taste. You should simply adopt a strategy that makes your proof easier to read: sometimes writing a contradiction argument is really your best choice; other times (as in this example) the extra twist needed to turn the argument into a contradiction argument does not improve its readability. ⌟

It is important to realize that these are all skeletal examples. In practice, proofs are *not* written by putting together a series of statements labeled p, q, r, etc., or by compiling truth tables. This would make for unbearably dry style, and in the end it would not convey your thoughts very effectively. But it is good to keep in mind that, *in principle,* one could take a proof and break it down to show its inner working. At some level, this would involve discussions analogous to those summarized above.

There is another frequent technique of proof, called *induction,* involving an ingredient we have not encountered yet. We will cover induction in the next chapter (§2.2.3).

1.3.4 Homework

1.3.1: The fact that $\neg(\forall x\colon p(x))$ is logically equivalent to $(\exists x\colon \neg p(x))$ may evoke one of De Morgan's laws: $\neg(p \wedge q)$ is logically equivalent to $(\neg p) \vee (\neg q)$. Find out in what sense the latter is a particular case of the former.

▷ **1.3.2:** The useful abbreviation $\forall x \in A$, $p(x)$ stands for *'the statement $p(x)$ is true for all x in the set A'.* In other words, *'for all x, if x is an element of A, then $p(x)$ is true'.* More formally, it is shorthand for

$$\forall x \quad \big((x \in A) \implies p(x)\big) \quad .$$

What is the formal way to interpret the abbreviation $\exists x \in A, p(x)$?

1.3.3: Find[14] a perfect number other than 6.

▷ **1.3.4:** Are

$$\forall x > 0 \; \exists N : n > N \implies |s_n| < x$$

and

$$\forall x > 0 \; \exists N : n > N \implies |s_n| < 3x$$

equivalent statements, in the sense that they express the same condition on any indexed set $\{s_n\}_{n\in\mathbb{N}}$? If so, prove that they are. If not, find an example of an assignment s_n which satisfies one of the statements but not the other.

▷ **1.3.5:** Explain why the truth table for $\implies$ tells us that if p is a statement that implies a false statement (that is, if it is *true* that $p \implies q$, where q is *false*), then p is itself false.

○ In the following three problems, pay attention to the specific request. The point is not to show that you can prove these statements—they should be straightforward enough that this should not be an issue. The point is to show that you can write proofs in the styles reviewed in this section. An integer n is *even* if there is an integer k such that $n = 2k$; it is *odd* if there is an integer k such that $n = 2k + 1$.

[14] What does 'find' mean? I timed myself, and it took me 11 seconds (without rushing) to *find* another perfect number by looking it up on google. This is not what you are supposed to do. Can you find another perfect number *without outside help?* The same clarification will apply to every instance when you are required to 'find' something.

There is a two-digit perfect number, and it is not that large, so finding it shouldn't take you long. *After* you do this, you may enjoy reading the wikipedia page on perfect numbers.

do not distribute

▷ **1.3.6:** Prove 'directly' that

$$(\forall x \in \mathbb{R}\,,\, x^2 \in S) \implies S \text{ contains every nonnegative real number.}$$

(Cf. Example 1.3.2.)

▷ **1.3.7:** Give a 'proof by contrapositive' of the following statement: *Let n be an integer; if n^2 is even, then n is even.*

1.3.8: Give both a 'direct proof' and a 'proof by contradiction' of the following statement: *Let a, b be integers. If a is even and b is odd, then $a + b$ is odd.*

do not distribute

do not distribute

2
Sets and relations

2.1 Naive set theory.

2.1.1 Sets: Basic definitions

Now that we are a little more experienced with the basic language of logic, we can go back to our definition of 'set' and review several related notions.

Recall that, for us, a *set* S is a collection of *elements*; we write $a \in S$ to mean: *'The element a belongs to the set S'*. We write $a \notin S$ to mean: *'The element a does not belong to the set S'*. We put no restriction on what sort of things the 'elements' should be; for the naive considerations we are dealing with, this is unimportant. As it happens, in order to build a consistent *set theory* one has to be much more careful in dealing with these issues, but that will not concern us here. If you are curious about this, spend some time with Exercise 2.1.12.

If S is a finite set (that is, it contains only finitely many elements), the number of its elements is denoted $|S|$.

Definition 2.1.1. The *cardinality* of a finite set S is the number $|S|$ of elements in S. ⌟

It is actually possible to talk about the cardinality of *infinite* sets. We will deal carefully with this in Chapter 5 (cf. Definition 5.1.1).

Remember (from §1.1.1) that two sets A, B are equal precisely when they consist of the same elements. I can state this fact by writing that if A and B are sets, then $A = B$ precisely when

$$\forall x \qquad x \in A \iff x \in B \quad .$$

So 'equality of sets' is an embodiment of the ' $\iff$ ' logical connective. We will see that all logical operations have a counterpart in naive set theory. Let's start with $\implies$.

Definition 2.1.2. Let A, B sets. We say that A is a *subset* of B if

$$\forall x \qquad x \in A \implies x \in B \quad .$$

To denote that A is a subset of B we write $A \subseteq B$. ⌟

Incidentally, a *definition* in mathematics is a prescription which assigns a precise meaning to a term. It is not a theorem. This is a little confusing, since we often write definitions as 'if…' statements, just as we write theorems. They are different beasts: in a theorem, *'if…then…'* relates mathematical statements; *'We say that A is a subset of B'* is not a mathematical statement. The 'if' in Definition 2.1.2 just introduces a piece of language ('A is a subset of B') as a way to refer to a mathematical statement ('$\forall x,\ x \in A \implies x \in B$'), and

do not distribute

offers a notation for it ('$A \subseteq B$'). No new 'truth' is discovered in the process. All Definition 2.1.2 does is to give a shorthand that will allow us to write other mathematical statements more efficiently.

Intuitively, A is a subset of B if it is 'contained in' B. We can also write $B \supseteq A$, and say that 'B contains A'. Some mathematicians prefer the symbol $\subset$ to $\subseteq$, and $\supset$ to $\supseteq$, with exactly the same meaning. I prefer the version decorated with the extra horizontal line because this turns out to be an *order* relation (cf. §2.2.2) and $\leq$ (*with* the horizontal line) is my prototype of an order relation.

Note that the empty set $\emptyset$ (see §1.1.2) is contained in *every* set: for all sets S,

$$\emptyset \subseteq S \quad .$$

Indeed, the 'law' $x \in \emptyset \implies x \in S$ is never challenged, since '$x \in \emptyset$' is F no matter what x is. So $\emptyset \subseteq S$ by virtue of our convention that $p \implies q$ should be T if p is F (cf. §1.2.3). Another direct consequence of the truth table for $\implies$ is that

$$A \subseteq A$$

for every set A. Indeed, from the truth table we see that $p \implies p$ is T both when p is T and when p is F; so $x \in A \implies x \in A$ no matter what x is (that is, $\forall x$).

To express that A is contained in B but it is not equal to it, one writes $A \subsetneq B$.

The following natural-looking fact is also a direct consequence of the basic considerations on logic we reviewed in §1.2:

Proposition 2.1.3. *Let A, B be sets. Then*

$$A = B \iff \big(A \subseteq B \wedge B \subseteq A\big) \quad .$$

This follows from the fact that $A \subseteq B$ is a counterpart to $\implies$, while $A = B$ is a counterpart to $\iff$. Explicitly:

Proof. Let p be the statement $x \in A$, and let q be the statement $x \in B$. The tautology

$$(p \iff q) \iff ((p \implies q) \wedge (q \implies p))$$

was observed and verified towards the end of §1.2.3. It translates to

$$(x \in A \iff x \in B) \iff \big((x \in A \implies x \in B) \wedge (x \in B \implies x \in A)\big) \tag{2.1}$$

and therefore to

$$A = B \iff \big(A \subseteq B \wedge B \subseteq A\big) \tag{2.2}$$

which is the statement[1]. □

The uniqueness of the empty set follows from Proposition 2.1.3, so we can call it a 'corollary':

Corollary 2.1.4. *There is only one empty set. That is: if $\emptyset_1$ and $\emptyset_2$ are both empty sets, then $\emptyset_1 = \emptyset_2$.*

Proof. We have observed that empty sets are subsets of every set. In particular,

$$\emptyset_1 \subseteq \emptyset_2$$

[1] Here the reader sees the usefulness of definitions: (2.1) and (2.2) mean precisely the same thing, but (2.2) is much easier to read.

do not distribute

since $\emptyset_1$ is an empty set; and

$$\emptyset_2 \subseteq \emptyset_1$$

since $\emptyset_2$ is an empty set. But then we have obtained that

$$(\emptyset_1 \subseteq \emptyset_2) \wedge (\emptyset_2 \subseteq \emptyset_1) \quad ,$$

and by Proposition 2.1.3 we can conclude that $\emptyset_1 = \emptyset_2$ as needed. □

Actually, the uniqueness of the empty set already follows from the fact that two sets are the same if they have the same elements: if $\emptyset_1$ and $\emptyset_2$ are both empty sets, then they both have no elements, so they vacuously satisfy this requirement.

2.1.2 *Sets: Basic operations*

Definition 2.1.5. Let A, B be sets.

- The *intersection* $A \cap B$ of A and B is the set

$$A \cap B = \{x \mid (x \in A) \wedge (x \in B)\} \quad .$$

- The *union* $A \cup B$ of A and B is the set

$$A \cup B = \{x \mid (x \in A) \vee (x \in B)\} \quad .$$

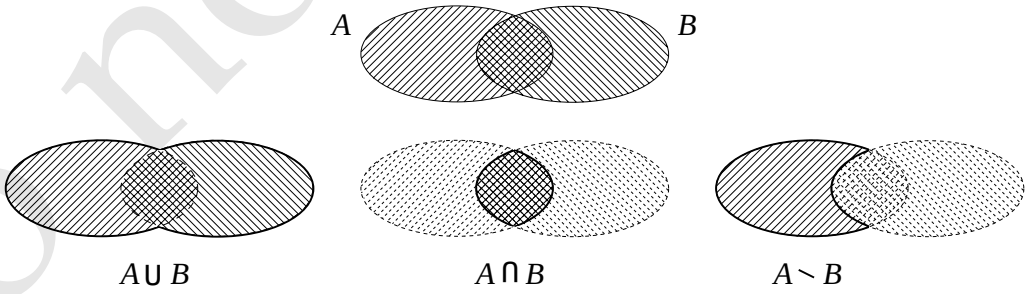
- The *difference* $A \smallsetminus B$ is the set

$$A \smallsetminus B = \{x \in A \mid x \notin B\} \quad .$$

If $B \subseteq A$, then $A \smallsetminus B$ is called the *complement* of B in A.

Remark 2.1.6. Thus, naive set theory simply 'concretizes' logic: the operation $\cap$ is a counterpart of $\wedge$, $\cup$ is a counterpart of $\vee$ (note the similarity in the symbols we use for these notions!), and the complement is a counterpart of $\neg$, in the sense that it expresses the negation $\neg(x \in B)$. ⌟

Remark 2.1.7. One may visualize the set operations rather effectively by drawing so-called 'Venn diagrams':



These should be self-explanatory. They may help keeping track of what is going on, but in most cases such pictures do not prove theorems. ⌟

Do quantifiers also have a counterpart? Yes: one may take intersections and unions of arbitrary families of sets (i.e., sets of sets), and these operations correspond to the universal and existential quantifiers.

do not distribute

Definition 2.1.8. Let S_i be sets, where the index i ranges over a set I. Then the *intersection* and *union* of the sets S_i are defined respectively by

$$\bigcap_{i\in I} S_i = \{x \,|\, \forall i \in I,\, x \in S_i\}$$
$$\bigcup_{i\in I} S_i = \{x \,|\, \exists i \in I,\, x \in S_i\} \quad .$$

The bit 'S_i, where the index i ranges over a set I' may be confusing at first. I could have written: *'Let $\{S_i\}_{i\in I}$ be a family of sets'*, or more properly an 'indexed set of sets', to mean exactly the same thing.

The above definitions offer very flexible notation. For example, if $I = \{1, 2\}$, then we are just talking about two sets S_1, S_2, and the notation $\bigcap_{i\in I} S_i$, $\bigcup_{i\in I} S_i$ just reproduces $S_1 \cap S_2$, $S_1 \cup S_2$: for instance, in this case '$\forall i \in I,\, x \in S_i$' just says that $x \in S_1$ *and* $x \in S_2$, so indeed we recover the definition of $S_1 \cap S_2$.

On the other hand, I could be any set, even an infinite one, and the new Definition 2.1.8 will still make sense.

Example 2.1.9. Take $I = \mathbb{N}$, and for $n \in \mathbb{N}$ let $S_n \subseteq \mathbb{R}$ be the interval $[-n, n]$. Then

$$\bigcup_{n\in\mathbb{N}} [-n, n] = \mathbb{R} \quad .$$

Indeed, by Proposition 2.1.3 we can prove this equality by proving both inclusions $\subseteq$ and $\supseteq$. The inclusion $\subseteq$ is automatic (since all the sets $S_n = [-n, n]$ are subsets of $\mathbb{R}$, so their union also is). As for the inclusion $\supseteq$: if $r \in \mathbb{R}$, let n be any integer larger than $|r|$; then $r \in [-n, n]$. This shows that $\exists n$, $r \in S_n$, and therefore $r \in \bigcup_{n\in\mathbb{N}} S_n$ as needed, by the definition of $\bigcup_{n\in\mathbb{N}} S_n$.

Example 2.1.10. Take $I = (0, \infty)$, that is, the interval $x > 0$ in $\mathbb{R}$. For $r \in I$, let S_r be the interval $(0, r)$. Then

$$\bigcup_{r\in(0,\infty)} S_r = (0, \infty) \quad , \quad \bigcap_{r\in(0,\infty)} S_r = \emptyset \quad .$$

Indeed: the first assertion is very similar to Example 2.1.9, so it is left to the reader. As for the second one, we have to prove that *no* real number x is an element of $\bigcap_{r\in(0,\infty)} S_r$, that is, that the condition

$$\forall r \in \mathbb{R}\,,\; x \in S_r \tag{2.3}$$

is *not* satisfied by any $x \in \mathbb{R}$. Now keep in mind that $S_r = (0, r)$; if $x \leq 0$, $x \notin S_r$ for *any* r, so (2.3) is certainly not satisfied in this case. If $x > 0$, let r be any positive number no larger than x (for example, we can let r equal x itself, or $r = x/2$ for extra room if we wish). Then $x \notin (0, r)$. This shows that

$$\exists r \in \mathbb{R}\,,\; x \notin S_r$$

that is, the negation of (2.3)). So (2.3) is indeed false for every x, as claimed.

2.1.3 *Properties of* $\cap$, $\cup$, $\setminus$

The operations we have defined in §2.1.2 satisfy several elementary properties:

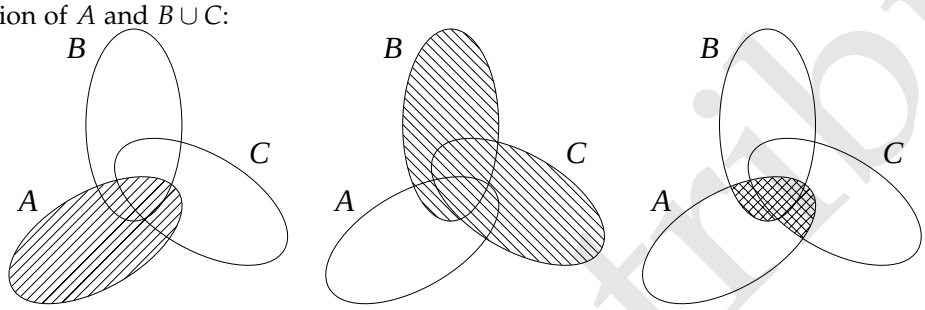
Theorem 2.1.11. *Let A, B, C be subsets of a set S. Then*

1. $A \cap B = B \cap A$
2. $A \cup B = B \cup A$

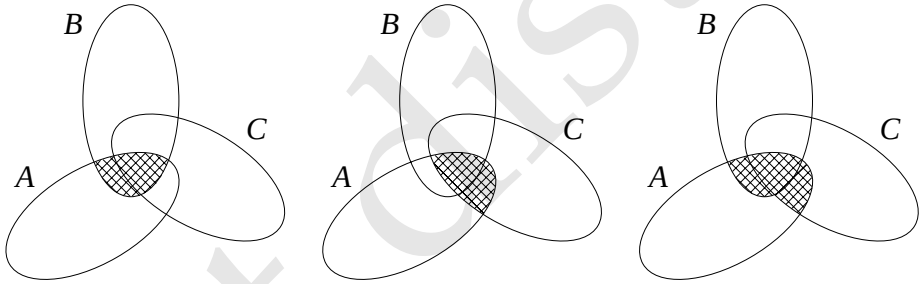
do not distribute

3. $(A \cap B) \cap C = A \cap (B \cap C)$
4. $(A \cup B) \cup C = A \cup (B \cup C)$
5. $A \cap (B \cup C) = (A \cap B) \cup (A \cap C)$
6. $A \cup (B \cap C) = (A \cup B) \cap (A \cup C)$
7. $S \smallsetminus (A \cap B) = (S \smallsetminus A) \cup (S \smallsetminus B)$
8. $S \smallsetminus (A \cup B) = (S \smallsetminus A) \cap (S \smallsetminus B)$

All the facts listed here (and many more) are transparent consequences of simple logical statements that can be checked, for example, by constructing appropriate truth tables. You may get some intuition for such properties of set operations by drawing corresponding Venn diagrams. For example, the left-hand side of (5) is the intersection of A and $B \cup C$:



The right-hand side is the union of $A \cap B$ and $A \cap C$:



The results are visibly the same. This is not a proof, however: sets are not, after all, drawings of ovals in a plane; maybe we have drawn our pictures in some special way that would make the fact accidentally true in this particular case. A real proof could go as follows:

Proof of (5). We have to compare the sets $A \cap (B \cup C)$ and $(A \cap B) \cup (A \cap C)$. An element x is in the first set if and only if

$$(x \in A) \wedge (x \in B \cup C) \quad ,$$

by definition of $\cap$. By definition of $\cup$, this means

$$(x \in A) \wedge ((x \in B) \vee (x \in C)) \quad . \tag{2.4}$$

Now, we have verified in §1.2.2 (see (1.3)) that for all statements p, q, r

$$p \wedge (q \vee r) \iff (p \wedge q) \vee (p \wedge r)$$

is T. Applying this tautology with $x \in A$ standing for p, $x \in B$ for q, and $x \in C$ for r, we see that (2.4) is equivalent to

$$\big((x \in A) \wedge (x \in B)\big) \vee \big((x \in A) \wedge (x \in C)\big) \quad .$$

do not distribute

By definition of $\cap$, this means

$$(x \in A \cap B) \vee (x \in A \cap C) \quad ,$$

and by definition of $\cup$ this says

$$x \in (A \cap B) \cup (A \cap C) \quad .$$

Thus, we have verified that

$$x \in A \cap (B \cup C) \quad \Longleftrightarrow \quad x \in (A \cap B) \cup (A \cap C) \quad ,$$

and this says precisely that

$$A \cap (B \cup C) = (A \cap B) \cup (A \cap C) \quad ,$$

as needed. □

This is as direct a proof as they come. All the other facts listed in Theorem 2.1.11 may be proved by the same technique, and the reader will practice this in the homework. Several of these facts have a counterpart for operations on families of sets. For example, let $\{S_i\}_{i\in I}$, $\{S_j\}_{j\in J}$ be two families of sets. Then

$$(\bigcap_{i\in I} S_i) \cap (\bigcap_{j\in J} S_j) = \bigcap_{k\in I\cup J} S_k \tag{2.5}$$

$$(\bigcup_{i\in I} S_i) \cup (\bigcup_{j\in J} S_j) = \bigcup_{k\in I\cup J} S_k \quad . \tag{2.6}$$

Proofs of such formulas may be obtained by manipulating logic statements involving $\exists$ and $\forall$ quantifiers. It is an excellent exercise (in fact, it is Exercise 2.1.9)—not really difficult, but you will need a steady hand to work it out successfully. Even if you do not feel like proving these formulas rigorously, try to make a mental image of what is happening; for example, make sure you understand why k ends up ranging over $I \cup J$ in both cases.

2.1.4 *The power set*

If you read about Russell's paradox as I am suggesting in the homework listed below, then you will find out that unbridled freedom in defining sets as 'collections of elements' satisfying given properties may lead to disaster: nothing would prevent you from considering the *set of all sets that are not elements of themselves*, or similar pathologies, producing logical inconsistencies. Mind you, there is nothing pathological about considering sets as elements of other sets: recall the example of the empty set, $\emptyset$, viewed as the sole element of the set $\{\emptyset\}$ (which is not empty!); or the families of sets $\{S_i\}_{i\in I}$ over which we defined extended operations $\bigcap_{i\in I}$, $\bigcup_{i\in I}$. Problems such as those raised by Russell's paradox are at a deeper level. They have been addressed by laying down firmer foundations of set theory, via (for example) the *Zermelo-Fraenkel* axioms.

Luckily, however, in the common practice of mathematics most of us do not run into the type of pathologies raised by Russell, and 'actual' set theory agrees with the 'naive' one for run-of-the-mill sets. So we need not worry too much about these problems. For the purpose of the discussion started in §2.1.2, we have an easy way out: just assume that *every* set we run across is a subset of some gigantic 'universe' set U, containing all the elements we may ever want to consider. The operations $\cap$, $\cup$, $\smallsetminus$ work then between subsets of this set S, and no logical inconsistencies arise.

There is one important operation on sets for which this approach does not solve the inherent problems of naive set theory, and with which we need to become familiar: the *set of all subsets* of a given set.

do not distribute

Definition 2.1.12. The 'power set', or 'set of parts' of a given set S is the set of *subsets* of S. It may be denoted $\mathcal{P}(S)$ or 2^S. ⌟

Example 2.1.13. The empty set $\emptyset$ has exactly one subset, that is itself. Thus, $\mathcal{P}(\emptyset) = \{\emptyset\}$ is a set with one element.

A set $\{*\}$ with exactly one element has *two* subsets: $\emptyset$, and $\{*\}$. Thus $\mathcal{P}(\{*\}) = \{\emptyset, \{*\}\}$ is a set with two elements.

A set $\{\bullet, \circ\}$ with *two* elements has *four* subsets:

$$\mathcal{P}(\{\bullet, \circ\}) = \{\emptyset, \{\bullet\}, \{\circ\}, \{\bullet, \circ\}\} \quad .$$

And so on. ⌟

As it happens, the set $\mathcal{P}(S)$ defies our easy way out, in the sense that in general it really cannot all consist of elements of some fixed universe U. We will come back to this interesting observation later on; we will be able to prove that the set $\mathcal{P}(S)$ is necessarily 'larger' than S, in a sense that we will make precise (Theorem 5.2.2). This is clear if S is finite, but requires a careful discussion in case S is an infinite set. So if U is a universe set, it cannot contain $\mathcal{P}(U)$ as a subset. If S were the 'set of all sets', then we would be able to produce a *larger* set $\mathcal{P}(S)$, and this is a contradiction.

These problems are fortunately immaterial in the common practice of mathematics, in the sense that the sets one considers in ordinary fields of analysis, algebra, topology, etc., are 'small' enough (even if they may be infinite!) to be safe from pathologies such as Russell's paradox. But it is good to be aware that the situation is less idyllic and clear-cut than it may seem at this stage. You may learn (much) more about this by reading specialized texts in logic or set theory.

2.1.5 *The cartesian product*

Towards the end of §1.1.1, we mentioned briefly that round parentheses are used to denote *ordered* sets. Thus,

$$(a, b)$$

is the 'ordered pair' of elements a, b. While the sets $\{a, b\}$ and $\{b, a\}$ are *equal,* the ordered pairs (a, b) and (b, a) differ—that is, the order in which a and b are mentioned matters.

Definition 2.1.14. Let A, B be sets. We will denote by $A \times B$ the set of ordered pairs of elements of A and B:

$$A \times B := \{(a, b) \mid (a \in A) \wedge (b \in B)\} \quad .$$

This set is the *cartesian product* of A and B. ⌟

(The adjective 'cartesian' is derived from the name of the mathematician René Descartes.)

Example 2.1.15. Let $A = \{0, 1, 2\}$ and $B = \{\circ, \bullet\}$ (where $\circ$ and $\bullet$ are any two distinct elements). Then $A \times B$ is the set

$$\{(0, \circ), (1, \circ), (2, \circ), (0, \bullet), (1, \bullet), (2, \bullet)\}$$

Since there are 3 choices for the first element of the pair, and 2 choices for the second element, it should not be surprising that the set $A \times B$ consists of precisely $3 \cdot 2 = 6$ elements. ⌟

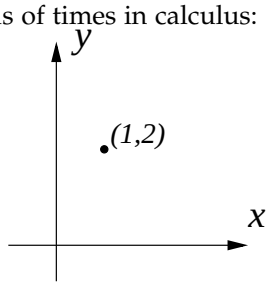
Recall (Definition 2.1.1) that if A is a finite set, we denote by $|A|$ the number of elements in A. It is hopefully clear that if A and B are finite sets, then

$$|A \times B| = |A| \cdot |B| \quad :$$

do not distribute

there are $|A|$ choices for the first element of the ordered pairs (a, b) in $A \times B$, and for each of these there are $|B|$ choices for the second element.

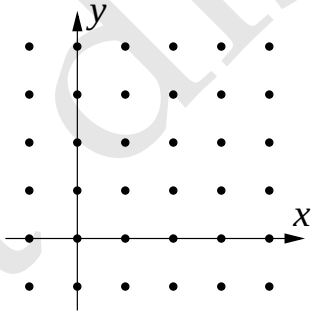
Example 2.1.16. For $A = B = \mathbb{R}$, the set of real numbers, the set $\mathbb{R} \times \mathbb{R}$ is the set of (ordered) *pairs of real numbers.* You have drawn this set hundreds of times in calculus:



The information of a 'point' on this 'plane' is precisely the same as the information of a pair of real numbers (x, y). You have likely reasonably used the notation $\mathbb{R}^2$ for $\mathbb{R} \times \mathbb{R}$. Likewise, in more advanced calculus courses or in linear algebra you have dealt with 'vectors in $\mathbb{R}^3$'; this was also a cartesian product: $\mathbb{R}^3 = \mathbb{R} \times \mathbb{R} \times \mathbb{R}$ would be the set of ordered *triples* (x, y, z) of real numbers. Of course you can think of a triple (x, y, z) as the information of a point (x, y) on the plane $\mathbb{R}^2$, together with a 'height' z. This amounts to viewing $\mathbb{R}^3$ as $\mathbb{R}^2 \times \mathbb{R}$. ⌟

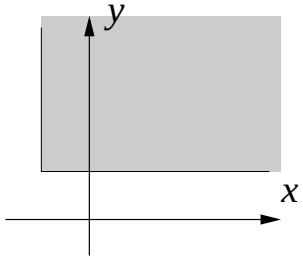
Example 2.1.17. How would you draw $\mathbb{R}^4$? You wouldn't. This does not prevent us from working with this (or much 'larger') sets. ⌟

Example 2.1.18. If A' is a subset of A, and B' is a subset of B, then you can view the product $A' \times B'$ as a subset of $A \times B$. For example, you can visualize the set $\mathbb{Z} \times \mathbb{Z}$ of ordered pairs of *integers* as a subset of $\mathbb{R} \times \mathbb{R}$: it is the set of pairs (x, y) for which both x and y are integers, and looks like this:



This is called the *integer lattice* in the plane. ⌟

Example 2.1.19. The subset $\{(x, y) \in \mathbb{R}^2 \mid -1 \leq x < \infty, 1 \leq y < \infty\}$ could be drawn like this:

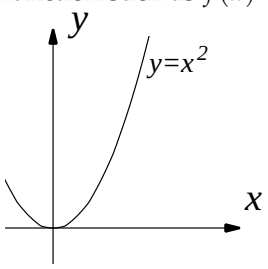


Note that this subset is itself a product: it can be viewed as the product $[-1, \infty) \times [1, \infty)$ of the two intervals $[-1, \infty)$ and $[1, \infty)$. ⌟

Example 2.1.20. On the other hand, many interesting subsets of a cartesian product are *not* themselves

do not distribute

products. In calculus, you would represent a function such as $f(x) = x^2$ by drawing its *graph:*



This is the set of points (x, y) such that $y = f(x)$ (or whatever part of this graph fits in your picture). In general, this is not a product. ⌟

In Chapter 3 we will discuss extensively *functions* at the level of generality taken in this note. Preview: a 'function' will in fact be defined in terms of its 'graph'.

2.1.6 *Homework*

▷ **2.1.1:** Prove that if A, B are both subsets of C, then $A \cup B$ is a subset of C.

▷ **2.1.2:** Prove that if A and B are sets, then $(A \smallsetminus B) \cup (A \cap B) = A$.

2.1.3: Let A and B be sets. If $A \cup B = A \cap B$, what can we conclude about A and B?

2.1.4: What can we conclude about A and B, if $A \smallsetminus B \subseteq B$? What if $A \smallsetminus B = B \smallsetminus A$?

2.1.5: Prove part (3) of Theorem 2.1.11, that is: $(A \cap B) \cap C = A \cap (B \cap C)$ for all sets A, B, C.

2.1.6: Prove parts (7) and (8) of Theorem 2.1.11.

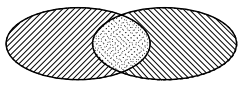
▷ **2.1.7:** State and prove versions of parts (7) and (8) of Theorem 2.1.11 for arbitrary families of sets.

▷ **2.1.8:** What is $\bigcap_{n\in\mathbb{N}}[n, \infty)$, and why? What about $\bigcap_{a\in(0,1)}(-a, a)$?

▷ **2.1.9:** Prove (2.5) and (2.6): Let $\{S_i\}_{i\in I}$, $\{S_j\}_{j\in J}$ be two families of sets, and prove that

$$\Big(\bigcap_{i\in I} S_i\Big) \cap \Big(\bigcap_{j\in J} S_j\Big) = \bigcap_{k\in I\cup J} S_k$$
$$\Big(\bigcup_{i\in I} S_i\Big) \cup \Big(\bigcup_{j\in J} S_j\Big) = \bigcup_{k\in I\cup J} S_k \quad .$$

2.1.10: Denote by $A + B$ the *symmetric difference* of the sets A and B, defined by $A + B = (A \cup B) \smallsetminus (A \cap B)$:



(The use of the symbol $+$ may surprise you, but there are good reasons to adopt it: this operation is 'commutative' and 'associative', as you will prove below; these properties are typical of operations we denote using a conventional $+$ sign.)

- Find the truth table for the logical connective corresponding to this set-theory operation.
- Using this definition for $+$, prove that $A + B = B + A$. (This is what makes the operation $+$ 'commutative'.)
- Using this definition for $+$, prove that $(A + B) + C = A + (B + C)$. (We say that $+$ is 'associative'.)
- Find a set O such that $A + O = A$ for all sets A.

do not distribute

- For any given set A, find a set B such that $A + B = O$, where O is the set you found in the previous question.

2.1.11: Let S be a fixed set. For subsets A, B of S, humor me and denote $A \cdot B$ the intersection $A \cap B$.

- Prove that $(A \cdot B) \cdot C = A \cdot (B \cdot C)$. (So $\cdot$ is an associative operation.)
- Find a set I such that $A \cdot I = A$ for all sets $A \subseteq S$.
- With $+$ denoted as in Exercise 2.1.10, prove that for all subsets A, B, C of S, $(A + B) \cdot C = A \cdot C + B \cdot C$. (Use a truth table.)

Combined with the results you proved in Exercise 2.1.10, this shows that the operations $+$, $\cdot$ you considered here satisfy several properties analogous to those satisfied by the conventional operations $+$ and $\cdot$ among (for example) integers. This makes the power set $\mathcal{P}(S)$ into a *ring*. In fact, $A \cdot B = B \cdot A$ (check!), so $\mathcal{P}(S)$ is a *commutative* ring. Rings are one of the main pillars of Abstract Algebra.

▷ **2.1.12:** Find a discussion of *Russell's paradox*, and understand it.

2.1.13: Prove that $A \times (B \cup C) = (A \times B) \cup (A \times C)$.

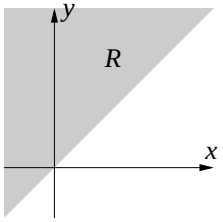
▷ **2.1.14:** Let $S = \{s_1, \ldots, s_n\}$ be a finite set, and let $S' = \{s_1, \ldots, s_{n-1}\}$. Explain in what sense $\mathcal{P}(S')$ is a subset of $\mathcal{P}(S)$. Describe the complement $\mathcal{P}(S) \setminus \mathcal{P}(S')$. How large is this set, compared to $\mathcal{P}(S')$? (Hint: Work this out for $n = 4$ first.)

▷ **2.1.15:** Let S be a finite set with $|S| = n$ elements. Prove that $|\mathcal{P}(S)| = 2^n$. (Hint: If A is a subset of S, color every element $s \in S$ Teal or Fuchsia according to whether $s \in A$ or $s \notin A$.) This is one reason why 2^S may be used as an alternative notation for $\mathcal{P}(S)$: so if S is a finite set, we get the reasonable-looking formula $|2^S| = 2^{|S|}$.

2.2 *Relations and induction*

2.2.1 *Definition of 'relation'*

Here is a drawing of another subset of $\mathbb{R} \times \mathbb{R}$, in the same style as the examples seen above:



More precisely, this is the set

$$R = \{(x, y) \mid x < y\}$$

(Note the font: this is R, a random name for a set; not $\mathbb{R}$, the set of real numbers.) The point I want to make here is that this set R contains precisely the same information as the 'relation' $x < y$: if you give me two real numbers x and y, I can figure out if the first is the smaller than the second by putting together the pair (x, y) and checking whether $(x, y) \in R$.

Intuitively, a 'relation' between elements a, b of two sets A, B is some kind of kinship between the elements, which may hold for some choices of $a \in A$ and $b \in B$ and not for others. Here we run into a difficulty similar

do not distribute

to the one we faced when we discussed logic: how do we extract the 'mathematical meaning' of something like kinship, relation, affinity, association, liaison...? In the case of the example given above, we know what $x < y$ 'means', so this is not a problem; but if we want to consider a completely general kind of relation, thinking in terms of 'meaning' injects an extraneous and possibly subjective component into what should just be a technical tool.

Truth tables were the solution to this problem when we dealt with logic: we managed to codify effectively our logical operations without having to think about 'meaning' at all. The solution here is analogous. Rather than thinking of a relation in terms of what it may mean, we just take at face value the fact that some elements a, b will be 'related', and some won't. The information in the 'relation' should be no more and no less than the information of which pairs $(a, b) \in A \times B$ are such that a is related with b.

Definition 2.2.1. Let A, B be two sets. A *relation* between elements of A and elements of B is a subset $R \subseteq A \times B$ of the cartesian product $A \times B$. We say that *'a is related to b with respect to R'*, or simply *'$a\,R\,b$'* if $(a, b) \in R$. We say that *a is not related to b* (with respect to R) if $(a, b) \notin R$, and we may write $a \not R b$.

If $A = B$, we say that the relation is 'defined on A'. ⌟

From this point of view, the name for the set R pictured above should be '$<$'. Admittedly, it would be a little awkward to use math symbols as names for sets[2], so we may use a fancy math symbol like $\sim$ to denote the way a relation acts: $a \sim b$; and a plain letter like R to denote the corresponding subset of $A \times B$. In this case $(a, b) \in R$ and $a \sim b$ are just two different ways to denote the same statement.

Example 2.2.2. Going back to finite examples, let $A = B = \{1, 2, 3, 4\}$, the set of integers between 1 and 4, inclusive. Then the relation on A that we would commonly denote $\geq$ is the subset

$$\{(1,1), (2,1), (3,1), (4,1), (2,2), (3,2), (4,2), (3,3), (4,3), (4,4)\} \subseteq A \times A$$

consisting of all pairs (a_1, a_2) with $a_1, a_2 \in A$ and $a_1 \geq a_2$. ⌟

The language of relations will allow us to formalize several important notions. In order to define these notions, we will need to consider several properties that relations may or may not satisfy.

Definition 2.2.3. Let A be a set. We say that a relation $R \subseteq A \times A$ is

- *reflexive* if $\forall a \in A$:
$$a\,R\,a \quad .$$
- *symmetric* if $\forall a \in A$, $\forall b \in A$:
$$(a\,R\,b) \implies (b\,R\,a) \quad .$$
- *antisymmetric* if $\forall a \in A$, $\forall b \in A$:
$$(a\,R\,b) \wedge (b\,R\,a) \implies (a = b) \quad .$$
- *transitive* if $\forall a \in A$, $\forall b \in A$, $\forall c \in A$:
$$(a\,R\,b) \wedge (b\,R\,c) \implies (a\,R\,c) \quad .$$

(Alternately, we could write that R is *reflexive* if $\forall a \in A, (a, a) \in R$; etc.). ⌟

Various combinations of these requirements (and more) define different types of relations. We will deal with *order* relation next (§2.2.2); *equivalence* relations (§2.3 and §2.4); and *functions* (Chapter 3).

[2] There's nothing wrong with that, actually.

do not distribute

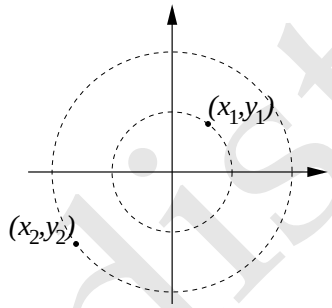
2.2.2 *Order relations*

Definition 2.2.4. An *order* relation is a relation that is reflexive, antisymmetric, and transitive. ⌟

Example 2.2.5. The relation $<$ on $\mathbb{R}$ (that is, the subset R defined at the beginning of this subsection) is transitive (because if x, y, z are real numbers, and $x < y$ and $y < z$, then $x < z$), and it is vacuously antisymmetric (because it never happens that $x < y$ and $y < x$ are both true, so the premise of the stated condition is always F); but it is not reflexive (because it is not true that $x < x$). So, maybe a little unfortunately, $<$ is *not* considered an order relation according to the definition given above. ⌟

Example 2.2.6. On the other hand, the relation $\leq$ (for example on the set $\mathbb{R}$ of real numbers, or on $\mathbb{Z}$, $\mathbb{Q}$, etc.) *is* an order relation. Indeed, $x \leq x$ for every x: that is the benefit obtained by adding the little extra $_$ underneath $<$. It is antisymmetric, and not vacuously: if it so happens that $x \leq y$ and $y \leq x$, then we may indeed conclude that $x = y$. And it is transitive: if $x \leq y$ and $y \leq z$, then indeed $x \leq z$. ⌟

Example 2.2.7. Define a relation $\preceq$ on $\mathbb{R}^2$ by declaring that $(x_1, y_1) \preceq (x_2, y_2)$ precisely when $x_1^2 + y_1^2 \leq x_2^2 + y_2^2$. Do you see what this means? (Think about it a moment.) $(x_1, y_1) \preceq (x_2, y_2)$ says that the distance from (x_1, y_1) to the origin is $\leq$ the distance from (x_2, y_2) to the origin: indeed, the distance from (x, y) to the origin is $\sqrt{x^2 + y^2}$.



This relation is reflexive and transitive; however, it is not an order relation, because it is not antisymmetric: for example,

$$(4, 3) \preceq (5, 0) \quad \text{and} \quad (5, 0) \preceq (4, 3)$$

(since $x^2 + y^2 = 25$ for both points) and yet $(4, 3) \neq (5, 0)$. ⌟

Definition 2.2.8. If $\preceq$ is an order relation on a set A, and $B \subseteq A$ is a subset, an element $m \in B$ is a *minimum* for B if $m \preceq b$ for every $b \in B$. An element $M \in B$ is a *maximum* for B if $b \preceq M$ for every $b \in B$. ⌟

A 'minimum' for an order relation $\preceq$ on A is a minimum for A with respect to $\preceq$. For example, the conventional $\leq$ relation on $\mathbb{Z}$ or $\mathbb{R}$ does not have a minimum (there isn't a 'least' integer or real number), while it does on $\mathbb{N}$: 0 is the 'smallest' natural number. Similarly, 0 is the minimum in the set $\mathbb{R}^{\geq 0}$ of nonnegative real numbers.

The set $\mathbb{N}^{>0}$ of *positive* natural numbers has a minimum: 1 is the smallest positive integer. On the other hand, the set $\mathbb{R}^{>0}$ of positive real numbers does *not* contain a minimum for $\leq$. Indeed, 0 is not an option since it is not positive, and if $r > 0$, then $r/2$ is smaller than r no matter how small r may be. You may be tempted to think that 0 should qualify as a minimum for this set anyway, but since 0 is not an element of the set we use a different name for it. We say that 0 is an 'infimum' (or 'greatest lower bound') for $\mathbb{R}^{>0}$ in $\mathbb{R}$. Similarly, 0 is a 'supremum' (or 'least upper bound') for $\mathbb{R}^{<0}$. But that is a story for another time—we will come back to these notions in Chapter 4.

An *equivalence relation* is a relation that is *reflexive, symmetric, and transitive.* This is a very important notion,

do not distribute

almost ubiquitous in mathematics. We will spend quite a bit of time exploring it, later in this chapter.

2.2.3 *The well-ordering principle and induction*

We will come back to set theory and relations very soon, but first we take a moment to cover another powerful 'technique of proof', which I left out of the early discussion in §1.3.3 since it required more confidence with the basic language of sets. This technique is based on the notion of 'well-ordered sets', and on the fact that $\mathbb{N}$ is 'well-ordered' by $\leq$. What does this mean?

The conventional order relation $\leq$ has a minimum (that is, 0) on both $\mathbb{N}$ and $\mathbb{R}^{\geq 0}$, but there are important differences in these two situations. While *not every* nonempty subset of $\mathbb{R}^{\geq 0}$ has a minimum with respect to $\leq$ (for example, $\mathbb{R}^{>0}$ does not have a minimum, as we saw at the end of §2.2), *every* nonempty subset of $\mathbb{N}$ does have a minimum.

Definition 2.2.9. An order relation $\preceq$ on a set A defines a *well-ordering* on A if every nonempty subset of A has a minimum with respect to $\preceq$. ⌟

Thus, we say that $\mathbb{N}$ is 'well-ordered' by $\leq$, while $\mathbb{R}^{\geq 0}$ is not. In fact, the first observation is so important that it deserves a name:

Claim 2.2.10 (Well-ordering principle)**.** *The set $\mathbb{N}$ is well-ordered by $\leq$. That is, every nonempty subset of $\mathbb{N}$ has a minimum with respect to the relation $\leq$.*

Our familiarity with $\mathbb{N}$ should make this fact appear very reasonable. In fact, if one is setting up arithmetic from scratch and has to *define* natural numbers (for example as 'cardinalities' of finite sets, about which we will learn in Chapter 5), then this is something that should be proven rigorously; in certain approaches, it must be written in as one of the key axioms defining $\mathbb{N}$. I am not going to attempt any such foundational work here; I will just take the well-ordering principle for granted. And by the way, you will encounter the well-ordering principle again when you tackle abstract algebra, particularly when you study the algebraic properties of $\mathbb{Z}$.

Granting that we believe Claim 2.2.10, we can extract from it the powerful new technique of proof mentioned above. This is called 'induction'. Technically, it is based on the following consequence of the well-ordering principle:

Proposition 2.2.11. *Let $S \subseteq \mathbb{N}$ be a subset of $\mathbb{N}$. Assume that*

- *$0 \in S$;*
- *If $n \in S$, then $n + 1 \in S$.*

Then $S = \mathbb{N}$.

Please pause a moment and appreciate what this statement is saying; once you understand what the statement is claiming, it will likely appear completely obvious to you. However, 'obvious' is a forbidden word in mathematics, and we have to understand *why* something like this is true; it turns out to be a straightforward consequence of Claim 2.2.10.

Proof. Let $T = \mathbb{N} \setminus S$ be the complement of S. We have to prove that $T = \emptyset$. Arguing by contradiction, assume that $T \neq \emptyset$. By the well-ordering principle, T has a least element $t \in T$.

By hypothesis, $0 \in S$: so $0 \notin T$, and this tells us $t > 0$. But then we can consider the natural number $n = t - 1$. Since t is the least element in T, and $n < t$, we have $n \in S$. On the other hand, by the second

do not distribute

hypothesis in the proposition,

$$n \in S \implies n+1 \in S \quad . \tag{2.7}$$

Since $t = n+1$, this would mean that $t \in S$. This is a contradiction, since $t \in T$ and t cannot belong both to S and its complement T.

Thus, the assumption $T \neq \emptyset$ leads to a contradiction. It follows that $T = \emptyset$, hence $S = \mathbb{N}$, as needed. □

More generally, (Exercise 2.2.7), if $S \subseteq \mathbb{N}$ contains a 'seed' k, and satisfies the hypothesis (2.7) appearing in Proposition 2.2.11, then we can conclude that S contains all numbers larger than or equal to k.

Why should these statements look 'obvious'? If the seed k is 0, as in Proposition 2.2.11, start from $0 \in S$; by the hypothesis

$$n \in S \implies n+1 \in S \quad ,$$

this implies $0+1 \in S$, so $1 \in S$. By the hypothesis

$$n \in S \implies n+1 \in S \quad ,$$

this implies $1+1 \in S$, so $2 \in S$. By the hypothesis

$$n \in S \implies n+1 \in S \quad ,$$

this implies $2+1 \in S$, so $3 \in S$. By the hypothesis

$$n \in S \implies n+1 \in S \quad ,$$

this implies $3+1 \in S$, so $4 \in S$. By the hypothesis

$$n \in S \implies n+1 \in S \quad ,$$

this implies $4+1 \in S$, so $5 \in S$. By the hypothesis

$$n \in S \implies n+1 \in S \quad ,$$

this implies $5+1 \in S$, so $6 \in S$. Etc., etc., etc.: if we apply this reasoning enough times, we will prove that $1000 \in S$, and $1000000 \in S$, and $10000000000000000 \in S$, and so on *ad libitum.* The problem with this is that, while it seems reasonable that it should be so (that's why it seems 'obvious' once you understand what we are talking about), we cannot *really* apply this reasoning 10000000000000000 times, can we? The real problem is that our box of tools to infer logical consequences is limited to the logical connectives $\wedge$, $\vee$, etc., and they do not include a tool to apply a reasoning 'arbitrarily many' times. What Proposition 2.2.11 shows is that the well-ordering principle gives us just such a tool. We do not need to explicitly apply the key implication '$n \in S \implies n+1 \in S$' a million times to prove that $1000000 \in S$; the implication itself and the well-ordering principle prove that $S = \mathbb{N}$ once and for all (provided $0 \in S$ to begin with). *All* numbers are shown to be in S, in one shot.

What does this have to do with 'techniques of proof'?

Suppose you are trying to prove a statement for all integers, such as the validity of some formula. Assume that you know that

- (*Seed.*) The formula holds for $n = 0$; and
- (*Induction.*) If the formula holds for a fixed (but otherwise arbitrary) number n, then it necessarily holds for $n+1$.

do not distribute

Then you can conclude that the formula holds for all $n \in \mathbb{N}$.

Indeed: Let S be the set of numbers for which the formula holds; then S satisfies the hypotheses of Proposition 2.2.11, therefore $S = \mathbb{N}$, that is, the formula holds for all numbers.

Proofs constructed in this way are called *'Proofs by induction'*.

Note that in the induction hypothesis (2.7) no one is choosing n: you have to verify the induction hypothesis for a general, unspecified n. This usually involves some abstract argument, often an algebraic manipulation of some kind.

2.2.4 *Two examples*

Example 2.2.12. Consider the following equality, which I will call $P(N)$ for reference:

$$P(N) \quad : \quad 1+2+3+\cdots+N = \frac{(N+1)N}{2} \quad .$$

The proof of this formula is a classical illustration of induction, and the formula itself is also somewhat useful. (It is used, for example, to prove that the integral of x is $\frac{x^2}{2}$ by applying the definition of Riemann integral.) To focus ideas, note that for $N = 0$ there are no summands on the left-hand side, so the left-hand side equals 0 for $N = 0$. So $P(0)$ is the assertion that

$$0 = \frac{(0+1)\cdot 0}{2} \quad .$$

Similarly, $P(1)$ states that

$$1 = \frac{(1+1)\cdot 1}{2} \quad ,$$

and $P(2)$ is the formula

$$1+2 = \frac{(2+1)\cdot 2}{2} \quad ,$$

and so on.

Claim 2.2.13. *$P(N)$ is true for all $N \geq 0$.*

Proof. To prove this statement by induction, we have to verify that

- $P(0)$ is true;
- for all $n \geq 0$, the truth of $P(n)$ implies the truth of $P(n+1)$.

The first point is immediate: as we have seen above, $P(0)$ states that $0 = \frac{(0+1)\cdot 0}{2}$, which indeed is the case. So $P(0)$ is true.

The second point is much more interesting. Let $n \geq 0$ be a fixed, but unspecified number, for which formula $P(n)$ happens to hold. Formula $P(n+1)$ deals with the sum of the first $n+1$ integers; write that sum as follows:

$$1+2+\cdots+n+(n+1) = (1+2+\cdots+n)+(n+1) \quad .$$

The first summand on the right-hand side is precisely the left-hand side of $P(n)$. Since we are assuming that $P(n)$ is true for this specific n, it follows that

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

do not distribute

This is what the assumption $P(n)$ implies. Now we perform a little algebra on the right-hand side:

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

The conclusion is that assuming that $P(n)$ is true, we obtain

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

But this new formula is *precisely* $P(n+1)$. Thus we have verified that the truth of $P(n)$ implies the truth of $P(n+1)$, i.e., the second point listed at the beginning of the proof.

We are done: $P(N)$ is true for all $N \geq 0$, *by induction.* □

There are other ways to convince oneself that the given formula is true. For example,

$$\begin{aligned} 2(1+2+3+\cdots+N) &= (N + (N-1) + (N-2) + \cdots + 1) + (1+2+3+\cdots+N) \\ &= (N+1) + ((N-1)+2) + ((N-2)+3) + \cdots + (1+N) \\ &= (N+1) + (N+1) + (N+1) + \cdots + (N+1) \\ &= N(N+1) \quad ; \end{aligned}$$

the formula follows by dividing both sides by 2.

This looks clever (it allegedly occurred to Gauss at age 3). Induction is not particularly clever, and works just as well. ⌟

Example 2.2.14. Another classical application of induction is the 'binomial theorem'. It is a very basic fact, which should be in everyone's toolbox. At this point we know enough to see precisely why it is true.

The binomial theorem will use numbers denoted as follows:

$$\binom{n}{k}$$

for n, k integers and[3] $n \geq 0$. The number $\binom{n}{k}$ is assumed to be 0 if $k < 0$ or $k > n$; if $0 \leq k \leq n$, it is given by the formula

$$\binom{n}{k} = \frac{n!}{k!(n-k)!} \quad ,$$

where the ! sign denotes 'factorial': $n! = 1 \cdots n$. (And we set $0! = 1$. Can you think of why this is a reasonable convention?)

There is a nice interpretation of these numbers: $\binom{n}{k}$ is *the number of subsets with k elements of a set with n elements.* (Why? Exercise 2.2.13!) For example, $\binom{4}{2} = \frac{4!}{2!2!} = \frac{24}{4} = 6$, and here are the 6 subsets with 2 elements in the set $\{1,2,3,4\}$:

$$\{1,2\} \quad , \quad \{1,3\} \quad , \quad \{1,4\} \quad , \quad \{2,3\} \quad , \quad \{2,4\} \quad , \quad \{3,4\} \quad .$$

This set-theoretic interpretation may be used to give a 'combinatorial' proof of the following formula; I'll take a more algebraic approach.

[3] There are several interesting generalizations allowing for negative n and also extending the notation to numbers other than integers or to polynomials.

do not distribute

Lemma 2.2.15. *For all $n \geq 0$ and all k,*

$$\binom{n+1}{k} = \binom{n}{k-1} + \binom{n}{k} \quad .$$

Proof. The formula reduces to $0 = 0$ for $k < 0$ or $k > n+1$ and to $1 = 1$ for $k = 0$ or $k = n+1$. For $1 \leq k \leq n$, it is verified by simple algebra of factorials, which you are well advised to refresh if it seems rusty:

$$\begin{aligned}
\binom{n}{k-1} + \binom{n}{k} &= \frac{n!}{(k-1)!(n-(k-1))!} + \frac{n!}{k!(n-k)!} \\
&= \frac{n!}{(k-1)!(n-k)!}\left(\frac{1}{n-k+1} + \frac{1}{k}\right) \\
&= \frac{n!}{(k-1)!(n-k)!} \cdot \frac{k+n-k+1}{k \cdot (n-k+1)} \\
&= \frac{(n+1)!}{k!(n+1-k)!} = \binom{n+1}{k}
\end{aligned}$$

as claimed. □

As a consequence of Lemma 2.2.15, the numbers $\binom{n}{k}$ can be computed 'recursively'. Arrange them in a triangular array, as follows:

$$\begin{array}{ccccccccccc}
 & & & & \binom{0}{0} & & & & & & \\
 & & & \binom{1}{0} & & \binom{1}{1} & & & & & \\
 & & \binom{2}{0} & & \binom{2}{1} & & \binom{2}{2} & & & & \\
 & \binom{3}{0} & & \binom{3}{1} & & \binom{3}{2} & & \binom{3}{3} & & & \\
\binom{4}{0} & & \binom{4}{1} & & \binom{4}{2} & & \binom{4}{3} & & \binom{4}{4} & & \\
\cdots & \cdots & & \cdots & & \cdots & & \cdots & & \cdots &
\end{array}$$

Since $\binom{n}{0} = \binom{n}{n} = 1$, the two visible sides of this triangle consist of 1's:

$$\begin{array}{ccccccccccc}
 & & & & 1 & & & & & & \\
 & & & 1 & & 1 & & & & & \\
 & & 1 & & \binom{2}{1} & & 1 & & & & \\
 & 1 & & \binom{3}{1} & & \binom{3}{2} & & 1 & & & \\
1 & & \binom{4}{1} & & \binom{4}{2} & & \binom{4}{3} & & 1 & & \\
\cdots & \cdots & & \cdots & & \cdots & & \cdots & & \cdots &
\end{array}$$

All the other numbers can now be filled recursively: they are arranged so that $\binom{n+1}{k}$ sits right underneath $\binom{n}{k-1}$ and $\binom{n}{k}$:

$$\begin{array}{ccccccc}
 & \cdots & & \cdots & & \cdots & \\
\cdots & & \binom{n}{k-1} & & \binom{n}{k} & & \cdots \\
 & \cdots & & \binom{n+1}{k} & & \cdots & \\
 & & \cdots & & \cdots & &
\end{array}$$

do not distribute

and Lemma 2.2.15 tells us that $\binom{n+1}{k}$ is the sum of $\binom{n}{k-1}$ and $\binom{n}{k}$. Therefore, every number in the triangular grid is the sum of the two numbers right above it. Applying this observation over and over and over produces the following grid:

$$
\begin{array}{ccccccccccccc}
 & & & & & & 1 & & & & & & \\
 & & & & & 1 & & 1 & & & & & \\
 & & & & 1 & & 2 & & 1 & & & & \\
 & & & 1 & & 3 & & 3 & & 1 & & & \\
 & & 1 & & 4 & & 6 & & 4 & & 1 & & \\
 & 1 & & 5 & & 10 & & \boxed{10} & & \boxed{5} & & 1 & \\
1 & & 6 & & 15 & & 20 & & \boxed{15} & & 6 & & 1 \\
\ldots & & & & \ldots & & \ldots & & & & \ldots & &
\end{array}
$$

For example, $15 = \binom{6}{4}$ is the sum of $10 = \binom{5}{3}$ and $5 = \binom{5}{4}$.

This arrangement of numbers is called *Pascal's triangle.*

You were likely already familiar with the grid I drew above, and you know that the numbers that appear in each row are the numbers needed to expand the power of a 'binomial'; for instance,

$$(x+y)^3 = x^3 + 3\,x^2y + 3\,xy^2 + y^3$$
$$(x+y)^4 = x^4 + 4\,xy^3 + 6\,x^2y^2 + 4\,xy^3 + y^4$$

and so on. Why is this true?

Theorem 2.2.16 (Binomial theorem)**.** *For all $n \geq 0$,*

$$(x+y)^n = \sum_{k=0}^{n} \binom{n}{k} x^k y^{n-k} \quad .$$

Because of this result, the numbers $\binom{n}{k}$ are usually called *'binomial coefficients'*.

The proof of the binomial theorem is a prime example of a proof by induction.

Proof. For $n = 0$, the formula states that

$$(x+y)^0 = \binom{0}{0} x^0 y^0$$

which is true: $1 = 1$. By induction, in order to prove the formula for all $n \geq 0$, I need to verify that if the formula is true for a *fixed* (but unspecified) $n \geq 0$, then it is also true for $n+1$. That is, I may assume that I know

$$(x+y)^n = \sum_{k=0}^{n} \binom{n}{k} x^k y^{n-k}$$

for a certain (unspecified) $n \geq 0$, and I have to verify that I can then deduce that

$$(x+y)^{n+1} = \sum_{k=0}^{n+1} \binom{n+1}{k} x^k y^{n+1-k}$$

for the same n. I will work on the right-hand side, and prove it equals the left-hand side. Use Lemma 2.2.15

do not distribute

to rewrite the right-hand side:

$$\sum_{k=0}^{n+1}\binom{n+1}{k}x^k y^{n+1-k} = \sum_{k=0}^{n+1}\left(\binom{n}{k-1}+\binom{n}{k}\right)x^k y^{n+1-k}$$
$$= \sum_{k=0}^{n+1}\binom{n}{k-1}x^k y^{n+1-k} + \sum_{k=0}^{n+1}\binom{n}{k}x^k y^{n+1-k} \quad .$$

As k goes from 0 to $n+1$, $k-1$ goes from -1 to n. I use this to rewrite the first sum, letting $\ell = k-1$:

$$\sum_{k=0}^{n+1}\binom{n+1}{k}x^k y^{n+1-k} = \sum_{\ell=-1}^{n}\binom{n}{\ell}x^{\ell+1} y^{n-\ell} + \sum_{k=0}^{n+1}\binom{n}{k}x^k y^{n+1-k} \quad .$$

Now note that $\binom{n}{\ell} = 0$ for $\ell = -1$, and $\binom{n}{k} = 0$ for $k = n+1$. Taking this into account, and replacing again ℓ by k, we get

$$\sum_{k=0}^{n+1}\binom{n+1}{k}x^k y^{n+1-k} = \sum_{k=0}^{n}\binom{n}{k}x^{k+1} y^{n-k} + \sum_{k=0}^{n}\binom{n}{k}x^k y^{n+1-k} \quad .$$

Next, note that $k+1 \geq 1$ in the first sum, and $n+1-k \geq 1$ in the second one. Thus, we can factor a variable out of each summand:

$$\sum_{k=0}^{n+1}\binom{n+1}{k}x^k y^{n+1-k} = x\left(\sum_{k=0}^{n}\binom{n}{k}x^k y^{n-k}\right) + y\left(\sum_{k=0}^{n}\binom{n}{k}x^k y^{n-k}\right) \quad .$$

And now the induction hypothesis tells us that the two terms in parenthesis on the right-hand-side both equal $(x+y)^n$, and this allows us to complete the proof:

$$\sum_{k=0}^{n+1}\binom{n+1}{k}x^k y^{n+1-k} = x\,(x+y)^n + y\,(x+y)^n = (x+y)(x+y)^n = (x+y)^{n+1} \quad ,$$

since this verifies the induction step. □

I spelled out this proof in great detail. You should go through it and practice these manipulations until you see that they are *all immediate:* in a self-respecting textbook in algebra or analysis, a proof like this would surely be omitted and left to the reader. On p. 183 of Lang's *Algebra* you can find a statement about the factoriality of polynomial rings over a factorial ring, and about the units and prime elements of such rings. (Whatever this all means.) The proof given in that text is:

Proof. Induction. □

This is a bit extreme, but not completely atypical. ⌟

Remark 2.2.17. Many amazing identities involving the numbers $\binom{n}{k}$ follow very easily from the binomial theorem. For example, add up the numbers on any line of Pascal's triangle, and you will get a power of 2:

$$1+6+15+20+15+6+1 = 64 = 2^6 \quad .$$

A conceptual explanation for this fact is that since $\binom{n}{k}$ equals the number of subsets of n with k elements (Exercise 2.2.13), this sum counts the number of *all* subsets of a set with n elements. So the sum must equal 2^n since this is the number of elements of $\mathcal{P}(S)$, as you verified in Exercise 2.1.15. The binomial theorem gives

do not distribute

an easier argument, because it does not require us to know anything sophisticated about the numbers $\binom{n}{k}$ or to know about $\mathcal{P}(S)$: just let $x = y = 1$ in the binomial theorem, and you get

$$2^n = (1+1)^n = \sum_{k=0}^{n} \binom{n}{k} 1^k \cdot 1^{n-k} = \sum_{k=0}^{n} \binom{n}{k} \quad .$$

Similarly, look at this:

$$1 + 4 \cdot 2 + 6 \cdot 2^2 + 4 \cdot 2^3 + 1 \cdot 2^4 = 1 + 8 + 24 + 32 + 16 = 81 = 3^4 \quad .$$

Do you see why this must work for every line of Pascal's triangle? Think about it a moment before reading on...

Here is a proof of the corresponding general formula, as an immediate consequence of the binomial theorem:

$$3^n = (2+1)^n = \sum_{k=0}^{n} \binom{n}{k} 2^k \quad .$$

You should try other examples like this—you will discover many other amazing identities. ⌟

2.2.5 *Homework*

2.2.1: Let A be a set consisting of exactly 4 elements. How many different relations on A are there?

2.2.2: Let S be a set, and let $A = \mathcal{P}(S)$ be its power set. Consider the 'inclusion' relation on subsets of S: so this is a relation defined on A, and two subsets U, V are related when $U \subseteq V$. What properties does this relation satisfy, among those listed in Definition 2.2.3?

2.2.3: Let $\preceq$ be an order relation on a set A. Prove that if a minimum a exists for $\preceq$, then it is unique. That is, prove that if a_1 and a_2 are both minima, then necessarily $a_1 = a_2$.

▷ **2.2.4:** Let $\sim_A$ be a relation on a set A, and let B be a subset of A. Define a relation $\sim_B$ on B using the information of $\sim_A$. (There is only one reasonable way to do this! We then say that $\sim_B$ is 'induced' by $\sim_A$.) If $\sim_A$ is reflexive, is $\sim_B$ necessarily reflexive? (Why/Why not?) Study the same question for the other standard properties of relations.

2.2.5: Let $\preceq$ be a well-ordering on a set S. Prove that it induces a well-ordering on every subset of S. (See Exercise 2.2.4 for the meaning of 'induced' relation.)

2.2.6: An order relation $\preceq$ on a set S is a 'total' ordering if for every two elements a and b in S, either $a \preceq b$ or $b \preceq a$.

- Give an example of an order relation that is not total.
- Give an example of a total ordering that is not a well-ordering.
- Prove that every well-ordering is total.

▷ **2.2.7:** Use the well-ordering principle to show that if $S \subseteq \mathbb{N}$ is a subset containing a given number $k \in \mathbb{N}$, and such that $n \in S \implies n+1 \in S$ for all $n \geq k$, then S contains all numbers $\geq k$.

2.2.8: Prove by induction, or disprove, that $10^n - 1$ is a multiple of 3 for all $n \geq 0$.

2.2.9: Prove by induction that if S has n elements, then $|\mathcal{P}(S)| = 2^n$; this will reproduce the result of Exercise 2.1.15. (Hint: In Exercise 2.1.14 you should have verified that if S' is obtained from S by removing one element, then $|\mathcal{P}(S')| = |\mathcal{P}(S) \smallsetminus \mathcal{P}(S')|$.)

do not distribute

2.2.10: Prove by induction, or disprove:

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

for all $n \geq 0$.

2.2.11: Prove by induction, or disprove, that $n^2 - n + 41$ is a *prime* number for all $n \geq 0$.

2.2.12: What is wrong with the following proof that all cats have the same color?

I am going to argue that if $n \geq 1$, and I have a set of n cats, then all of the cats in that set have the same color. If $n = 1$, then there is only one cat in the set, so the fact is trivially true in this case. Assuming that the fact has been proven for n cats, I am going to prove it for $n + 1$ cats. Let C be a set with $n + 1$ cats, and let a, b be any two distinct elements of C. I have to prove that a and b have the same color. Let c be another cat in C, different from a and b. I am going to show that the color of a is the same as the color of c, and that the color of c is the same as the color of b, and this will show that a, b have the same color. To see that a and c have the same color, consider the set $C \setminus \{b\}$. This is a set of n cats, and contains both a and c. By the induction hypothesis, all the cats in this set have the same color, so a and c have the same color. Applying the same reasoning to $C \setminus \{a\}$ shows that b and c have the same color. It follows that a and b have the same color, as claimed.

▷ **2.2.13:** Let $C(n,k)$ be the number of subsets with k elements of a set with n elements. Prove by induction that $C(n,k) = \binom{n}{k}$. (Hint: First prove that $C(n+1,k) = C(n,k-1) + C(n,k)$; then use this observation, Lemma 2.2.15, and induction to show that $C(n,k) = \binom{n}{k}$.)

2.2.14: Give a 'combinatorial' proof of the result of Exercise 2.2.13, using the following guide. You have to find the number of ways to choose k elements out of n. How many choices do you have for the first element? For the second? For the third? ...Once you have chosen your k elements, you have to count as one all choices that only differ by the order of the k elements. Therefore...

2.2.15: Explain why adding up every other number on a line of Pascal's triangle gives a power of 2. For example, from the line for $n = 6$:

$$1 + 0 + 15 + 0 + 15 + 0 + 1 = 32 = 2^5 \quad .$$

(Use the binomial theorem!)

2.2.16: Prove that for $n > 0$

$$\sum_{k=0}^{n} k\binom{n}{k} = n \cdot 2^{n-1} \quad .$$

For example, from the line for $n = 6$:

$$0 \cdot 1 + 1 \cdot 6 + 2 \cdot 15 + 3 \cdot 20 + 4 \cdot 15 + 5 \cdot 6 + 6 \cdot 1 = 192 = 6 \cdot 2^5 \quad .$$

(Hint: Take a derivative.)

▷ **2.2.17:** Put your calculus hat on for this exercise. Let f, g be real functions of one variables that can be differentiated infinitely many times, and denote by $f^{(k)}$ the k-th derivative of f. (For example, $f^{(0)} = f$, $f^{(1)} = f'$, $f^{(2)} = f''$.) Prove the following formula:

$$(fg)^{(n)} = \sum_{k=0}^{n} \binom{n}{k} f^{(k)} g^{(n-k)} \quad .$$

for all integers $n \geq 0$.

do not distribute

2.3 *Equivalence relations and partitions.*

2.3.1 *Equivalence relations*

The notion of *order* relation led us to consider the *well-ordering principle,* which we have recognized as the underpinning of a powerful technique of proof, called *induction.* This was a little scenic detour; we are now ready to get back to the main road and resume the discussion of interesting types of relations.

Among all relations, *equivalence* relations are especially important; their definition was mentioned in passing at the end of §2.2.2 and will be recalled below. Many constructions in mathematics, ranging from algebra to analysis to geometry, stem from considering the 'quotient' of a set with respect to an equivalence relation. We are going to define and study this notion in the rest of this chapter. This construction is so pervasive that this topic is one of the most important covered in this text; if you can only retain *one* notion among all the ones we cover in this text, let it be the notion of 'quotient by an equivalence relation'.

The prototypical equivalence relation is equality: $=$. Equality is clearly reflexive ($a = a$), symmetric (if $a = b$, then $b = a$), and transitive (if $a = b$ and $b = c$, then $a = c$). These properties are abstracted in the definition of equivalence relation.

Definition 2.3.1. Let A be a set, and let $\sim$ be a relation defined on A. We say that $\sim$ is an *equivalence relation* if $\sim$ is reflexive, symmetric, and transitive. ⌟

That is, $\sim$ is an equivalence relation if $\forall a, b, c \in A$:

- $a \sim a$;
- $(a \sim b) \implies (b \sim a)$;
- $((a \sim b) \wedge (b \sim c)) \implies (a \sim c)$.

While $=$ is an equivalence relation as observed ab, it is by no means the only example of an equivalence relation.

Example 2.3.2. Consider the relation $\sim$ defined on $\mathbb{Z}$ by

$$a \sim b \iff b - a \text{ is an even number.}$$

For example: $3 \sim 7$, while $5 \not\sim 10$. I claim that this is an equivalence relation. Indeed:

- $\forall a \in \mathbb{Z}$, $a - a = 0$ is even; therefore, $a \sim a$: $\sim$ is **reflexive.**
- $\forall a, b \in \mathbb{Z}$, if $b - a = 2k$ is even then so is $a - b = 2(-k)$; therefore, if $a \sim b$, then $b \sim a$: $\sim$ is **symmetric.**
- $\forall a, b, c \in \mathbb{Z}$, if $b - a = 2k$ is even and $c - b = 2\ell$ is even, then so is $c - a = (c - b) + (b - a) = 2(k + \ell)$; therefore, if $a \sim b$ and $b \sim c$, then $a \sim c$: $\sim$ is **transitive.**

Here is an alternative way to describe the same relation: $a \sim b$ if a and b are both even or both odd, and $a \not\sim b$ if one of the two numbers is even and the other is odd.

Yet another way: $a \sim b$ precisely when the 'remainder after division by 2' is the same for both a and b. You will define a more general equivalence relation in the same style in Exercise 2.3.5. ⌟

Example 2.3.3. Consider the relation $\sim$ defined on $\mathbb{R}$ by

$$a \sim b \iff b - a \in \mathbb{Z} \quad .$$

This is an equivalence relation. For example, let's verify transitivity: let a, b, c be real numbers, and assume $a \sim b$ and $b \sim c$. Then $m = b - a$ and $n = c - b$ are both integers. It follows that

$$c - a = (c - b) + (b - a) = n + m$$

do not distribute

is also an integer, and hence we conclude that $a \sim c$, as needed. Reflexivity and symmetry are (even) easier to verify. ⌟

Example 2.3.4. The relation $\leq$ on $\mathbb{R}$ is *not* an equivalence relation. Indeed, it is not symmetric: for example, $2 \leq 3$, and yet $3 \not\leq 2$. ⌟

Example 2.3.5. The relation $\preceq$ on $\mathbb{R}^2$ considered in Example 2.2.7 is also not an equivalence relation, again because it is not symmetric. On the other hand, consider the relation $\sim$ on $\mathbb{R}^2$ defined by

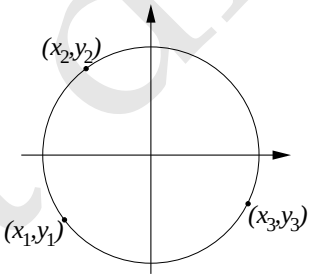
$$(x_1, y_1) \sim (x_2, y_2) \quad \Longleftrightarrow \quad x_1^2 + y_1^2 = x_2^2 + y_2^2 \quad .$$

Then

- $\forall (x, y) \in \mathbb{R}^2$, $x^2 + y^2 = x^2 + y^2$, and hence $(x, y) \sim (x, y)$: $\sim$ is **reflexive**.
- $\forall (x_1, y_1), (x_2, y_2) \in \mathbb{R}^2$, if $x_1^2 + y_1^2 = x_2^2 + y_2^2$, then $x_2^2 + y_2^2 = x_1^2 + y_1^2$; therefore, if $(x_1, y_1) \sim (x_2, y_2)$, then $(x_2, y_2) \sim (x_1, y_1)$. Therefore, $\sim$ is **symmetric**.
- $\forall (x_1, y_1), (x_2, y_2), (x_3, y_3) \in \mathbb{R}^2$, if $x_1^2 + y_1^2 = x_2^2 + y_2^2$ and $x_2^2 + y_2^2 = x_3^2 + y_3^2$, then $x_1^2 + y_1^2 = x_3^2 + y_3^2$; therefore, if $(x_1, y_1) \sim (x_2, y_2)$ and $(x_2, y_2) \sim (x_3, y_3)$, we can conclude that $(x_1, y_1) \sim (x_3, y_3)$. Therefore, $\sim$ is **transitive**.

This tells us that $\sim$ *is* an equivalence relation. ⌟

If you are like me, you do not particularly enjoy reading such verifications; this last one is particularly dense from the notational point of view. For an alternative viewpoint, notice that $(x_1, y_1) \sim (x_2, y_2)$ precisely when (x_1, y_1), (x_2, y_2) are on the same circle centered at the origin. With this in mind, think again through the verification that $\sim$ is an equivalence relation, and you should be able to carry it out without all the awkward notation. For example, transitivity amounts to the fact that if (x_1, y_1) and (x_2, y_2) are on the same circle, and (x_2, y_2) and (x_3, y_3) are on the same circle, then (x_1, y_1) and (x_3, y_3) are on the same circle. Big deal.



This is a useful viewpoint: it views an equivalence relation in terms of the *equivalence classes* it determines; that will be the topic of the next subsection. As we will see, thinking in terms of 'equivalence classes' clarifies the notion of equivalence relation considerably. For example: How many different equivalence relations are there on a set with 3 elements? If the set is $\{a, b, c\}$, you could try to count all equivalence relations such that a is only in relation with itself, those for which a is in relation with itself and exactly another element, and so on. This can be carried out, but will likely look somewhat complicated. I promise that once you absorb the material that we are going to explore, this question will look very easy (but you should try it right now, before I spill the beans).

2.3.2 *Definition of equivalence classes*

The circle drawn above consists of all the elements of the set ($\mathbb{R}^2$ in that example) which are in relation with each other: if two points (x_1, y_1) and (x_2, y_2) are on the same circle centered at the origin, then

do not distribute

$(x_1, y_1) \sim (x_2, y_2)$, and conversely. We abstract this example with the notion of 'equivalence class'. Here is the official definition, for any set and any equivalence relation:

Definition 2.3.6. Let $\sim$ be an equivalence relation on a set A, and let $a \in A$. The subset of A defined by

$$[a]_\sim = \{b \in A \mid a \sim b\}$$

is called the *equivalence class* of a (with respect to, or 'modulo' $\sim$). The element a is called a *representative* of the class $[a]_\sim$. ⌟

The notation $[a]_\sim$ is one of several possibilities. The subscript $\sim$ may be omitted if it is clear from context. In this case, it is not uncommon to denote the equivalence class of a by just decorating a in some way, e.g. $\overline{a}$, $\underline{a}$, $\hat{a}$, $\tilde{a}$, ... For now I'll try to be consistent and use $[a]_\sim$.

Keep in mind that

$$a \sim b \quad \iff \quad b \in [a]_\sim \quad :$$

this is what the definition of equivalence class says.

In fact, two elements a, b are related by $\sim$ precisely when the two equivalence classes coincide. This deserves to be highlighted and proved formally.

Lemma 2.3.7. *Let A be a set, and let $\sim$ be an equivalence relation on A. Then $[a]_\sim = [b]_\sim$ if and only if $a \sim b$.*

The (easy) proof of this lemma will use everything we know about equivalence relations: their defining properties (reflexive, symmetric, transitive) all play a role in this proof.

Proof. $\implies$) Assume $[a]_\sim = [b]_\sim$. Since $\sim$ is an equivalence relation, it is reflexive; thus, $b \sim b$, and this implies $b \in [b]_\sim$ by definition of equivalence class. Since $[a]_\sim = [b]_\sim$ by assumption, and $b \in [b]_\sim$, we know that $b \in [a]_\sim$. We can then conclude that $a \sim b$, again by definition of equivalence class.

$\impliedby$) Now assume that $a \sim b$. I will first prove that $[b]_\sim \subseteq [a]_\sim$. For this, let $c \in [b]_\sim$. Then $b \sim c$ by definition of equivalence class. Since $a \sim b$ and $b \sim c$, we know that $a \sim c$, because $\sim$ is transitive as it is an equivalence relation. Since $a \sim c$, we can conclude that $c \in [a]_\sim$, once more by definition of equivalence class.

This shows that if $a \sim b$, then $[b]_\sim \subseteq [a]_\sim$.

On the other hand, $\sim$ is symmetric, since it is an equivalence relation. Thus, as $a \sim b$, we know that $b \sim a$. By the same argument given above, this implies that $[a]_\sim \subseteq [b]_\sim$.

Finally, since $[b]_\sim \subseteq [a]_\sim$ and $[a]_\sim \subseteq [b]_\sim$, we can conclude that $[a]_\sim = [b]_\sim$, by Proposition 2.1.3. □

This is a potentially confusing point, so let me stress what is going on here: $[a]_\sim = [b]_\sim$ does *not* mean or imply that $a = b$. What it means is that $a \sim b$. *One* equivalence class will in general have *many* representatives. If all you know about an equivalence class $[a]_\sim$ is the equivalence *class* itself, you have in general no way to recover the information of the element a that I may choose to represent it.

The set of equivalence classes in A modulo an equivalence relation $\sim$ is the 'quotient' $A/\sim$. This is a very basic and very important construction, used extensively in abstract algebra, where you learn about 'quotient rings', 'quotient groups', and the like. We will explore it more fully in §2.4. For now, let's focus on firming up our understanding of equivalence classes.

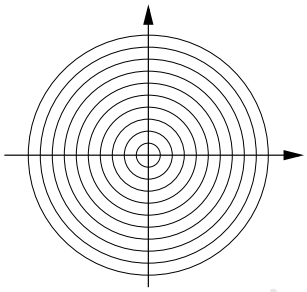
Example 2.3.8. For the equivalence relation considered in Example 2.3.2, there are exactly two distinct equivalence classes: $[0]_\sim$ and $[1]_\sim$. Indeed, if a number a is even, then $0 \sim a$, since $a = a - 0$ is even; if a number a is odd, then $1 \sim a$, since $a = a - 1$ is even in this case. Therefore, every number a is in either the equivalence class $[0]_\sim$ or the equivalence class $[1]_\sim$.

do not distribute

In fact, for this equivalence relation on $\mathbb{Z}$, $[0]_\sim$ is precisely the set of even numbers, and $[1]_\sim$ is precisely the set of odd numbers. Again, if I am thinking of a number a, and I tell you that its equivalence class is $[0]_\sim$ (that is, I tell you that $[a]_\sim = [0]_\sim$), you have no way to figure out the specific number a. All you will know is that the number must be *even*.

Note that the subsets $[0]_\sim$ and $[1]_\sim$ of $\mathbb{Z}$ are nonempty; disjoint; and their union is the whole of $\mathbb{Z}$. This makes the quotient $\mathbb{Z}/\sim = \{[0]_\sim, [1]_\sim\}$ a *partition* of $\mathbb{Z}$. ⌟

Example 2.3.9. Going back to Example 2.3.5, the equivalence classes are concentric circles: $[(x,y)]_\sim$ is the circle centered at the origin and with radius $\sqrt{x^2+y^2}$. This includes the general case of $[(0,0)]_\sim$, which is just the origin (a 'circle of radius 0').



Two points (x_1, y_1), (x_2, y_2) determine the same equivalence class when $[(x_1, y_1)]_\sim = [(x_2, y_2)]_\sim$, that is, when they belong to the same circle, that is, when $x_1^2 + y_1^2 = r^2$ and $x_2^2 + y_2^2 = r^2$ for the same radius r. In other words, $[(x_1, y_1)]_\sim = [(x_2, y_2)]_\sim$ if and only if $(x_1, y_1) \sim (x_2, y_2)$, as prescribed by Lemma 2.3.7.

Also note that each equivalence class is nonempty; any two are disjoint; and the union of the equivalence classes is the set $A = \mathbb{R}^2$ on which the equivalence relation is defined. As in Example 2.3.8, the set of equivalence classes $A/\sim$ is a partition of A. We will verify soon that this is a general feature of equivalence classes. ⌟

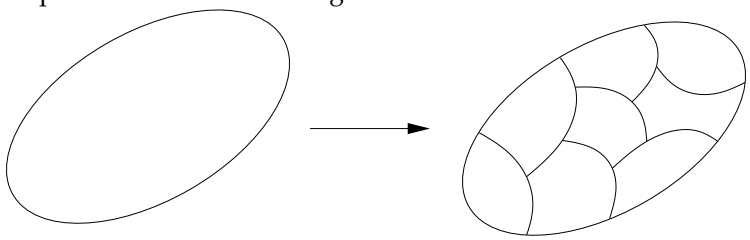
2.3.3 *Partitions and equivalence relations*

A 'partition' of a set A is a splitting of A into a collection of disjoint subsets. More precisely:

Definition 2.3.10. A *partition* of a set A is a family of subsets of A, $\{A_i\}_{i \in I}$, such that

- each A_i is nonempty;
- every two distinct subsets in the family are disjoint: $A_i \cap A_j = \emptyset$ for all i and j in I such that $i \neq j$;
- the union of all the subsets A_i is A: $\cup_{i \in I} A_i = A$. ⌟

My mental image of a partition looks something like

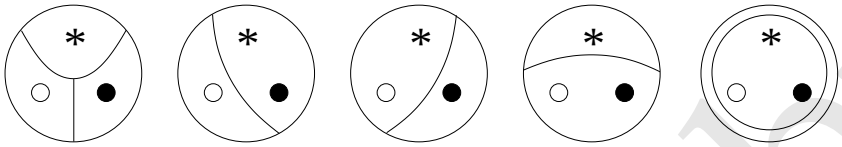


do not distribute

Example 2.3.11. Let $A = \{*, \bullet, \circ\}$ be a set with three elements. Here are *all* the possible partitions of A:

$$\{\{*\},\{\bullet\},\{\circ\}\}$$
$$\{\{*,\bullet\},\{\circ\}\}$$
$$\{\{*,\circ\},\{\bullet\}\}$$
$$\{\{\circ,\bullet\},\{*\}\}$$
$$\{\{*,\bullet,\circ\}\}$$

Pictorially:



⌟

Now I claim that *giving an equivalence relation on a set A is 'the same as' giving a partition of A.* That is, the situations reviewed in Examples 2.3.8 and 2.3.9 are typical: *the equivalence classes always form a partition of A.* Further, *every partition of A arises in this way.* We will prove this in the two theorems that follow (Theorem 2.3.12 and 2.3.13).

—*From equivalence relations to partitions*

Theorem 2.3.12. *Let A be a set, and let $\sim$ be an equivalence relation on A. Then*

- *Every equivalence class is nonempty.*
- *Every two different equivalence classes are disjoint.*
- *The union of the equivalence classes is A.*

In other words, the set of equivalence classes of elements of A with respect to an equivalence relation is a partition of A.

The proof of this fact is straightforward, but it's worth going through it carefully. The proof relies on Lemma 2.3.7, which uses in an essential way all three defining properties of equivalence relations. This is the case for Theorem 2.3.12 as well. For example, the first point holds because equivalence relations are assumed to be *reflexive:* for all $a \in A$, $a \sim a$, and then this means that $a \in [a]_\sim$ by definition of equivalence class. In particular, $[a]_\sim$ is not empty. If we tried to set up a theory of equivalence classes for relations that are not reflexive, we would not be able to draw this conclusion!

Proof. The class $[a]_\sim$ contains the element a, since $\sim$ is reflexive. This shows that equivalence classes are nonempty, establishing the first point.

The second point follows from Lemma 2.3.7. Indeed, let $[a]_\sim$ and $[b]_\sim$ be two equivalence classes. Arguing contrapositively, assume that $[a]_\sim$ and $[b]_\sim$ are not disjoint; so $\exists c$ such that $c \in [a]_\sim \cap [b]_\sim$. This implies $a \sim c$ and $b \sim c$. By Lemma 2.3.7, this would in turn imply $[a]_\sim = [c]_\sim$ and $[b]_\sim = [c]_\sim$, and hence $[a]_\sim = [b]_\sim$. Therefore, if $[a]_\sim$ and $[b]_\sim$ are not disjoint, then in fact $[a]_\sim = [b]_\sim$, so they are not distinct. This proves the second point.

For the third point, let U be the union of the equivalence classes. Every element $a \in A$ is in an equivalence class, i.e., $[a]_\sim$; again this is because $\sim$ is reflexive. It follows that every $a \in A$ is contained in U, hence $A \subseteq U$. On the other hand each equivalence class is a subset of A, and therefore $U \subseteq A$. We can conclude that $U = A$ as needed, by Proposition 2.1.3. □

do not distribute

Summarizing: Every equivalence relation $\sim$ on a set A determines a specific partition of A: the partition consisting of the equivalence classes modulo $\sim$. The 'quotient' of A modulo $\sim$ will be this partition (as we will formalize in Definition 2.4.1).

—From partitions to equivalence relations

We have just verified that every equivalence relation on a set A determines a partition of A. Next, I claim that *every* partition arises in this way. That is, I claim that if $\mathcal{P}$ is a partition of A, then I can find an equivalence relation $\sim$ on A such that the elements of $\mathcal{P}$ are precisely the equivalence classes modulo $\sim$.

This is actually pretty easy. If $\mathcal{P}$ is a partition of A, define $\sim$ by declaring that $\forall a, b \in A$, a and b belong to the same element of $\mathcal{P}$. That is:

$$a \sim b \quad \iff \quad \exists P \in \mathcal{P}\,,\ (a \in P) \wedge (b \in P)$$

Theorem 2.3.13. *Let $\mathcal{P}$ be a partition of a set A. Then the relation $\sim$ we have just defined is an equivalence relation, and the corresponding equivalence classes are precisely the elements of the partition $\mathcal{P}$.*

Proof. We have to verify that $\sim$ is reflexive, symmetric, and transitive.

To see that $\sim$ is reflexive, let $a \in A$. Since $\mathcal{P}$ is a partition, the union of the elements of $\mathcal{P}$ is A. This implies that some element $P \in \mathcal{P}$ contains a. It follows that $a \sim a$, since 'both a and a' belong to $P \in \mathcal{P}$.

To see that $\sim$ is symmetric, assume that $a \sim b$. Then a and b are in the same element of $\mathcal{P}$, and it trivially follows that b and a are in the same element of $\mathcal{P}$. This means that $b \sim a$, as needed.

To see that $\sim$ is transitive, assume that $a \sim b$ and $b \sim c$. Then a and b are in the same element $P \in \mathcal{P}$, and b and c are in the same element $Q \in \mathcal{P}$. Since b is in both P and Q, these two elements P and Q are not disjoint. Since $\mathcal{P}$ is a partition, it follows that $P = Q$. But then a and c both belong to $P = Q \in \mathcal{P}$, and this proves that $a \sim c$.

Thus, $\sim$ is indeed an equivalence relation.

To conclude the proof of the theorem, just observe that $[a]_\sim$ consists of all $b \in A$ contained in the same element P of $\mathcal{P}$ which contains a, by definition of $\sim$. This shows that $[a]_\sim = P$, proving that indeed the equivalence classes modulo $\sim$ are the elements of the given partition. □

Taken together, Theorem 2.3.12 and Theorem 2.3.13 prove that there is a complete dictionary between equivalence relations and partitions: every equivalence relation on a set A determines a partition of A and is determined by this partition.

Example 2.3.14. There are five different equivalence relations on a set $A = \{*, \bullet, \circ\}$ with three elements. Indeed, we have seen in Example 2.3.11 that there are five different partitions of A. The first partition listed in Example 2.3.11 corresponds to the equivalence relation $=$. ⌟

2.3.4 *Homework*

2.3.1: Describe the subset of $A \times A$ corresponding to the relation $=$. Draw its picture in $\mathbb{R}^2$, for the case $A = \mathbb{R}$.

2.3.2: What's wrong with the following 'proof' that symmetry and transitivity imply reflexivity? Assume that the relation $\sim$ is symmetric and transitive, and argue that $a \sim a$ for all a, as follows: (1) since $\sim$ is symmetric, $a \sim b$ implies $b \sim a$; (2) since $a \sim b$ and $b \sim a$, by transitivity it follows that $a \sim a$. Why doesn't this argument work?

▷ **2.3.3:** Say that two squares of a chessboard are 'knight-equivalent' if one can be reached from the other by a sequence of knight moves. How many equivalence classes are there?

do not distribute

Now say that two squares of a chessboard are 'bishop-equivalent' if one can be reached from the other by a sequence of bishop moves. How many equivalence classes are there?

2.3.4: Say that two notes on a piano are equivalent if they are any number of octaves apart. How many equivalence classes are there?

▷ **2.3.5:** Let n be a fixed positive integer. Define a relation $\sim$ on $\mathbb{Z}$ by:

$$a \sim b \quad \iff \quad a \text{ and } b \text{ have the same remainder after division by } n.$$

Prove that for every given n, $\sim$ is an equivalence relation. (The case $n = 2$ is worked out in Example 2.3.2.)

▷ **2.3.6:** Describe explicitly the equivalence classes for the relation $\sim$ defined in Exercise 2.3.5. How many *distinct* equivalence classes are there? (This gives a description of the set $\mathbb{Z}/\sim$ in this case.)

2.3.7: Consider the set $A = \{1, 2, 3, 4, 5, 6, 7, 8, 9, 10\}$, and the equivalence relation $\sim$ defined on A by declaring that $a \sim b$ if and only if $b - a$ is a multiple of 3. Find the partition corresponding to this relation.

2.3.8: Consider the following relation in the plane $\mathbb{R}^2$:

$$(x_1, y_1) \sim (x_2, y_2) \iff x_1 = x_2 \quad .$$

Prove that this is an *equivalence* relation, and describe the equivalence classes.

▷ **2.3.9:** List all partitions of a set with 4 elements. How many partitions are there?

2.3.10: How many different equivalence relations are there on a set with 4 elements?

2.4 *Quotients modulo an equivalence relation.*

2.4.1 *Quotient of a set by an equivalence relation*

Now that we know about the link between equivalence relations and partitions, we can give a better look at the notion of 'quotient' set. The definition of quotient was given in passing in §2.3.2. Here it is again, more officially.

Definition 2.4.1. The set of equivalence classes determined by an equivalence relation $\sim$ on a set A is called the *quotient of A modulo $\sim$*, denoted $A/\sim$. ⌟

Thus, the quotient of a set with respect to an equivalence relation is just the partition associated to that relation as we worked out in §2.3.3.

Example 2.4.2. If $\sim$ is the relation on $\mathbb{Z}$ defined in Example 2.3.2, the quotient

$$\mathbb{Z}/\sim \; = \{[0]_\sim, [1]_\sim\} = \{\{\text{even integers}\}, \{\text{odd integers}\}\}$$

consists of two elements[4]. Note that 0, 1 are respectively the 'remainder after division by 2' of even, respectively odd integers. ⌟

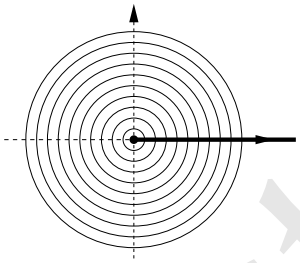
[4] Potential objection: if $\mathbb{Z}/\sim \; = \{\{\text{even integers}\}, \{\text{odd integers}\}\}$, doesn't that mean that $\mathbb{Z}/\sim$ consists of all integers? No! There is a vast difference between the set {even integers} and the set {{even integers}}. The first is the set of even integers; elements of this set are $\ldots, -2, 0, 2, 4, \ldots$; this set contains infinitely many elements. The second set contains *only one* element, and that element happens to be the set of even integers. The first set is a box containing lots and lots of things. The second set is a box which contains one box. This latter box contains lots of things, but that's a different story.

do not distribute

It is often the case that a quotient admits a compelling description, capturing all the information carried by the equivalence classes. *Warning:* Making this notion of 'description' precise requires more sophisticated language than we have developed thus far. We will be able to think again about these examples (and your homework) more precisely once we cover the notions presented in Chapter 3, and indeed we will come back to this issue in §3.5, once we learn about 'canonical decompositions'. Still, thinking about several examples now with the more limited language at our disposal is quite valuable—it will prepare us well for later encounters with this type of examples.

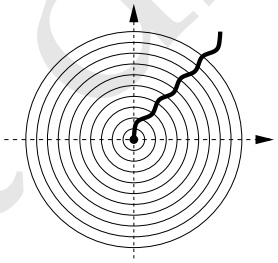
Example 2.4.3. In the case of Example 2.3.5, $\mathbb{R}^{\geq 0}$ could be taken as a model for the quotient set $\mathbb{R}^2/\!\sim$. Indeed, we have observed (Example 2.3.9) that the equivalence classes in this case are concentric circles of any radius ≥ 0; there is exactly one such circle (hence, exactly one equivalence class) for each real number $r \geq 0$.

If you want to 'picture' this, here are several equivalence classes (as in Example 2.3.9) juxtaposed to the line of nonnegative real numbers in the x-axis:



You see? There is exactly one point on the fat line for each equivalence class (that is, for each circle), and conversely one equivalence class for each point on the fat line.

Note that I could extend the half-line from the origin in any direction I want, with the same result. There is nothing special about $\mathbb{R}^{\geq 0}$ being 'horizontal, pointing to the right' as in this picture. It could even be a squiggly meandering curve like this:



so long as it meets *every* equivalence class in *exactly one* point. The set $\mathbb{R}^2/\!\sim$ may admit several reasonable descriptions, but none of them is 'special'. The set exists on its own, regardless on how we try to depict it. ⌟

Example 2.4.4. Consider again the equivalence relation $\sim$ defined on $\mathbb{R}$ by

$$a \sim b \quad \Longleftrightarrow \quad b - a \in \mathbb{Z}$$

(see Example 2.3.3). The equivalence classes corresponding to this relation are 'copies of $\mathbb{Z}$' shifted by a real number:

$$[r]_\sim = \{r + n \,|\, n \in \mathbb{Z}\} = \{\dots, r-2, r-1, r, r+1, r+2, \dots\} \quad .$$

Similarly, think of

$$\{\{\text{even integers}\}, \{\text{odd integers}\}\}$$

as a box containing two boxes. These two boxes are full of stuff, but when you open the largest box you only see two things, that is, two boxes. In this sense it is a set 'with two elements'.

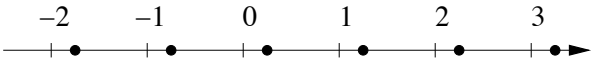
do not distribute

Indeed,

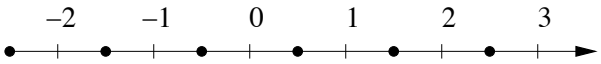
$$s \in [r]_\sim \iff r \sim s \iff s - r \in \mathbb{Z} \iff \exists n \in \mathbb{Z} : s = r + n \quad .$$

For example, $[0]_\sim$ is $\mathbb{Z}$ itself:

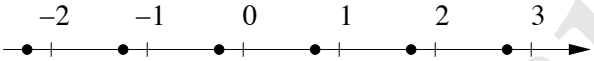
while here is $[0.25]_\sim$:



here is $[0.5]_\sim$:



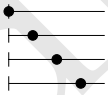
and here is $[0.75]_\sim$:



What about $[1]_\sim$? that will look exactly like $[0]_\sim$ again:

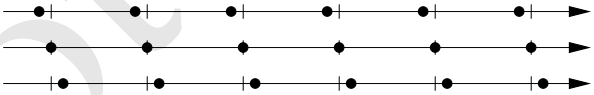
The reason why $[1]_\sim$ looks like $[0]_\sim$ is that it *is* $[0]_\sim$: we figured out in Lemma 2.3.7 that $[a]_\sim = [b]_\sim$ precisely when $a \sim b$, and here $1 \sim 0$, so of course $[1]_\sim = [0]_\sim$. By the same token, $[1.25]_\sim = [0.25]_\sim = [17.25]_\sim = \cdots$, and more generally $[r]_\sim = [r']_\sim$, where r' is the 'decimal part' of r: the unique number r' such that $0 \leq r' < 1$ and that $r - r' \in \mathbb{Z}$.

We could conclude that therefore a good description of the quotient $\mathbb{R}/\sim$ is the interval $[0, 1)$: there is exactly one point in this interval for every equivalence class modulo $\sim$. Here are again $[0]_\sim$, $[0.25]_\sim$, $[0.5]_\sim$, $[0.75]_\sim$, viewed as elements of this model for $\mathbb{R}/\sim$:



(I just chopped off from the previous pictures the part < 0 and ≥ 1, leaving the trace on the interval $[0, 1)$.)

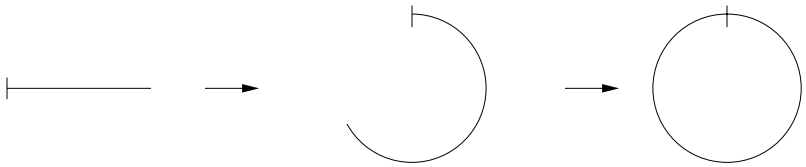
We can do a little better, however. From the point of view of $\mathbb{R}/\sim$, the class $[0]_\sim$ is not really special, in the sense that I can 'approach it' from left to right, and I see nothing particularly drastic happening:



But if I use my model $[0, 1)$, something drastic does seem to happen:

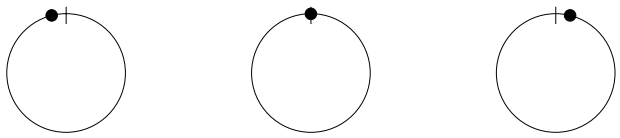
So maybe my model is not so good after all. Can you think of a way to improve it?

You have the same problem in representing time: time does not jump as you go from minute 59 of an hour to minute 00 of the next hour. That's why clocks are round, to make 60 be the same as 00. Here we can do the same: bend your interval $[0, 1)$, joining the two ends to get a circle:



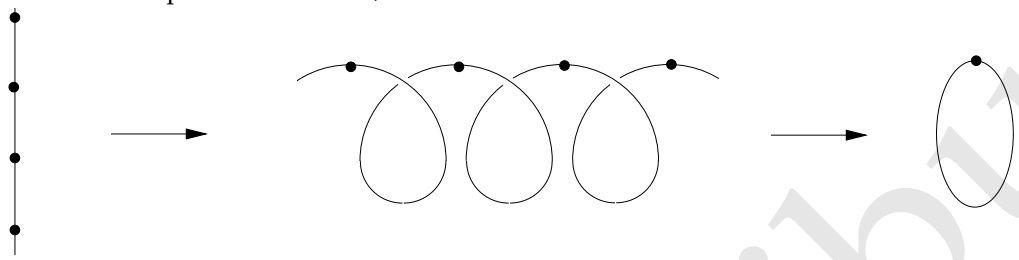
Then again nothing too drastic happens as I approach $[0]_\sim$

do not distribute



I would consider the circle a better representation for the quotient $\mathbb{R}/\sim$.

In a sense, what we are doing is 'wrapping the real line around a circle' in such a way that equivalent numbers go to the same point of the circle;



(Please forgive my attempts at drawing in perspective: that thing on the right is still a circle.) ⌟

I can sympathize with you if you feel a bit of discomfort here: at some stage (in Chapter 5) we will understand that all the sets we have encountered in this subsection are 'isomorphic' as sets—from the point of view of set theory alone, they cannot really be distinguished one from the other. The efficacy of a particular description of a quotient $A/\sim$ depends on the context in which it arises. My preference for the circle rather than the interval $[0,1)$ in the last example is due to the fact that it is more natural in a topological context, and indeed you will run into it again when, one happy day, you learn about universal covers while studying topology. You may also run into it in an algebra context: for reasons that you will learn when the time is ripe, the quotient $\mathbb{R}/\sim$ is usually denoted $\mathbb{R}/\mathbb{Z}$, and you will write things such as '$\mathbb{R}/\mathbb{Z} \cong S^1$' ($S^1$ is the standard notation for the circle). Of course, it is very advisable that by the time you approach topology or algebra you feel quite comfortable with equivalence relations and quotients.

2.4.2 *One more example:* $\mathbb{Q}$ *as a quotient*

Here is another example of quotient construction, which will look different from the 'geometric' examples reviewed above. Denote by $\mathbb{Z}^{\neq 0}$ the set of *nonzero* integers. I am going to define a relation on the set $\mathbb{Z} \times \mathbb{Z}^{\neq 0}$, that is, on the set of pairs (p,q) where p and q are integers and q is further required to be $\neq 0$.

Here is the relation:

$$(p_1,q_1) \sim (p_2,q_2) \quad \Longleftrightarrow \quad p_1q_2 = p_2q_1 \quad . \tag{2.8}$$

Lemma 2.4.5. *This is an equivalence relation.*

Proof. Reflexivity and symmetry are straightforward:

- $\sim$ is reflexive: taking $p_1 = p_2 = p$, $q_1 = q_2 = q$, (2.8) says $pq = pq$, which is correct; so indeed $(p,q) \sim (p,q)$.
- $\sim$ is symmetric: if $(p_1,q_1) \sim (p_2,q_2)$, then $p_1q_2 = p_2q_1$. But then $p_2q_1 = p_1q_2$, so we can conclude $(p_2,q_2) \sim (p_1,q_1)$.

Proving transitivity requires more work. Assume that $(p_1,q_1) \sim (p_2,q_2)$ and $(p_2,q_2) \sim (p_3,q_3)$. Then $p_1q_2 = p_2q_1$ and $p_2q_3 = p_3q_2$. Multiplying both terms in the first identity by q_3 gives

$$p_1q_2q_3 = p_2q_1q_3 \quad ;$$

do not distribute

multiplying both terms in the second identity by q_1 gives

$$p_2 q_3 q_1 = p_3 q_2 q_1 \quad .$$

Together, these identities show that

$$p_1 q_2 q_3 = p_3 q_2 q_1 \quad .$$

Since $q_2 \neq 0$ by assumption, this implies that

$$p_1 q_3 = p_3 q_1 \quad ,$$

and hence $(p_1, q_1) \sim (p_3, q_3)$, as needed. □

Note that in the proof of transitivity we have actually used the fact that the second element of the pairs (p, q) we are considering is nonzero. We really need $\mathbb{Z} \times \mathbb{Z}^{\neq 0}$ to make this work.

How would we describe the quotient $(\mathbb{Z} \times \mathbb{Z}^{\neq 0})/\sim$?

The equivalence class of the pair $(1, 2)$ is, by definition, the set of all pairs (p, q) with $q \neq 0$ and such that

$$1 \cdot q = 2 \cdot p \quad ,$$

or simply

$$\frac{p}{q} = \frac{1}{2} \quad .$$

So, $[(1, 2)]_\sim$ is just the set of all pairs (p, q) of integers such that the fraction p/q determines the rational number $1/2$. More generally, if $r \in \mathbb{Q}$ is a rational number, and we write $r = a/b$ in any way as a quotient of two integers, then $[(a, b)]_\sim$ is precisely the set of all pairs (p, q) and such that the fraction p/q equals r as a rational number: $[(a, b)]_\sim$ is the set of all representations of the rational number $r = a/b$ as a quotient of two integers.

A moment's thought reveals that therefore there is exactly one equivalence class for every rational number. Indeed, a rational number $\frac{p}{q}$ is represented as the equivalence class $[(p, q)]_\sim$, and there are no 'duplicates' in the sense that two classes representing the same number necessarily agree:

$$\frac{p_1}{q_1} = \frac{p_2}{q_2} \iff p_1 q_2 = p_2 q_1 \iff (p_1, q_1) \sim (p_2, q_2) \iff [(p_1, q_1)]_\sim = [(p_2, q_2)]_\sim$$

(where in the last equality I again used Lemma 2.3.7).

Therefore: The quotient $(\mathbb{Z} \times \mathbb{Z}^{\neq 0})/\sim$ may be described as a copy of the set $\mathbb{Q}$ of rational numbers.

In fact, this works well as a *definition* for the set $\mathbb{Q}$. In courses in abstract algebra, this construction is given as a particular case of the construction of the 'field of fractions' of any given 'integral domain'. This more general construction is itself a particular case of the process of 'localization', which is very important in commutative algebra, number theory, and algebraic geometry. ⌟

2.4.3 *Homework*

All the problem listed here deal with plane geometry. This is very reductive—the notions we covered in this chapter are extremely general. In fact, you are more likely to end up dealing more extensively with 'quotients' in algebra, and those examples may end up looking more along the lines of Exercises 2.3.5 and 2.3.6 or Example 2.4.2. Still, it seemed nice to illustrate that even such a simple-minded environment as elementary geometry allows for a rich collection of different examples.

do not distribute

2.4.1: Consider the equivalence relation $\sim$ on the set $\mathbb{R}^2$ defined by declaring

$$(x_1, y_1) \sim (x_2, y_2) \quad \iff \quad y_1 - x_1^2 = y_2 - x_2^2 \quad .$$

What kind of geometric object is the equivalence class $[(a, b)]_\sim$ of a point (a, b)? Describe the quotient $\mathbb{R}^2/\sim$.

2.4.2: Consider the relation $\sim$ on $\mathbb{R}^2 \smallsetminus \{(0, 0)\}$ defined by declaring that $(x_1, y_1) \sim (x_2, y_2)$ if and only if $\exists r \in \mathbb{R}^{>0}, (x_2, y_2) = (rx_1, ry_1)$.

Verify that $\sim$ is an equivalence relation. Describe the equivalence classes 'geometrically'. Describe the quotient $(\mathbb{R}^2 \smallsetminus \{(0, 0)\})/\sim$.

▷ **2.4.3:** Let A be the set of *squares* in the plane. Define an equivalence relation on A by letting $S_1 \sim S_2$ if there is a rigid motion in the plane taking the square S_1 to the square S_2. Describe the quotient $A/\sim$.

2.4.4: The same as in Exercise 2.4.3, but with *rectangles.* Let A be the set of rectangles in the plane; declare that $R_1 \sim R_2$ if there is a rigid motion in the plane taking the rectangle R_1 to the rectangle R_2. Describe the quotient $A/\sim$.

▷ **2.4.5:** Let A be the set of 'oriented segments' $\overrightarrow{PQ}$ in the plane. Define an equivalence relation $\sim$ on A by prescribing that $\overrightarrow{P_1Q_1} \sim \overrightarrow{P_2Q_2}$ if the segments are parallel, point in the same direction, and have the same length. (It should be clear that this is an equivalence relation.) Describe the quotient $A/\sim$.

2.4.6: The same as in Exercise 2.4.5, but now let A be the set of *unoriented* segments PQ in the plane. We set $P_1Q_1 \sim P_2Q_2$ if the two segments are parallel and they have the same length; we cannot talk about 'pointing in the same direction', since the segments are not oriented. Describe the quotient $A/\sim$.

▷ **2.4.7:** (1) Define a relation $\sim'$ on $\mathbb{R}^2$ by prescribing

$$(x_1, y_1) \sim' (x_2, y_2) \quad \iff \quad x_1 = x_2 \text{ and } y_2 - y_1 \in \mathbb{Z}.$$

(This is easily checked to be an equivalence relation.) Find a good geometric description of the quotient $\mathbb{R}^2/\sim'$.

(2) Do the same for the equivalence relation $\sim''$ on $\mathbb{R}^2$ defined by

$$(x_1, y_1) \sim'' (x_2, y_2) \quad \iff \quad x_2 - x_1 \in \mathbb{Z} \text{ and } y_2 - y_1 \in \mathbb{Z}.$$

do not distribute

do not distribute

3
Functions and canonical decomposition

3.1 Functions.

3.1.1 Functions: basic definitions

Next we move to another fundamental type of relations: *functions.* You have used functions extensively in calculus, and you have drawn many *graphs* of real functions of one real variable: that is, subsets of $\mathbb{R}^2 = \mathbb{R} \times \mathbb{R}$ of the type

$$\{(x, y) \in \mathbb{R}^2 \,|\, y = f(x)\} \quad .$$

I have already pointed this out to you in Example 2.1.20. In calculus, you have probably gotten used to thinking of a function as some kind of 'recipe' that gives you a number when you plug in a number: so the output of $f(x) = x^2$ is the square of the input. From a more general point of view, however, this is not really adequate: there are many functions for which such a simple recipe just does not exist. Abstractly, the datum of a function is captured completely by its graph, which is the most detailed description of 'which y is the output $f(x)$' for all given x. Thus, we think of functions in terms of the relations they define.

Not all relations qualify as functions, though, just like not all relations are equivalence relations. The official definition is as follows. Note the ! following $\exists$ in the definition: this stands for 'unique': so $\exists! b \in B \ldots$ means 'there exists a unique element b of $B \ldots$', that is, there is *only one* $b \in B \ldots$.

Definition 3.1.1. Let A, B be sets. A *function* f 'from A to B' is the datum of a relation

$$\Gamma_f \subseteq A \times B$$

such that

$$(\forall a \in A)\,(\exists! b \in B): \quad (a, b) \in \Gamma_f \quad . \tag{*}$$

We call Γ_f the *graph* of the function f. ⌟

If f is a function, then for each $a \in A$ there is *only one* b such that $(a, b) \in \Gamma_f$. Therefore, we can invent a notation for this element $b \in B$ depending on a. This notation is the familiar

$$b = f(a) \quad ,$$

which you should view as a shorthand for $(a, b) \in \Gamma_f$. Think about this for a moment—if $b_1 \neq b_2$ and both (a, b_1) and (a, b_2) were in Γ_f, then we could not really write '$f(a)$' *as an element of B:* we would not know whether $f(a)$ refers to b_1 or b_2. (Maybe $f(a)$ would make sense as a subset of B, but that's not what the

do not distribute

definition says: the definition requires $f(a)$ to be a specific *element* of B.) The requirement of uniqueness is what allows us to use such a notation.

In 'functional' notation, the requirement on functions expressed in Definition 3.1.1 is

$$(\forall a \in A)\,(\exists! b \in B): \quad f(a) = b \quad .$$

Again: for f to be a function, we are requiring that *for every $a \in A$, there exists a unique b* such that $b = f(a)$. This is implicit in the very use of the notation $f(a)$.

To indicate that f is a function from A to B, we draw the diagram

$$f : A \to B$$

or

$$A \xrightarrow{f} B \quad .$$

If the function f is understood, the fact that $f(a) = b$ may be denoted by the diagram

$$a \mapsto b$$

(note the different type of arrow).

Terminology: A is the *source*, or *domain*, of $f : A \to B$; B is the *target*, or *codomain*. If $b = f(a)$, we say that b is the *image* of a, and that a is a *preimage* of b, or that a belongs to the *inverse image* of b.

Note carefully: $\boxed{\textit{the}}$ *image*, because for every a there is a *unique* b such that $f(a) = b$; but $\boxed{a}$ *preimage*, because b could be the image of several elements a.

In fact, every function $f : A \to B$ determines several interesting subsets of A and B:

Definition 3.1.2. Let $f : A \to B$ be a function.

- For every $b \in B$, the *inverse image of b*, denoted $f^{-1}(b)$, is the set[1] of preimages of b:
$$f^{-1}(b) = \{a \in A \mid f(a) = b\} \quad .$$
- More generally, if $T \subseteq B$ is a subset, then $f^{-1}(T)$ denotes the set of preimages of elements of T:
$$f^{-1}(T) = \{a \in A \mid f(a) \in T\} \quad .$$
- For every subset $S \subseteq A$, the *image of S*, denoted $f(S)$, is the set of images of elements of S:
$$f(S) = \{b \in B \mid (\exists s \in S) : b = f(s)\} \quad .$$
- The *image of f*, denoted $\operatorname{im} f$, is the subset of B consisting of all images of elements of A. That is, $\operatorname{im} f = f(A)$. ⌟

Remark 3.1.3. This terminology is somewhat unfortunate. We will soon introduce a *different* use of 'f^{-1}'; and the image 'of a function' is not the image 'of one element'. The only way not to get lost in the maze of terms is to always be very aware of the context in which they are used, and to pay attention to details. If I say 'image of an element', then I am talking about one element in the target; if I say 'image of a function', then I am talking about a whole subset of the target. Of course these are different concepts. ⌟

[1] Note: a whole subset of A, not just one element! Contrast this situation with the preceding discussion leading to the notation $f(a)$.

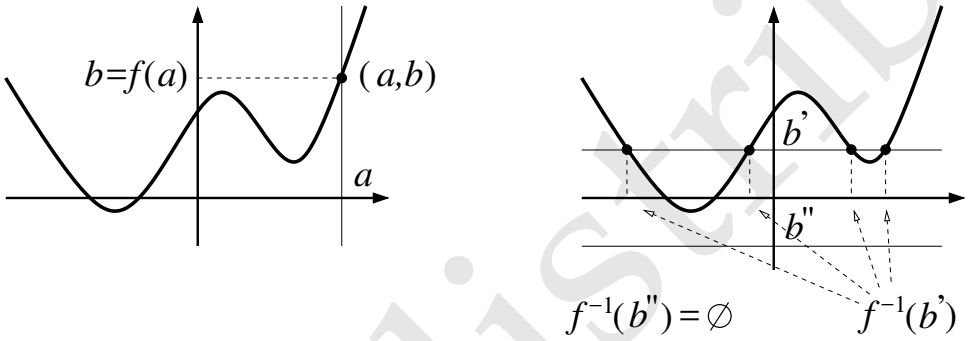
do not distribute

The term 'range' is also used, but with some inconsistencies: some use it to denote the image of a function, others use it for the target. I don't like inconsistencies, so I will stick to the terminology introduced above.

According to Definition 3.1.1, the graph of a function $f : \mathbb{R} \to \mathbb{R}$ consists of all points (x, y) in the plane $\mathbb{R}^2 = \mathbb{R} \times \mathbb{R}$ such that $y = f(x)$. This is the familiar notion of 'graph' from calculus. Several of the terms introduced above should also be familiar from calculus. But keep in mind that we are now in a more general setting, and the source A and target B of a function can be any sets.

Remark 3.1.4. In the case of functions $\mathbb{R} \to \mathbb{R}$, the condition (*) in Definition 3.1.1 is easy to visualize. The point (a, b) of the graph corresponding to any $a \in \mathbb{R}$ is a point on the line $x = a$, a 'vertical' line. So condition (*) says that *every vertical line meets the graph at exactly one point:* this is the same as saying that for every a there is exactly one (that is, 'there exists a unique') b such that (a, b) is a point of the graph.

On the other hand, *horizontal* lines may meet the graph of a function $f : \mathbb{R} \to \mathbb{R}$ in many points; in fact, the line $y = b$ meets the graph of f precisely at the points (a, b) with $a \in f^{-1}(b)$. (Right?) There may be many such points: $f^{-1}(b)$ is a whole subset of the source.



The horizontal line $y = b$ may also miss the graph entirely, in which case $f^{-1}(b) = \emptyset$. ⌟

3.1.2 *Examples (and more definitions)*

Example 3.1.5. We can define a function $f : \mathbb{R} \to \mathbb{R}$ by setting $f(x) = x^2$. For every real number r, the function prescribes that $f(r) = r^2$, and this number r^2 is uniquely determined by r. On the other hand, we have e.g. $f(-2) = f(2) = 4$: so -2 and 2 are preimages of 4; in fact they are *all* its preimages, so the inverse image of 4 consists of $\{2, -2\}$.

The image $\operatorname{im} f$ of this function is the set of nonnegative real numbers: $\operatorname{im} f = \mathbb{R}^{\geq 0}$. We can also write $f(\mathbb{R}) = \mathbb{R}^{\geq 0}$. Note that $f(\mathbb{R}^{\geq 0})$ also equals $\mathbb{R}^{\geq 0}$: every nonnegative real number is the square of some nonnegative real number. In fact, every nonnegative real number is the square of some *nonpositive* real number as well, a fact that we could express by writing (correctly) $f(\mathbb{R}^{\leq 0}) = \mathbb{R}^{\geq 0}$. If we want to express the fact that no negative number is the square of a real number, we can write $f^{-1}(\mathbb{R}^{<0}) = \emptyset$.

The same formula $f(x) = x^2$ would also define a function $\mathbb{R} \to \mathbb{R}^{\geq 0}$, since $x^2 \geq 0$ for all $x \in \mathbb{R}$. If we are being precise, then the two functions $f : \mathbb{R} \to \mathbb{R}$ and $f : \mathbb{R} \to \mathbb{R}^{\geq 0}$ defined by the same formula $f(x) = x^2$ should be considered as *different* functions, because they have different targets. (In fact, the first function is determined by a subset of $\mathbb{R} \times \mathbb{R}$; the second by a subset of $\mathbb{R} \times \mathbb{R}^{\geq 0}$.) It is not uncommon to gloss over such distinctions, but I will try to be careful about such things, and you should too. ⌟

Example 3.1.6. We can*not* define a function $f : \mathbb{R} \to \mathbb{R}$ by setting $f(x) = \sqrt{x}$: indeed, $\sqrt{a}$ is not defined as a real number if $a < 0$, so it is not true that *for every* a there is a b such that $b = f(a)$.

We can however define a function $f : \mathbb{R}^{\geq 0} \to \mathbb{R}$ by the same formula, where $\mathbb{R}^{\geq 0}$ denotes the set of

do not distribute

nonnegative real numbers, because it is true that *every* nonnegative real number has a square root.

The prescription $f(x) = \pm\sqrt{x}$ does not define a function, because we are giving *two* values $f(a)$ for $a > 0$: so it is not true that for every a there exists a *unique* b such that $b = f(a)$. (By convention, $\sqrt{a}$ denotes the *nonnegative* square root of a.) At best, this is shorthand for the *two* functions $\sqrt{x}$ and $-\sqrt{x}$. ⌟

Example 3.1.7. Giving an $m \times n$ matrix M with real entries, as in linear algebra, is the same as giving a function

$$(\{1, \ldots, m\} \times \{1, \ldots, n\}) \to \mathbb{R} \quad .$$

(And indeed, *this* is how matrices were defined in the linear algebra class I took in college.) Indeed, elements of $\{1, \ldots, m\} \times \{1, \ldots, n\}$ are pairs (i, j), where i and j are integers and $1 \leq i \leq m$, $1 \leq j \leq n$. The value of the function at (i, j) can be interpreted as the (i, j) entry of a matrix, and conversely. For instance, the matrix

$$\begin{pmatrix} 0 & 17 & 3 \\ 2 & 4 & -\pi \end{pmatrix}$$

would correspond to the function $f : \{1, 2\} \times \{1, 2, 3\} \to \mathbb{R}$ defined by setting

$$\begin{cases} f(1,1) = 0 \\ f(1,2) = 17 \\ f(1,3) = 3 \\ f(2,1) = 2 \\ f(2,2) = 4 \\ f(2,3) = -\pi \end{cases} .$$

Nothing deep is going on here, but talking about matrices in this way avoids awkward terms such as 'arrays of numbers…'. ⌟

Example 3.1.8. Very early on in these notes (§1.1.3) we run into the notion of 'indexed set,' and back then I mentioned that they are best understood as functions.

You probably see now what I meant then. The notation introduced in §1.1.3, $\{s_n\}_{n \in \mathbb{N}}$, with elements s_n taken from a set S, is simply a different notation for the function $\mathbb{N} \to S$ sending $n \in \mathbb{N}$ to $s_n \in S$.

We can similarly consider indexed sets $\{s_i\}_{i \in I}$, where I is any set of indices. This would simply be a different notation for a function $I \to S$. ⌟

Example 3.1.9. Let $A = \{F, T\}$ and $B = \{*, \circ, \bullet\}$. How many functions $f : A \to B$ are there? We can write all of them out. Note that $A \times B$ has 6 elements, so there are $2^6 = 64$ subsets of $A \times B$ (Exercise 2.1.15); therefore, there are 64 relations among elements of A and B. I am asking: How many of these 64 relations are functions?

The point is that, in this example, *'for every $a \in A$' means for $a = F$ and for $a = T$;* for each of these two choices it must be true that *there is a unique $b \in B$* such that $f(a) = b$, and this expresses the choice of one element out of three ($*$, $\circ$, or $\bullet$). Therefore there are $3^2 = 9$ possibilities for $\Gamma_f \subseteq A \times B$:

$$\begin{array}{lll} \{(F,*),(T,*)\} & \{(F,*),(T,\circ)\} & \{(F,*),(T,\bullet)\} \\ \{(F,\circ),(T,*)\} & \{(F,\circ),(T,\circ)\} & \{(F,\circ),(T,\bullet)\} \\ \{(F,\bullet),(T,*)\} & \{(F,\bullet),(T,\circ)\} & \{(F,\bullet),(T,\bullet)\} \end{array}$$

These are all the (graphs of the) functions $A \to B$. For example, the graph in the middle corresponds to the function $f : A \to B$ defined by $f(F) = \circ$, $f(T) = \circ$. ⌟

do not distribute

Example 3.1.10. On the other hand, with the same $A = \{F, T\}$ and $B = \{*, \circ, \bullet\}$, there are $2^3 = 8$ functions $B \to A$:

$$\begin{array}{ll} \{(*,F),(\circ,F),(\bullet,F)\} & \{(*,F),(\circ,F),(\bullet,T)\} \\ \{(*,F),(\circ,T),(\bullet,F)\} & \{(*,F),(\circ,T),(\bullet,T)\} \\ \{(*,T),(\circ,F),(\bullet,F)\} & \{(*,T),(\circ,F),(\bullet,T)\} \\ \{(*,T),(\circ,T),(\bullet,F)\} & \{(*,T),(\circ,T),(\bullet,T)\} \end{array}$$

Indeed, we have to choose one of two options (F or T) for each of three possibilities ($*$, $\bullet$, $\circ$); there are $2 \cdot 2 \cdot 2$ ways to do this. The bottom graph in the first column shown above corresponds to the function defined by setting $f(*) = T$, $f(\circ) = T$, and $f(\bullet) = F$. ⌟

Of course the reasoning used in these examples generalizes to all finite sets.

Lemma 3.1.11. *Let A, B be finite sets. Then there are exactly $|B|^{|A|}$ functions $A \to B$.*

Proof. Every function $f : A \to B$ determines and is determined by which $b = f(a) \in B$ is selected for every $a \in A$. That amounts to selecting one of $|B|$ choices for every one of $|A|$ possibilities. Therefore there are $\underbrace{|B| \cdots |B|}_{|A| \text{ times}} = |B|^{|A|}$ possible functions $A \to B$. □

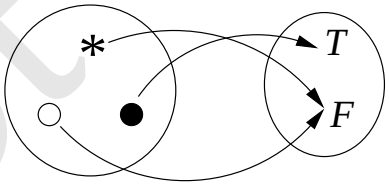
Lemma 3.1.11 inspires the following terminology.

Definition 3.1.12. If A and B are any sets, B^A denotes the set of functions from A to B. ⌟

This notation is used regardless of whether A and B are finite or infinite. By Lemma 3.1.11, *if* A and B are finite sets, then $|B^A| = |B|^{|A|}$.

Remark 3.1.13. You may remember from Definition 2.1.12 that 2^S is another symbol used to denote the power set $\mathcal{P}(S)$ consisting of all subsets of S. You will justify this choice of terminology in Exercise 3.1.4. ⌟

A picture may help visualize what a function is doing. For example,



could be a visual representation of the second function listed in Example 3.1.10 (top of second column).

Let's review a few basic examples of functions. These may not look much like your ordinary calculus functions...

Definition 3.1.14. Every set A determines one distinguished function $A \to A$, denoted 1_A or id_A, defined by

$$\forall a \in A : \quad 1_A(a) = a \quad .$$

This function is called the *identity on A*. ⌟

In other words, the identity on A leaves every element of A alone.

This may seem too simple-minded a notion to deserve a formal definition; on the contrary, it is very important. We will soon come back to it, and use it to define several other notions.

do not distribute

Example 3.1.15. If A is a subset of B, then we can define another simple-minded function, $\iota : A \to B$, by prescribing $\forall a \in A$

$$\iota(a) = a \quad .$$

This may be called the 'inclusion' function. Applying the inclusion function $A \to B$ amounts to viewing an element $a \in A$ as 'itself', but as an element of B rather than of A. If $A = B$, the corresponding inclusion function is the identity function of Definition 3.1.14. ⌟

Example 3.1.16. If $\sim$ is an equivalence relation on A, we have a function $\pi : A \to A/\sim$ from A to the quotient, defined by

$$\pi(a) = [a]_\sim$$

$\forall a \in A$. Indeed, every element of A determines exactly one equivalence class, that is, an element of $A/\sim$. I call this function the 'natural projection' from A to $A/\sim$.

For instance, for $A = \mathbb{Z}$ and the relation $\sim$ defined by $a \sim b \iff b - a$ is even, we verified that $\mathbb{Z}/\sim$ consists of two elements, say $E = [0]_\sim$ (= the set of even integers) and $O = [1]_\sim$ (= the set of odd integers). Cf. Example 2.4.2. Then the natural projection

$$\pi : \mathbb{Z} \to \mathbb{Z}/\sim = \{E, O\}$$

simply sends n to E if n is Even, and to O if n is Odd. ⌟

Example 3.1.17. If A, B are sets, and z is some fixed element of B, then we can define a function $c : A \to B$ by 'sending everything to z': $\forall a \in A, c(a) = z$. Such a function is called a 'constant' function.

For example, if A is any set, I can define a function $A \to \mathcal{P}(A)$ by sending everything to $\emptyset$, or everything to A: these are two constant functions.

There is a more interesting 'natural' function $\sigma : A \to \mathcal{P}(A)$: for every $a \in A$, define $\sigma(a) = \{a\}$. This is perfectly legal: given any a in A, $\{a\}$ is a subset of A, that is, an element of $\mathcal{P}(A)$. *Warning:* Writing $\sigma(a) = a$ would be incorrect—because a is an element of A, while the image $\sigma(a)$ must be an element of $\mathcal{P}(A)$, that is, a *subset* of A. It should be written as such. ⌟

Example 3.1.18. Let S, T be sets, and consider the Cartesian product $S \times T$. Then there are two natural[2] functions, $\pi_S : S \times T \to S$ and $\pi_T : S \times T \to T$, defined by

$$\pi_S((s,t)) = s \quad , \quad \pi_T((s,t)) = t$$

$\forall (s,t) \in S \times T$. These are called *projections* onto the first, resp. the second factor. ⌟

3.1.3 Homework

3.1.1: For every set A, there is exactly one *function* $f : A \to A$ whose graph determines an *equivalence relation* on A. What is this function?

3.1.2: Let A, B be sets, and $f : A \to B$ a function. Let S and T be subsets of A.

- Is $f(S \cap T)$ necessarily equal to $f(S) \cap f(T)$? Give a proof or find a counterexample.
- Is $f(S \cup T)$ necessarily equal to $f(S) \cup f(T)$? Proof or counterexample.

3.1.3: Let A, B be sets, and $f : A \to B$ a function. Let X and Y be subsets of B.

[2] In general there are of course many others, but these are 'natural' in the sense that they use the given information in the 'most natural' ('simplest') way.

do not distribute

- Is $f^{-1}(X \cap Y)$ necessarily equal to $f^{-1}(X) \cap f^{-1}(Y)$? Proof or counterexample.
- Is $f^{-1}(X \cup Y)$ necessarily equal to $f^{-1}(X) \cup f^{-1}(Y)$? Proof or counterexample.

▷ **3.1.4:** Let A denote a set $\{F, T\}$ with exactly two elements, and let S be any set. Prove that giving a function $S \to A$ (that is, an element of A^S) is 'the same as' giving a subset of S (that is, an element of $\mathcal{P}(S)$). In other words, explain how to define a subset of S for every function $S \to \{F, T\}$, and how to define a function $S \to \{F, T\}$ for every subset of S, in compatible ways. (This is the reason why $\mathcal{P}(S)$ is also denoted 2^S; cf. Definition 3.1.12. Here, '2' stands for a set with two elements.)

▷ **3.1.5:** Let $f : A \to B$ be a function. Prove that the set of nonempty inverse images of f forms a partition of A, and determine the corresponding equivalence relation.

3.2 *Functions: Composition, inverses, isomorphisms.*

3.2.1 *Functions, again*

Here is a reminder of the situation. *Functions* $f : A \to B$ (that is, 'from a set A to a set B') are (determined by) certain relations $\Gamma_f \subseteq A \times B$. We call Γ_f the 'graph' of f. Graphs are special kind of relations: if Γ_f is the graph of f, then we require that

$$(\forall a \in A)\,(\exists! b \in B) : \quad (a, b) \in \Gamma_f \quad .$$

The funny symbol $\exists!$ means 'there exists a unique…'. We write

$$f(a) = b$$

to mean that the pair (a, b) is an element of Γ_f.

Before we do anything else with functions, we should decide what it means for two functions $f_1, f_2 : A \to B$ to be *equal.* (For f_1 and f_2 to be comparable, they must have the same source A and the same target B.) Given that we are thinking of functions in terms of relations, the natural notion is that the graphs of f_1 and f_2 coincide: $\Gamma_{f_1} = \Gamma_{f_2}$. This translates into the following reasonable prescription once we take the 'functional' viewpoint:

Lemma 3.2.1. *Let f_1 and f_2 be functions $A \to B$. Then $f_1 = f_2$ (that is, $\Gamma_{f_1} = \Gamma_{f_2}$) if and only if*

$$\forall a \in A : \quad f_1(a) = f_2(a) \quad .$$

Proof. $\Longleftarrow$) By definition of graph of f_1, and for all $(a, b) \in A \times B$,

$$(a, b) \in \Gamma_{f_1} \iff b = f_1(a) \quad .$$

Since $f_1(a) = f_2(a)$, this statement is equivalent to

$$(a, b) \in \Gamma_{f_1} \iff b = f_2(a) \quad ,$$

and hence to

$$(a, b) \in \Gamma_{f_1} \iff (a, b) \in \Gamma_{f_2}$$

by the definition of the graph of f_2. This proves that $\Gamma_{f_1} = \Gamma_{f_2}$.

$\Longrightarrow$) The converse is similar, and left to the reader. □

do not distribute

3.2.2 *Composition of functions*

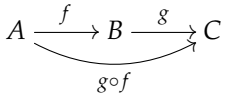
In appropriate situations, functions may be *composed* with one another. This is something else you have encountered in calculus (leading to the *chain rule*); the following definition should not surprise you.

Definition 3.2.2. Let A, B, C be sets, and let $f : A \to B$, $g : B \to C$ be function. The *composition* of f and g, denoted $g \circ f$ ('g composed with f') is the function $A \to C$ defined by

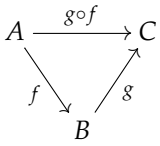
$$(g \circ f)(a) = g(f(a)) \tag{3.1}$$

$\forall a \in A$. ⌟

We can draw diagrams like this:



or like this:



to represent compositions. Such diagrams are *commutative*, in the sense that traveling from one point of the diagram to another point 'following the arrows' produces *the same* function, regardless of the path you take. In these small diagrams, the only options are to go 'directly' from A to C with the function $g \circ f$, or to go along f first, reaching B, and then along g, reaching C. The diagrams 'commute' because these two functions agree, as prescribed by (3.1).

Proposition 3.2.3. *The operation of composition of functions is associative. That is: if A, B, C, D are sets, and $f : A \to B$, $g : B \to C$, $h : C \to D$ are functions, then*

$$h \circ (g \circ f) = (h \circ g) \circ f \quad .$$

Proof. By Lemma 3.2.1, it suffices to prove that the effect of both sides on an arbitrary element $a \in A$ is the same. And indeed $\forall a \in A$,

$$(h \circ (g \circ f))(a) = h((g \circ f)(a)) = h(g(f(a))) \quad ;$$

while

$$((h \circ g) \circ f)(a) = (h \circ g)(f(a)) = h(g(f(a))) \quad :$$

these values agree, as needed. □

Note that the operation of composition is very much *not* 'commutative': the composition $g \circ f$ is only defined when the target of f coincides with the source of g; in general, if $g \circ f$ is defined there is no reason why $f \circ g$ should even be defined, let alone equal $g \circ f$.

Composing with the identity does not change a function:

Proposition 3.2.4. *Let $f : A \to B$ be a function. Then*

$$f \circ 1_A = f \quad \textit{and} \quad 1_B \circ f = f \quad .$$

do not distribute

Proof. By definition of composition and of the identity function,

$$(f \circ 1_A)(a) = f(1_A(a)) = f(a)$$

$\forall a \in A$; thus $f \circ 1_A = f$, by Lemma 3.2.1. The proof of the second equality is analogous, and left to the reader. □

Note again that the order in which we write the functions in a composition is important: if $f : A \to B$ is a function, and $A \neq B$, then $f \circ 1_A = f$ as observed in the proposition, but $1_A \circ f$ makes no sense, because the target of f (i.e., B) does not agree with the source of 1_A (i.e., A). On the other hand, $1_B \circ f$ *does* make sense, and equals f by the proposition. One subscript makes the difference between making sense and not making sense.

3.2.3 *Inverses*

My next goal is to introduce several interesting types of functions. There are many ways to do this, and I will choose to use *inverses* to streamline the discussion.

Definition 3.2.5. Let $f : A \to B$ be a function. A function $g : B \to A$ is

- a *left inverse* of f if $g \circ f = 1_A$;
- a *right inverse* of f if $f \circ g = 1_B$;
- a *two-sided inverse* of f if it is both a left and a right inverse. ⌟

Example 3.2.6. The function $f : \mathbb{R} \to \mathbb{R}$ defined by $f(x) = x^2$ has no left inverse and no right inverse. Indeed, suppose $g : \mathbb{R} \to \mathbb{R}$ is a left inverse, so that $g \circ f = 1_{\mathbb{R}}$. Then

$$g \circ f(-1) = -1 \quad \text{and} \quad g \circ f(1) = 1 \quad ,$$

that is

$$g(f(-1)) = -1 \quad \text{and} \quad g(f(1)) = 1 \quad ;$$

but $f(-1) = (-1)^2 = 1$ and $f(1) = 1^2 = 1$, so this would say

$$g(1) = -1 \quad \text{and} \quad g(1) = 1 \quad .$$

This is a problem, since if g is a function then there is a *unique* $b \in \mathbb{R}$ such that $g(1) = b$. Thus g is not a function, contrary to our assumption; this contradiction shows that f has no left inverse.

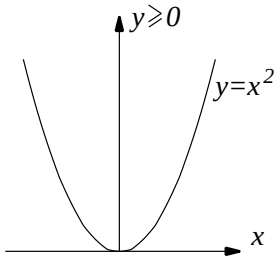
On the other hand, suppose that $g : \mathbb{R} \to \mathbb{R}$ is a right inverse to f. Then $f \circ g = 1_{\mathbb{R}}$, and in particular

$$-1 = 1_{\mathbb{R}}(-1) = f \circ g(-1) = f(g(-1)) = g(-1)^2 \quad .$$

This is a problem, because the square of a real number is necessarily nonnegative: so if $g(-1)^2 = -1$, then $g(-1)$ cannot be a real number. This shows that no function g can be a right inverse to f, so f has no right inverse. ⌟

Example 3.2.7. Now take 'the same' function $f(x) = x^2$, but defined after restricting the target: $\mathbb{R} \to \mathbb{R}^{\geq 0}$. (So this is not really the same function!)

do not distribute



We still do not have a left inverse, for the same reason as in Example 3.2.6. But now we *do* have right inverses, because we have eliminated the elements in the target (such as -1) that gave us a problem in Example 3.2.6. For example, if we define $g : \mathbb{R}^{\geq 0} \to \mathbb{R}$ by setting $g(y) = \sqrt{y}$ (which we can do, since y is now guaranteed to be nonnegative), then $\forall y \in \mathbb{R}^{\geq 0}$ we have

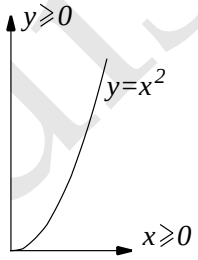
$$f \circ g(y) = f(g(y)) = g(y)^2 = (\sqrt{y})^2 = y \quad ,$$

so that indeed $f \circ g = 1_{\mathbb{R}^{\geq 0}}$. In fact, note that the function $h(y) = -\sqrt{y}$ is different from g, but is *also* a right inverse to f:

$$f \circ h(y) = f(h(y)) = h(y)^2 = (-\sqrt{y})^2 = y$$

$\forall y \geq 0$. This shows that the right inverse of a given function is not *unique* in general. (The same applies to left inverses: they may not exist, as in this example; and they may not be unique even when they exist.) ⌟

Example 3.2.8. We can go one step further, and consider $f(x) = x^2$ after restricting both the source and the target: $\mathbb{R}^{\geq 0} \to \mathbb{R}^{\geq 0}$:



Note that the appearance of the graph has changed now, because we are throwing away the negative part of the source.

I claim that this function has two-sided inverses. Indeed, the function $g(y) = \sqrt{y}$ is defined from $\mathbb{R}^{\geq 0}$ to $\mathbb{R}^{\geq 0}$, and

$$g \circ f(x) = g(x^2) = \sqrt{x^2} = x \quad , \quad f \circ g(y) = f(\sqrt{y}) = (\sqrt{y})^2 = y$$

$\forall x, y \in \mathbb{R}^{\geq 0}$: therefore, $g(y) = \sqrt{y}$ now works as a two-sided inverse of f. ⌟

Remark 3.2.9. Even such simple examples teach us a lot. Now we know that left or right inverses need not exist, and they need not be unique even when they exist. On the other hand, the reason why the right inverse is not unique in Example 3.2.7 won't work in Example 3.2.8; and the left and right inverse coincide in Example 3.2.8. You should convince yourself (and another skeptical reader if need be) that this is a general feature of the definitions. Do Exercise 3.2.4: you will establish that if a function f has both a left and a right inverse, then they coincide: and then it follows that f has a two-sided inverse and that it is unique). ⌟

Since (as observed in Remark 3.2.9) a two-sided inverse is unique if it exists, we can unambiguously assign a symbol to it.

do not distribute

Definition 3.2.10. Let $f : A \to B$ be a function and assume that f has a two-sided inverse. Then we say that f is an *isomorphism of sets.* The two-sided inverse of f is simply called the *inverse* of f, and it is denoted f^{-1}. ⌟

Thus, $f \circ f^{-1} = 1_B$ and $f^{-1} \circ f = 1_A$; but we can write such things *only if* f is an isomorphism. Don't confuse this new use of the symbol f^{-1} with the use introduced in Definition 3.1.2 (cf. Remark 3.1.3). Only if f is an isomorphism can we view f^{-1} as a function as indicated in Definition 3.2.10. If $f : A \to B$ is not an isomorphism, f^{-1} is just a notation associating with every $b \in B$ the subset of elements $a \in A$ such that $f(a) = b$; it is *not* a function $B \to A$. *If f is an isomorphism,* then this subset consists of a single element a, and we think of $f^{-1}(b)$ as an element of A rather than as the corresponding subset $\{a\} \subseteq A$.

Remark 3.2.11. Since we are (mostly) dealing with sets in these notes, I may just write 'isomorphism' rather than 'isomorphism of sets'. When you move on to fun subjects such as abstract algebra, you will run into isomorphisms *of groups, of rings,* and others. In each case, these will be (certain kind of) functions which admit two-sided inverses, in line with the example of sets we are exploring here. ⌟

3.2.4 *Homework*

3.2.1: Let f, g denote functions $\mathbb{R} \to \mathbb{R}$.

- Assume $(\forall x)\ f(x) = x - 1$. Does there exist a function g such that $g \circ f(x) = x^2 - 2x$? Find one, or prove that no such function exists.
- Assume $(\forall x)\ f(x) = x^2 - 2x$. Does there exist a function g such that $g \circ f(x) = x - 1$? Find one, or prove that no such function exists.

3.2.2: We have observed that if A, B, C are sets and $f : A \to B$, $g : B \to C$ are functions, then we can define $g \circ f$ but we cannot define $f \circ g$ in general. However, if $A = B = C$, then we can define both $g \circ f$ and $f \circ g$ as functions $A \to A$. Prove that necessarily $g \circ f = f \circ g$ in this case, or construct a counterexample.

▷ **3.2.3:** Let $f : A \to B$ and $g : B \to C$ be two functions.

- Prove that if $g \circ f$ has a left inverse, then f has a left inverse.
- Prove that if $g \circ f$ has a right inverse, then g has a right inverse.

▷ **3.2.4:** Let $f : A \to B$ be a function, and assume that f has a left inverse g and a right inverse h. Prove that $g = h$. (Hint: Use Proposition 3.2.3.) Deduce that if f has a left and a right inverse, then it has a two-sided inverse.

3.2.5: Let $f : A \to B$ and $g : B \to C$ be functions.

- Prove that if both f and g have a left inverse, then the composition $g \circ f$ has a left inverse.
- Prove that if both f and g have a right inverse, then the composition $g \circ f$ has a right inverse.

▷ **3.2.6:** Let $f : A \to B$ and $g : B \to C$ be two isomorphisms of sets (cf. Definition 3.2.10). Prove that $g \circ f$ is also an isomorphism, by finding its two-sided inverse.

3.3 *Injectivity and surjectivity.*

3.3.1 *Injections and surjections*

You are probably familiar with the definitions that follow. The surprise down the road will be that they have a lot to do with the subject of §3.2.3.

do not distribute

Definition 3.3.1. Let $f : A \to B$ be a function. We say that

- f is *injective*, or *one-to-one*, if $\forall a_1, a_2 \in A$

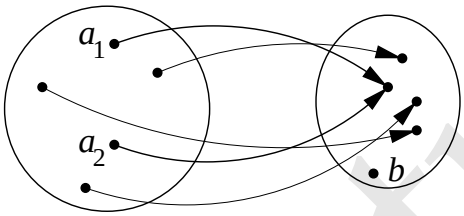
$$f(a_1) = f(a_2) \implies a_1 = a_2 \quad .$$

- f is *surjective*, or *onto*, if

$$(\forall b \in B)\,(\exists a \in A) : \quad f(a) = b \quad .$$

Injective functions are also called *injections;* surjective functions are called *surjections.* A function that is both injective and surjective is *bijective*, or a *bijection*, or a *one-to-one correspondence.* ⌟

In pictures, a function like this:



is neither injective nor surjective: the two arrows coming together tell us that there are elements $a_1, a_2 \in A$ such that $f(a_1) = f(a_2)$ while $a_1 \neq a_2$ (so f is not injective); and there is a $b \in B$ that is not reached by any arrow (so f is not surjective).

Example 3.3.2. For every set A, the identity function 1_A is both injective and surjective. ⌟

Injectivity/surjectivity can be translated easily in terms of the graph Γ_f of f:

- f is injective iff for all $b \in B$, there is *at most one* $a \in A$ such that $(a, b) \in \Gamma_f$;
- f is surjective iff for all $b \in B$, there is *at least one* $a \in A$ such that $(a, b) \in \Gamma_f$.

This is hopefully clear.

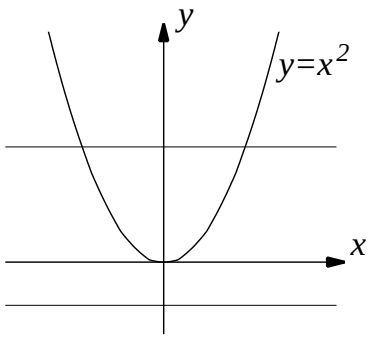
Remark 3.3.3. One can also view injectivity/surjectivity from the point of view of the notation introduced in Definition 3.1.2. A function $f : A \to B$ is injective if and only if for every $b \in B$ the inverse image $f^{-1}(b)$ consists of at most one element, and f is surjective if and only if $f^{-1}(b)$ consists of at least one element for every $b \in B$.

In fact, f is injective if and only if $f^{-1}(b)$ consists of *exactly one* element for every $b \in imf$. A function $f : A \to B$ is surjective if and only if $\operatorname{im} f = B$: indeed, $\operatorname{im} f$ consists of all the elements of B which are images by f of element in A, and f is surjective if and only if *every* element of B is the image of some element in A. ⌟

Example 3.3.4. In $\mathbb{R}^2 = \mathbb{R} \times \mathbb{R}$, the set of pairs (x, y) with fixed $y = b$ is a horizontal line, so points in a graph with fixed $y = b$ are intersections of the graph with a horizontal line. Thus, a function $f : \mathbb{R} \to \mathbb{R}$ is injective if and only if its graph in $\mathbb{R}^2$ meets every horizontal line in *at most one* point, while $f : \mathbb{R} \to \mathbb{R}$ is surjective if and only if its graph meets every horizontal line in *at least one* point[3].

[3] Recall from Remark 3.1.4 that the definition of 'function' prescribes that the graph meets every *vertical* line exactly once. Don't confuse vertical lines and horizontal lines.

do not distribute



Our running example of $f : \mathbb{R} \to \mathbb{R}$ defined by $f(x) = x^2$ is neither injective nor surjective, since there are horizontal lines meeting the graph in two points and horizontal lines missing the graph entirely. ⌟

There are 'prototypical' examples of injective/surjective functions with which we are already familiar (cf. Examples 3.1.15 and 3.1.16). They are both generalization of the identity function, which (as I have already observed: Example 3.3.2) is both injective and surjective. Here they are.

Example 3.3.5. [*Inclusions*] Let B be any set, and let $A \subseteq B$ be a subset. The 'inclusion function' of Examples 3.1.15 is the function $\iota : A \to B$ obtained by 'restricting the identity':

$$(\forall a \in A) : \quad \iota(a) = a \quad .$$

This function is clearly injective: if $\iota(a_1) = \iota(a_2)$, then $a_1 = a_2$. Although not much is going on here, it is occasionally useful to view an inclusion of sets as a function. In fact, an injective function $S \to T$ just reproduces a 'copy' of S as a subset of T, so in some loose sense every injective function looks like an inclusion. To stress this, we sometime use a hooked arrow to denote them: if I write

$$f : S \hookrightarrow T$$

then it is understood that f is injective. ⌟

Example 3.3.6. [*Projections*] Let $\sim$ be an equivalence relation on a set A. Then we can define a function as in Example 3.1.16,

$$\pi : A \to A/\!\sim \quad ,$$

by setting

$$(\forall a \in A) : \quad \pi(a) = [a]_\sim \quad .$$

This function is surjective: indeed, every equivalence class (that is, every element of $A/\!\sim$) is the equivalence class of some element of A. We call π the 'natural projection' from A to the quotient $A/\!\sim$.

This example is also typical of all surjective functions, as we will soon discover. Surjective functions are often denoted by a double-tipped arrow: if I write

$$f : S \twoheadrightarrow T$$

then it is understood that f is surjective. ⌟

3.3.2 *Injective vs. left inverses, surjective vs. right inverses*

If you think a moment about Examples 3.2.7 and 3.2.8, in view of Example 3.3.4, you will likely notice a relation between injectivity/surjectivity and the presence of left/right inverses. This relation is as neat as can be:

do not distribute

Theorem 3.3.7. *Let $f : A \to B$ be a function, and assume $A \neq \emptyset$. Then*

- *f is injective if and only if it has a left inverse.*
- *f is surjective if and only if it has a right inverse.*

A simple consequence of this statement is a characterization of isomorphisms:

Corollary 3.3.8. *Let $f : A \to B$ be a function. Then f is an isomorphism of sets if and only if it is a bijection (that is, it is both injective and surjective).*

Look back at Definitions 3.2.10 and 3.3.1! They don't much look like they are talking about the same concepts, but they are!

Proof of Corollary 3.3.8. By Theorem 3.3.7, f is both injective and surjective if and only if it has both a left and a right inverse, hence (Exercise 3.2.4) if and only if it has a two-sided inverse, hence (Definition 3.2.10) if and only if it is an isomorphism of sets. □

One tool in the proof of Theorem 3.3.7 will be the following straightforward lemma, whose proof is just an exercise in handling the notation.

Lemma 3.3.9. *Let $f : A \to B$ and $g : B \to C$ be functions. Then*

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

Proof. The first statement may be proven contrapositively. (It's just as easy to give a direct argument.) If f is *not* injective, then there exist $a_1, a_2 \in A$, with $a_1 \neq a_2$, such that $f(a_1) = f(a_2)$. But then

$$g \circ f(a_1) = g(f(a_1)) = g(f(a_2)) = g \circ f(a_2) \quad ;$$

since $a_1 \neq a_2$, this shows that $g \circ f$ is not injective, proving the contrapositive of the first statement.

As for the second, argue directly. Assume that $g \circ f$ is surjective. Then for all $c \in C$ there exists an $a \in A$ such that $g \circ f(a) = c$. This says that $g(f(a)) = c$, showing that there exists an element $b = f(a)$ of B such that $g(b) = c$. Thus for all $c \in C$ there exists a $b \in B$ such that $g(b) = c$, and this proves that g is surjective. □

You may be amused at noticing that once Theorem 3.3.7 is established, the statement of Lemma 3.3.9 is seen to be equivalent to the statement of Exercise 3.2.3, which you have (hopefully) worked out already. In solving Exercise 3.2.3, you should have used the definitions of left- and right inverse directly, without appealing to this lemma.

3.3.3 *Proof of Theorem 3.3.7*

Proof. —Let's first take care of injectivity.

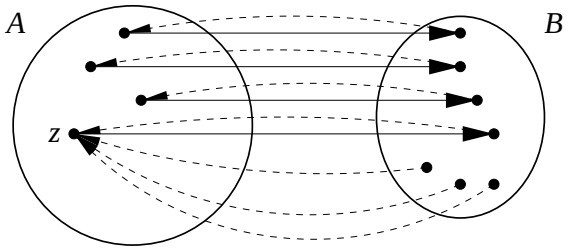
I claim that if f has a left inverse, then it is injective. Indeed: assume that f has a left inverse g. Then $g \circ f = 1_A$. We have observed that 1_A is injective (Example 3.3.2), so by Lemma 3.3.9 this proves that f is injective.

Conversely, assume that f is injective; we will construct a left inverse $g : B \to A$ for f. As $A \neq \emptyset$, I can find an element $z \in A$. To define g, let $b \in B$ be any element. If b is not in the image of f, then let $g(b) = z$. If b *is* in the image of f, then $b = f(a)$ for a *unique* $a \in A$ (since f is injective). In this case, define $g(b) = a$:

$$g(b) = \begin{cases} z & \text{if } b \notin \text{im}(f) \\ a & \text{if } b = f(a) \end{cases} .$$

do not distribute

The following picture attempts to depict this construction. The solid arrows stand for f, and the dashed ones for g; every element of B reached by a solid arrow is sent back by a dashed arrow to its (unique) preimage; the elements that are not reached by any solid arrow are all sent by a dashed arrow to the fixed element $z \in A$:



Now let $a \in A$ be any element. Then

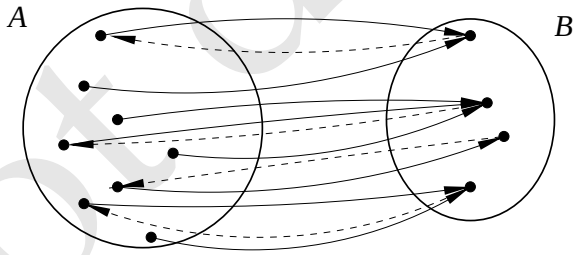
$$g \circ f(a) = g(f(a)) = a \quad :$$

in this computation only the second part of the prescription for g is used. This shows that $g \circ f(a) = 1_A(a)$ for all $a \in A$. It follows that $g \circ f = 1_A$ (by Lemma 3.2.1), and this proves that g is a left inverse to f.

—Next, let's deal with surjectivity.

I claim that if f has a right inverse, then it is surjective. Indeed: assume that f has a right inverse g. Then $f \circ g = 1_B$. We have observed that 1_B is surjective (Example 3.3.2), so by Lemma 3.3.9 this proves that f is surjective.

Conversely, assume that f is surjective; we will construct a right inverse $g : B \to A$ for f. Since f is surjective, for all $b \in B$ there exists some $a \in A$ such that $f(a) = b$. For every b, choose[4] arbitrarily an a_b such that $f(a_b) = b$, and then define $g(b) = a_b$. Again, here is a heuristic visual interpretation of this operation:



With this choice,

$$f \circ g(b) = f(g(b)) = f(a_b) = b$$

for all $b \in B$, showing that $f \circ g = 1_B$. This proves that f has a right inverse, as promised. □

3.3.4 *Homework*

▷ **3.3.1:** • Suppose A and B are finite sets, and $|B| < |A|$. Prove that there are no injective functions $A \to B$. (This is called the 'pigeonhole principle': if you have n pigeons in $m < n$ holes, then necessarily at least one hole must be shared by more than one pigeon.)

[4] This is actually a subtle point, involving the *axiom of choice.* We will run into this again, but you can really learn about it in texts and courses on foundation of mathematics.

do not distribute

• Suppose A and B are finite sets with the same number of elements, and that $f : A \to B$ is an injective function. Prove that f is also surjective.

• Now suppose that A and B are both *infinite* sets, and that $f : A \to B$ is an injective function. Is f necessarily surjective? Either give a proof or a counterexample.

3.3.2: How many functions $f : A \to B$ are *not* injective, if A and B have exactly 17 elements each? (Hint: There are 17! bijections $A \to B$, as we will see in Proposition 3.4.10.)

3.3.3: Let $f : A \to B$, $g : B \to C$ be functions.

- Assume that $g \circ f$ is injective and that f is surjective. Prove that g is injective.
- Assume that $g \circ f$ is surjective and that g is injective. Prove that f is surjective.

▷ **3.3.4:** Let $f : A \to B$ and $g : B \to C$ be functions.

- Prove that if both f and g are injective, then the composition $g \circ f$ is also injective.
- Prove that if both f and g are surjective, then the composition $g \circ f$ is also surjective.

(Do this directly from the definitions, without appealing to Exercise 3.2.5.)

3.3.5: Let $f : A \to B$ and $g : B \to C$ be two bijections. Prove that $g \circ f$ is also a bijection.

3.3.6: Let $f : A \to B$ be a function.

- Prove that f is injective if and only if $f^{-1}(f(S)) = S$ for all subsets $S \subseteq A$.
- Prove that f is surjective if and only if $f(f^{-1}(T)) = T$ for all subsetes $T \subseteq B$.

3.3.7: • Give an example of sets A, B, and functions $f : A \to B$, $g, h : B \to C$, such that $g \neq h$ and yet $g \circ f = h \circ f$.

• Prove that if f is *surjective,* then no such example exists.

▷ **3.3.8:** We have seen (cf. Remark 3.2.9) that right- or left inverses are not unique in general. Discuss this fact, in light of the proof of Theorem 3.3.7: what makes the functions constructed in that proof 'not unique'?

3.3.9: Let S be any set. Construct an injective function from S to its power set $\mathcal{P}(S)$.

3.4 *Isomorphisms of sets, a.k.a. bijections*

3.4.1 *Isomorphisms of sets*

As we observed above, Theorem 3.3.7 has the following consequence: A function $f : A \to B$ is an isomorphism of sets (that is, it has a two-sided inverse) if and only if it is a bijection (that is, both injective and surjective). This was Corollary 3.3.8. We are now going to study isomorphisms more in depth.

Definition 3.4.1. Two sets A, B are *isomorphic* if there exists an isomorphism (that is, by Corollary 3.3.8, a bijection) $A \to B$. We denote this condition by writing $A \cong B$. ⌟

We can also say that 'A is isomorphic to B (as sets)', or that 'A and B are isomorphic to each other (as sets)'; or 'B is an isomorphic copy of A'; or 'A, B are bijective'; or A, B are in one-to-one correspondence; or... (you can imagine other variations on the same theme).

Example 3.4.2. Two *finite* sets are isomorphic if and only if they have the same number of elements. Indeed, say $A = \{a_1, \ldots, a_m\}$ and $B = \{b_1, \ldots, b_n\}$, with $n = |A|$ and $m = |B|$. If $m = n$, then the function $f : A \to B$ defined by $f(a_i) = b_i$ is an isomorphism. Conversely, if there is an *injective* function $A \to B$, then $m \leq n$

do not distribute

(Exercise 3.3.1); so if there are injective maps $A \to B$ and $B \to A$, as is the case if A and B are isomorphic, then $m = n$. ⌟

In the homework (Exercise 3.4.1) you will verify that 'isomorphic' is an equivalence relation. In practice isomorphic sets are essentially indistinguishable as sets: if I have a bijection between A and B then I can use it to label elements of one sets by elements of the other, and conversely. I could say that an isomorphism between A and B allows me to 'identify' them as sets.

Remark 3.4.3. As I already mentioned, there are many notions of 'isomorphism', depending on the context: you will encounter isomorphisms of *groups*, of *rings*, of *fields*, and of many other structures, later on in your studies. Here we are just dealing with sets, so I'll use 'isomorphic' as shorthand for 'isomorphic *as sets*'. Occasionally I will opt for the synonym 'bijective', which avoids this ambiguity. ⌟

Example 3.4.4. The sets $\{*, \circ, \bullet\}$ and $\{a, b, c\}$ (where a, b, c are distinct elements) are isomorphic: here is an isomorphism

$$\begin{cases} * \mapsto a \\ \circ \mapsto b \\ \bullet \mapsto c \end{cases}.$$

Here is another one:

$$\begin{cases} * \mapsto b \\ \circ \mapsto c \\ \bullet \mapsto a \end{cases}.$$

Of course there are many functions $\{*, \circ, \bullet\} \to \{a, b, c\}$ that are *not* isomorphisms. For example,

$$\begin{cases} * \mapsto a \\ \circ \mapsto a \\ \bullet \mapsto b \end{cases}$$

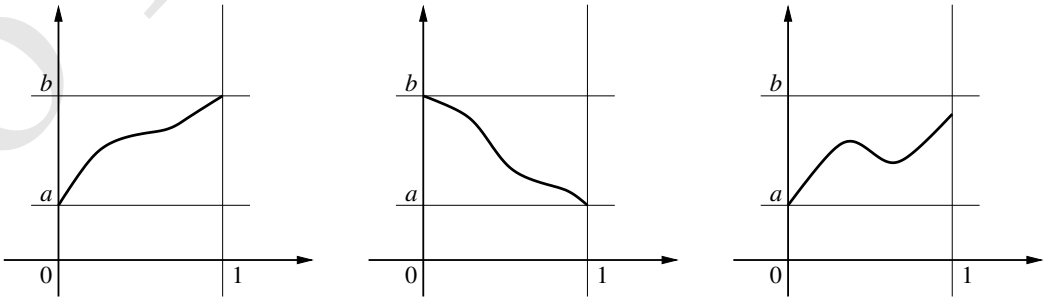
is not bijective. However, the existence of *one* bijection suffices to declare that the sets are isomorphic. ⌟

3.4.2 *Examples of isomorphisms*

Example 3.4.5. Any two nonempty open intervals (a_1, b_1), (a_2, b_2) in $\mathbb{R}$ are isomorphic, and they are isomorphic to $\mathbb{R}$.

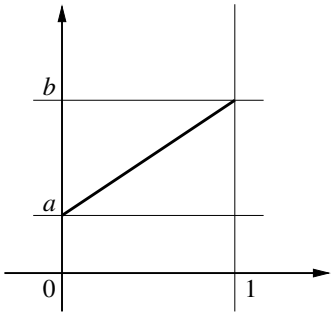
Indeed, it is enough (by transitivity of the isomorphism relation, which you will prove in Exercise 3.4.1) to verify that every nonempty open interval (a, b) is isomorphic to $(0, 1)$, and that $\mathbb{R}$ is isomorphic to $(0, 1)$.

To see that $(0, 1)$ and (a, b) are isomorphic, we only need to observe that there are bijections from $(0, 1)$ to (a, b). The first two graphs here



do not distribute

show two such bijections. There of course are lots of functions that are *not* bijections: the third graph is not the graph of a bijection (in fact, the corresponding function is neither injective nor surjective). If you want an explicit formula for an isomorphism, consider the simplest graph of a bijection $(0,1) \to (a,b)$:



You need a linear function $f : (0,1) \to (a,b)$ such that $f(0) = a$ and $f(1) = b$. This should do:

$$t \mapsto a + t(b-a) \quad .$$

If you need to convince a very skeptical person that this is a bijection, you can argue that it is both injective and surjective; or you can construct a two-sided inverse of f. Define $g : (a,b) \to (0,1)$ by

$$r \mapsto \frac{r-a}{b-a} \quad .$$

I claim that f and g are inverses of each other. Indeed,

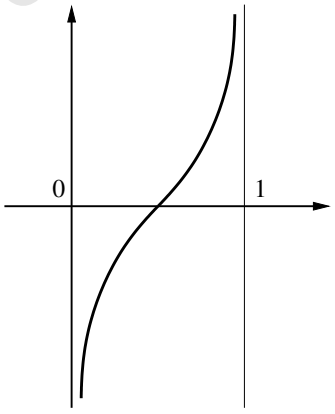
$$g \circ f(t) = g(f(t)) = g(a + t(b-a)) = \frac{a + t(b-a) - a}{b-a} = t \quad ,$$

proving that $g \circ f = \mathrm{id}_{(0,1)}$, and

$$f \circ g(r) = f(g(r)) = f\left(\frac{r-a}{b-a}\right) = a + \frac{r-a}{b-a}(b-a) = a + r - a = r \quad ,$$

proving that $f \circ g = \mathrm{id}_{(a,b)}$. This shows that f has a two-sided inverse, so it is an isomorphism.

To see that $(0,1)$ is isomorphic to $\mathbb{R}$, you can consider any function whose graph looks like this:

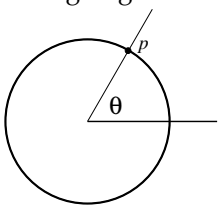


(that is: a monotone increasing function with asymptotes at 0 and 1). For example, $f(x) = \tan(\pi x - \pi/2)$ would do just fine. Viewed as a function $(0,1) \to \mathbb{R}$, this function is both injective and surjective, so it is an isomorphism. ⌟

do not distribute

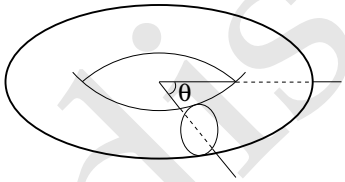
Example 3.4.6. Recall that S^1 is the official name for a circle. Of course all circles are isomorphic to one another (right?), and all ellipses are isomorphic to a circle (right?), so 'S^1' is actually a codename for a rather general concept[5].

The closed-open interval $[0, 2\pi)$ is bijective to S^1. Indeed, a point p on a circle is determined by an angle, expressed (in radians) by a number between 0 (included) and 2π (excluded, because the corresponding point of the circle already corresponds to 0). If you know about complex numbers, then you can realize this function $[0, 2\pi) \to S^1$ explicitly by $t \mapsto e^{it}$ (the image is the set of complex numbers with magnitude 1, a circle centered at the origin in the complex plane). But all that is going on is captured by the picture

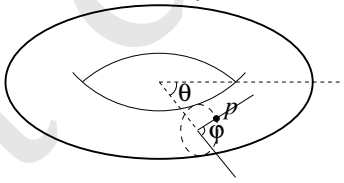


In a sense, we are taking the interval $[0, 2\pi)$ and bending it so as to identify 0 and 2π; we went through this already in Example 2.4.4. ⌟

Example 3.4.7. I claim that $S^1 \times S^1$ is isomorphic to a torus in a natural[6] way. To see this, think about what information you have to give in order to determine a point p on a torus T: once you decide a reference frame, you have to choose an angle $0 \leq \theta < 2\pi$ by which you are rotating about a vertical direction, and this determines a slice of T that is actually a circle:



Then you have to choose a point p on this circle, which you can do by selecting another angle $0 \leq \varphi < 2\pi$:



The choice of an angle between 0 and 2π amounts to the choice of a point of S^1 (Example 3.4.6). So the procedure illustrated above defines a function

$$S^1 \times S^1 \to T \quad ;$$

a moment's thought should reveal that this function is injective and surjective, so it is an isomorphism. ⌟

[5] I am cheating a little here: as we will see in a moment, S^1 is also isomorphic to an interval in $\mathbb{R}$, but we don't call such intervals 'S^1'. The point is that the most natural isomorphisms between a circle and, say, an ellipse preserve some structure that makes these objects 'look alike'; while the isomorphisms between S^1 and an interval do not preserve that structure. This has to do with *topology*. By the end of these notes we will know a tiny bit about topology, but to really give justice to these considerations would require more room than we will have.

[6] As it happens, $S^1 \times S^1$ will turn out to be 'abstractly' isomorphic (as sets) to lots of other things; even to $\mathbb{R}$ itself. The claim here is that there is a 'natural', *compelling* way to identify $S^1 \times S^1$ and a torus.

do not distribute

Remark 3.4.8. In §2.4 and elsewhere I have attempted to make the notion of quotient modulo an equivalence relation more concrete, describing a quotient $A/\sim$ by means of another set. I can now be a little more explicit as to what I meant by such a 'description': I meant a set isomorphic in a natural way to the given quotient. For instance, the quotient of $\mathbb{R}$ by the equivalence relation defined by $a \sim b \iff b - a \in \mathbb{Z}$ is *isomorphic* to a circle: this is the result of the discussion in Example 2.4.4. It is also bijective to the segment $[0, 1)$, as pointed out in the same example; my preference for the circle is due to more sophisticated 'geometric' types of information, that go beyond purely set-theoretic considerations.

We will soon see a more systematic way to concoct such bijections 'describing' a given quotient. ⌟

3.4.3 *The set of bijections from a set to itself*

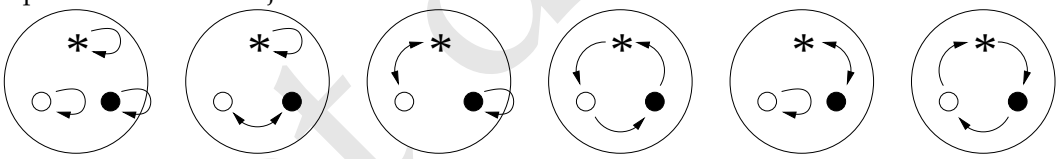
The set of bijections between two sets A and B has a very interesting structure, that is the prototype for one of the most important constructions in algebra. For elementary considerations nothing is lost if we assume that the two sets are *the same:* as pointed out earlier, if $A \cong B$, then we could use the elements of B to label the elements of A and conversely, so we might as well identify A and B, i.e., assume that $A = B$.

Let's then contemplate the set of bijections $f : A \to A$ from a set A to itself.

Example 3.4.9. Let $A = \{*, \circ, \bullet\}$ be a set with three elements. I claim that there are exactly six bijections $A \to A$, and here they are:

$$\begin{cases} * \mapsto * \\ \circ \mapsto \circ \\ \bullet \mapsto \bullet \end{cases} \quad \begin{cases} * \mapsto * \\ \circ \mapsto \bullet \\ \bullet \mapsto \circ \end{cases} \quad \begin{cases} * \mapsto \circ \\ \circ \mapsto * \\ \bullet \mapsto \bullet \end{cases} \quad \begin{cases} * \mapsto \circ \\ \circ \mapsto \bullet \\ \bullet \mapsto * \end{cases} \quad \begin{cases} * \mapsto \bullet \\ \circ \mapsto \circ \\ \bullet \mapsto * \end{cases} \quad \begin{cases} * \mapsto \bullet \\ \circ \mapsto * \\ \bullet \mapsto \circ \end{cases}$$

The first item in this list is the identity 1_A; the second swaps $\circ$ and $\bullet$, and so on. If you prefer pictures, here is a way to represent what each bijection does:



How do I know that this list is complete?

— $*$ can go to $*$, $\circ$, or $\bullet$;

— Once the image of $*$ has been decided, $\circ$ can only go to one of the remaining two elements in the target, because bijections are injective;

— By the same token, the image of $\bullet$ is determined once the images of $*$ and $\circ$ are determined: only one element in the target is left, so it must be the image of $\bullet$, by injectivity.

Running through these $3 \cdot 2 \cdot 1 = 6$ possibilities yields the list given above.

The composition of two bijections may be computed by following the fate of each element. For example, composing the second-to-last and the last items in the list:

$$\text{doing} \quad \begin{cases} * \mapsto \bullet \\ \circ \mapsto \circ \\ \bullet \mapsto * \end{cases} \quad \text{and then} \quad \begin{cases} * \mapsto \bullet \\ \circ \mapsto * \\ \bullet \mapsto \circ \end{cases}$$

do not distribute

will send $*$ to $\bullet$ first, and then $\bullet$ to $\circ$; $\circ$ to $\circ$ to $*$; and $\bullet$ to $*$ to $\bullet$. The result is

$$\begin{cases} * \mapsto \circ \\ \circ \mapsto * \\ \bullet \mapsto \bullet \end{cases}$$

which happens to be the third item in the list. Since the composition of two bijections is a bijection (Exercise 3.2.6) it is clear that composing any two items in the list will again produce an item in the list.

There is an annoyingly confusing point here: if f is the first mentioned bijection, and g is the second one, then the result we just computed is $g \circ f$, while our propensity to write from left to right might trick you into thinking it is $f \circ g$. (Some people pronounce $g \circ f$ as 'g after f' to remember this.) One has to keep track of these things, because the operation of composition is not 'commutative', even if the sources and targets of the bijections coincide. I already commented on this fact in the remarks following Proposition 3.2.3, and now we can verify it in practice in this concrete example. Indeed, compose again the last and the second-to-last items in the list, but in the opposite order:

$$\text{doing} \quad \begin{cases} * \mapsto \bullet \\ \circ \mapsto * \\ \bullet \mapsto \circ \end{cases} \quad \text{and then} \quad \begin{cases} * \mapsto \bullet \\ \circ \mapsto \circ \\ \bullet \mapsto * \end{cases}$$

sends $*$ to $\bullet$ to $*$, $\circ$ to $*$ to $\bullet$, and $\bullet$ to $\circ$ to $\circ$. This gives

$$\begin{cases} * \mapsto * \\ \circ \mapsto \bullet \\ \bullet \mapsto \circ \end{cases}$$

that is, the *second* item in the list. This is $f \circ g$, and it turns out to be different from $g \circ f$. The order in which you perform the operation matters.

Another general feature that we can check by hand in this example is the fact that bijections have inverses: for each item in the list, you will be able to 'compute' the inverse by simply implementing the prescription obtained by reversing the arrows. For example, the inverse of

$$\begin{cases} * \mapsto \circ \\ \circ \mapsto \bullet \\ \bullet \mapsto * \end{cases}$$

(the fourth element in the list) is

$$\begin{cases} * \leftarrow \circ \\ \circ \leftarrow \bullet \\ \bullet \leftarrow * \end{cases} \quad \text{that is,} \quad \begin{cases} * \mapsto \bullet \\ \circ \mapsto * \\ \bullet \mapsto \circ \end{cases}$$

(the sixth one), while the inverse of

$$\begin{cases} * \mapsto * \\ \circ \mapsto \bullet \\ \bullet \mapsto \circ \end{cases}$$

do not distribute

is itself.

Try a few more compositions. The last fact we have observed says that if s denotes the second bijection listed above, then $s \circ s = 1_A$. Verify that if t is the last one, then $t \circ t \circ t = 1_A$. (Thinking about the pictures makes this task particularly straightforward.) ⌟

Note that the second column in each diagram in Example 3.4.9 is just a permutation of the symbols $*$, $\circ$, $\bullet$: it must be so, since each symbol must appear, and each symbol must appear only once. Bijections from a set A to itself are often called 'permutations' of A; the set of bijections of A is usually denoted S_A, which is a common notation for the *permutation* (or *symmetric*) *group.*

The argument counting permutations, sketched in the example, generalizes to all finite sets.

Proposition 3.4.10. *Let A, B be finite sets, with $|A| = |B| = n$. Then there are exactly $n!$ bijections $A \to B$. Thus, if A has n elements, then S_A has $n!$ elements.*

Proof. Let $a_1, \ldots, a_n$ be the elements of A. To determine a bijection $f : A \to B$ we have to choose $f(a_1)$ among the n elements of B; then $f(a_2)$ must be a different element (by injectivity of f), so there are $n-1$ choices for $f(a_2)$; since $f(a_3) \neq f(a_1)$ and $f(a_3) \neq f(a_2)$ (again by injectivity), there are $n-2$ choices for $f(a_3)$. Arguing in the same way, we see that there are $n-k$ choices for $f(a_k)$ once $f(a_1), \ldots, f(a_{k-1})$ have been determined. In total, there are therefore[7]

$$n \cdot (n-1) \cdot (n-2) \cdots 2 \cdot 1 = n!$$

choices for f. □

The most essential properties of the operation of composition on the set S_A of bijections on a given set A are all mentioned in the simple instance examined in Example 3.4.9:

- The composition of two elements of S_A is an element of S_A;
- Composition is associative in S_A (Proposition 3.2.3);
- There exists an identity element 1_A such that $\forall f \in S_A$, $f \circ 1_A = 1_A \circ f = f$;
- Every $f \in S_A$ has an inverse: $\forall f \in S_A$, $\exists g \in S_A$ such that $f \circ g = g \circ f = 1_A$.

(Usually we denote g by f^{-1}.)

These four properties are extremely important; a *group* is a set endowed with an operation satisfying these same properties. The set S_A, considered with the operation of composition, is called the *symmetric group* on A. You already know a whole bunch of groups beside S_A: for example, the integers $\mathbb{Z}$ form a group when taken along with the operation $+$ of addition. Groups would be strong contenders for the title of 'most important notion in mathematics'. Entire fields of mathematics (e.g., representation theory) owe their existence to groups. Groups arise naturally in, for example, particle physics, chemistry, even music theory(!).

3.4.4 *Homework*

▷ **3.4.1:** Let $\mathcal{F}$ be a family of sets. Prove that the isomorphism relation $\cong$ defines an *equivalence* relation on $\mathcal{F}$. That is, show that (for all sets A, B, C in $\mathcal{F}$)

- $A \cong A$;
- If $A \cong B$, then $B \cong A$;

[7] Of course a formal version of this argument would use induction.

do not distribute

- If $A \cong B$ and $B \cong C$, then $A \cong C$.

3.4.2: • Is the function $f : \mathbb{R} \to \mathbb{R}$ defined by $f(x) = x^3$ a bijection? Why/why not?

- Is the function $f : \mathbb{R} \to \mathbb{R}$ defined by $f(x) = x^3 - x$ a bijection? Why/why not?

3.4.3: What common geometric shape would you choose as an isomorphic copy of $S^1 \times [0, 1]$?

3.4.4: A certain function $f : \mathbb{Z} \to \mathbb{Z}$ is defined by $f(x) = ax + b$, where a and b are integers. Assume that f is an isomorphism of sets, that is, a bijection. What can you say about a and b? What is the inverse of f?

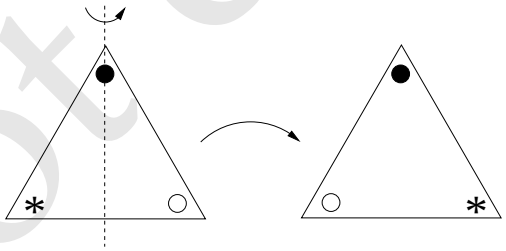
▷ **3.4.5:** List all the elements of the set of bijections of $A = \{a, b, c, d\}$ to itself, by indicating the image of each element. For example,

$$\begin{cases} a \mapsto a \\ b \mapsto c \\ c \mapsto b \\ d \mapsto d \end{cases}$$

denotes the bijection which swaps b and c.

- Find one element s that is not the identity and such that $s \circ s$ is the identity.
- Find one element t that is not the identity and such that $t \circ t \circ t$ is the identity.
- Find one element u that is not the identity, such that $u \circ u$ is not the identity, and such that $u \circ u \circ u \circ u$ is the identity.

▷ **3.4.6:** Make an equilateral triangle T out of cardboard (or imagine doing so!). Applying a (rigid) *symmetry* to T means taking your cardboard shape, moving it about in space as you wish, and then putting it back in its place, preserving its outline. Two symmetries are the same if the final positions of the triangle are the same. For example, leaving the triangle alone counts as one symmetry; flipping it about one of its axis counts as another one.



- Prove that there are exactly six distinct symmetries of an equilateral triangle.
- Find a one-to-one correspondence between the set of symmetries of an equilateral triangle and the set of bijections considered in Example 3.4.9.

In the second point, your one-to-one correspondence should preserve compositions, in the sense that performing one symmetry after the other should match the effect of composing the corresponding bijections.

3.4.7: You should emerge from Exercise 3.4.6 with a way to associate symmetries of a regular polygon with n sides with bijections from a set with n elements to itself. Which of the bijections you listed in Exercise 3.4.5 correspond to the symmetries of a square?

do not distribute

3.5 Canonical decomposition

3.5.1 Canonical decomposition

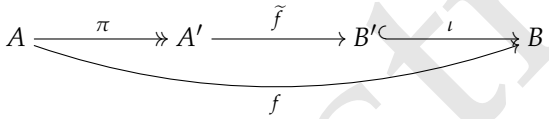
Now that we have acquired some familiarity with injective and surjective functions, we can appreciate a simple but important result. This is the prototype of a series of results known as 'first isomorphism theorems', which have manifestations in several contexts (groups, rings, modules—whatever these things may be). These more sophisticated forms are covered in courses in abstract algebra; what we are going to study here is the simpler set-theoretic counterpart.

Roughly, the result says that *every* function $f : A \to B$ may be written as the composition of a surjective function, a bijection, and an injective functions. Thus, injective and surjective functions are basic 'bricks' from which every other function can be constructed.

Here is a more precise version.

Theorem 3.5.1. *Given a function $f : A \to B$, there exist two sets A' and B', along with functions $\pi : A \to A'$, $\widetilde{f} : A' \to B'$, and $\iota : B' \to B$, such that:*

- $f = \iota \circ \tilde{f} \circ \pi$:



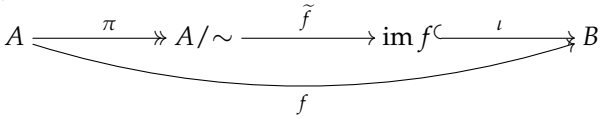
- *π is surjective, ι is injective, and $\widetilde{f}$ is an isomorphism (that is, it is bijective).*

This sounds impressive, but it is actually an extremely simple matter. The set B' will simply be the image of f in B, and the function ι will correspond (as in Example 3.3.5) to the inclusion $\operatorname{im} f \hookrightarrow B$; this is an injective function. As for A', it will be the quotient $A/\!\sim$ with respect to the equivalence relation $\sim$ on A defined by

$$a_1 \sim a_2 \quad \iff \quad f(a_1) = f(a_2) \quad .$$

You have already encountered this equivalence relation and the corresponding partition in your homework (Exercise 3.1.5), so you are hopefully already familiar with it. The function $\pi : A \twoheadrightarrow A/\!\sim$ will be the natural projection, as in Example 3.3.6; this is a surjective function.

How do I know that these are the 'right' ι and π? Because. You will know it too as soon as we prove that with these choices for ι and π we can indeed define an isomorphism $\widetilde{f} : A/\!\sim \to \operatorname{im} f$ so that $f = \iota \circ \widetilde{f} \circ \pi$:



In fact, we have reduced the proof of Theorem 3.5.1 precisely to the following statement:

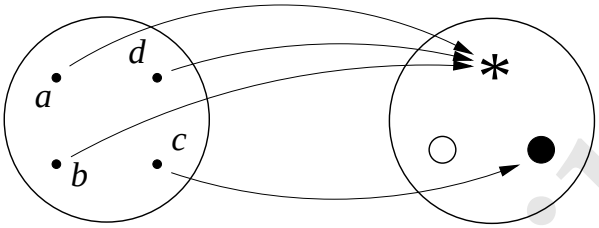
Lemma 3.5.2. *Given a function $f : A \to B$, and letting ι and π be as explained above, there exists a bijection $\widetilde{f} : A/\!\sim \to \operatorname{im} f$ such that $f = \iota \circ \widetilde{f} \circ \pi$.*

This is what we have to prove. To get a better feel for the statement, let's look at an example.

do not distribute

Example 3.5.3. Consider the (randomly chosen) function $\{a, b, c, d\} \to \{*, \circ, \bullet\}$ defined by

$$\begin{cases} a \mapsto * \\ b \mapsto * \\ c \mapsto \bullet \\ d \mapsto * \end{cases}$$



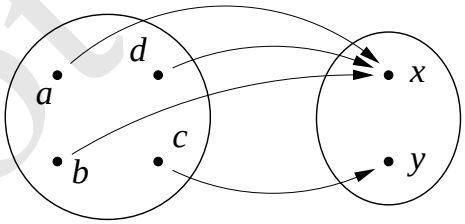
The equivalence relation determined by this function partitions $A = \{a, b, c, d\}$ into two equivalence classes:

$$A/\sim \; = \{\{a, b, d\}, \{c\}\} = \{[a]_\sim, [c]_\sim\} \quad .$$

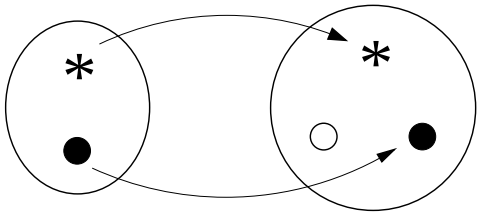
To fix ideas, call x the equivalence class $[a]_\sim = [b]_\sim = [d]_\sim$, and y the equivalence class $[c]_\sim$. The natural projection $A \twoheadrightarrow A/\sim$ (cf. Example 3.3.6) works like this:

$$\begin{cases} a \mapsto x \\ b \mapsto x \\ c \mapsto y \\ d \mapsto x \end{cases}$$

In pictures:

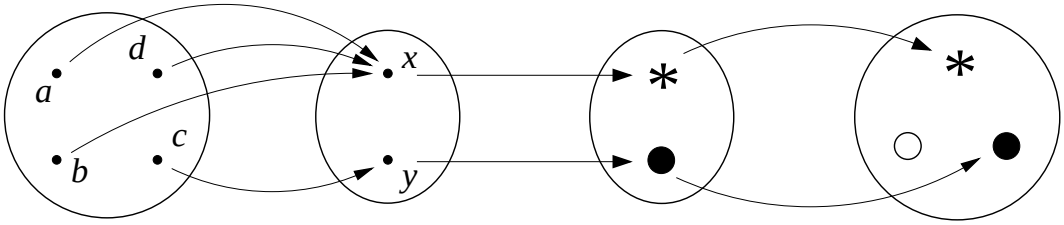


On the other hand, we have the inclusion of $\operatorname{im} f = \{*, \bullet\}$ into B, which we can view as an injective function (cf. Example 3.3.5) $\operatorname{im} f \hookrightarrow B$:



The point now is that we can visibly match the elements of $A/\sim$ and $\operatorname{im} f$:

do not distribute



in such a way that the effect of composing all these functions reproduces f: $a \in A$ goes to $x \in A/\sim$ to $* \in \operatorname{im} f$ to $* = f(a) \in B$, etc.

This is the diagram

$$A \xrightarrow{\ \pi\ } A/\sim \xrightarrow{\ \widetilde{f}\ } \operatorname{im} f \xhookrightarrow{\ \iota\ } B$$
$$A \xrightarrow[\ f\ ]{} B$$

in this case. ⌟

If this example is doing its job, you should readily believe that it generalizes to any function f from any set A to any set B. We always have a surjection

$$\pi : A \twoheadrightarrow A/\sim$$

from the source to the quotient modulo the equivalence relation mentioned above:

$$a_1 \sim a_2 \quad \Longleftrightarrow \quad f(a_1) = f(a_2) \quad ;$$

and we always have an injection

$$\iota : \operatorname{im} f \hookrightarrow B$$

simply realizing the inclusion $\operatorname{im} f \subseteq B$. What we need to verify is that f determines a function in the middle,

$$\widetilde{f} : A/\sim \longrightarrow \operatorname{im} f$$

and that this function is always a bijection. (We would say that f 'induces' $\widetilde{f}$.) If we do this (for an arbitrary $f : A \to B$) then we are done proving Theorem 3.5.1.

Now, there are two issues here:

- We have to define a function $\widetilde{f}$ such that $f = \iota \circ \widetilde{f} \circ \pi$;
- We have to prove that $\widetilde{f}$ is a bijection.

Until we have defined $\widetilde{f}$, we cannot prove anything about it. So, let's first focus on defining $\widetilde{f}$.

Look back at Example 3.5.3. The function $\widetilde{f}$ is the function in the middle, from $\{x, y\}$ to $\{*, \bullet\}$. How do we know that x should be sent to $*$, rather than $\bullet$? Because x is the image of (for example) b, and when the dust settles b has to reach $*$. Similarly, y must be sent to $\bullet$ because y comes from c, and $f(c) = \bullet$. Thinking about this should suggest a general prescription for where $\widetilde{f}$ should send an arbitrary element x of $A/\sim$:

$$\text{if } x = [a]_\sim, \text{ then define } \quad \widetilde{f}(x) = f(a) \quad . \tag{3.2}$$

If we can make this definition, then $\widetilde{f}$ will satisfy the requirement that $f = \iota \circ \widetilde{f} \circ \pi$. Indeed, we will have $\forall a \in A$

$$\iota \circ \widetilde{f} \circ \pi(a) = \iota(\widetilde{f}(\pi(a))) = \iota(\widetilde{f}([a]_\sim)) = \iota(f(a)) = f(a) \quad .$$

do not distribute

This is very nice, but your sharp eye notices that I have italicized the *If* in the previous sentence. There is one super-important subtlety that we have to address here, and that is the prototype of a very common situation.

Remember that in order to define a function we have to prescribe a unique element of the target for every element of the source. We have not quite done that yet, because our proposed prescription for $\widetilde{f}(x)$ depends on a *choice:* we have written: *"if $x = [a]_\sim$, then..."* and our recipe *appears to depend on the representative a.* What if $x = [a]_\sim = [b]_\sim$? Should we choose a, or should we choose b, when we define $\tilde{f}$ as in (3.2)? Will the result of our recipe depend on this? If it does, that's bad news: it would mean that our prescription potentially gives more than one value for $\widetilde{f}(x)$, and this is forbidden for functions.

Please make sure that you understand the point I am making in the previous paragraph, because it is very crucial. If you are defining a function by a prescription that depends on an arbitrary choice—here, it is the choice of a representative of a given equivalence class—then you have to make sure that the value you are proposing does *not* depend on the choice you have made. In math lingo, we say that you have to ensure that $\widetilde{f}$ is *well-defined:* that means that the definition of the induced function $\widetilde{f}$ appears to depend on a choice, but that you can show that this choice is in fact immaterial.

Now look back once more at Example 3.5.3 to appreciate what is going on. In deciding where to send x, you look up a representative for x—in this example, one of a, b, or d; and then see where f sends these representatives. The problem I am raising is the following: if it turned out that different representatives have different images in B, then we would be stuck: this definition would not work. But *it so happens that all the representatives are sent to $*$ by f,* and this is why we can safely send x to $*$.

We have to verify that this happens in complete generality, for an arbitrary function f from an arbitrary set A to an arbitrary set B.

Claim 3.5.4. *With notation as above, assume $[a_1]_\sim = [a_2]_\sim$. Then $f(a_1) = f(a_2)$. Therefore, $\widetilde{f}$ is well-defined.*

Proof. Once you really realize that this is what you have to prove, it proves itself. Indeed, it is a general fact about equivalence relations and equivalence classes that

$$[a_1]_\sim = [a_2]_\sim \quad \Longleftrightarrow \quad a_1 \sim a_2 \quad :$$

we proved this back in §2.3.2 (Lemma 2.3.7 to be precise). In the present context, this tells us that

$$[a_1]_\sim = [a_2]_\sim \quad \Longleftrightarrow \quad f(a_1) = f(a_2) \quad .$$

Our claim is the implication $\Longrightarrow$. □

This takes care of the first point mentioned above: now we *do* have an induced function $\widetilde{f} : A/\sim \to \operatorname{im} f$, (well-)defined by the prescription

$$(\forall a \in A): \quad \widetilde{f}([a]_\sim) = f(a) \quad ;$$

we have already observed that this implies $f = \iota \circ \widetilde{f} \circ \pi$ as needed.

In order to conclude the proof of Theorem 3.5.1, we have to prove that this function $\widetilde{f}$ is an isomorphism. Look again at Example 3.5.3: there is a perfect match between $A/\sim$ and $\operatorname{im} f$. This is what we have to prove in general.

Claim 3.5.5. *The induced function $\widetilde{f}$ is both injective and surjective.*

Proof. $\widetilde{f}$ is injective: Assume $\widetilde{f}([a_1]_\sim) = \widetilde{f}([a_2]_\sim)$. Then (by definition of $\widetilde{f}$) we have $f(a_1) = f(a_2)$; by definition of $\sim$, this means $a_1 \sim a_2$. By Lemma 2.3.7, it follows that $[a_1]_\sim = [a_2]_\sim$, and this proves that $\widetilde{f}$ is injective.

do not distribute

$\widetilde{f}$ is surjective: Let $b \in \operatorname{im} f$. This means that there exists some $a \in A$ such that $f(a) = b$. It follows (choosing $x = [a]_\sim$) that

$$\widetilde{f}(x) = f(a) = b \quad ,$$

proving that $\exists x \in A/\sim$ such that $\widetilde{f}(x) = b$. Thus $\widetilde{f}$ is surjective. □

This concludes the proof of Lemma 3.5.2, hence of Theorem 3.5.1!

3.5.2 *Applications*

Theorem 3.5.1 clarifies several issues we have encountered recently and not so recently.

Example 3.5.6. [*Inclusions and projections, again.*] Suppose $f : A \to B$ is a surjective function. Then $\operatorname{im} f = B$, hence the decomposition in Theorem 3.5.1 is simply

$$A \twoheadrightarrow A/\sim \xrightarrow{\sim} B \quad .$$

This shows that, up to the isomorphism $A/\sim \cong B$, the surjective function f may be identified with the natural projection $A \twoheadrightarrow A/\sim$. This is what I meant in Example 3.3.6 when I wrote that natural projections are 'typical of all surjective functions'.

The same point can be made concerning injective functions. If $f : A \to B$ is injective, then the equivalence relation $\sim$ is the equality relation: indeed, when we declare that

$$a_1 \sim a_2 \quad \Longleftrightarrow \quad f(a_1) = f(a_2) \quad ,$$

if f is injective, we are stating that $a_1 \sim a_2$ if and only if $a_1 = a_2$. Therefore, each equivalence class $[a]_\sim$ consists of the single element a. This says that (in this particular case!) the natural projection $A \to A/\sim$ is an isomorphism, and hence (in this particular case!) $A \cong \operatorname{im} f$. In this case, the canonical decomposition is

$$A \xrightarrow{\sim} A/\sim \xrightarrow{\sim} \operatorname{im} f \hookrightarrow B \quad .$$

In other words, up to the isomorphism $A \cong \operatorname{im} f$, the injective function f may be identified with the inclusion $\operatorname{im} f \hookrightarrow B$. This formalizes the (rather sloppy) claim made in Example 3.3.5 to the effect that 'every injective function looks like an inclusion'. ⌟

The most impressive use of Theorem 3.5.1 may be to the analysis of quotients. Suppose you are given a set A and an equivalence relation $\sim$ on A. You want to 'describe' the quotient $A/\sim$, that is (as we now know) find a set B which is isomorphic in some natural way to $A/\sim$. Here is what you do:

- Find a set B, and a function $f : A \to B$ such that the given equivalence relation $\sim$ matches the famous relation $\sim$: that is, such that $a_1 \sim a_2 \iff f(a_1) = f(a_2)$;
- Replace B with $\operatorname{im} f$ if necessary, to ensure that f is surjective;
- Then B is isomorphic to $A/\sim$, hence it describes the quotient.

This is just the first part of Example 3.5.6, recycled for the occasion. Indeed, f will be a surjective function $A \to B$, whose corresponding equivalence relation matches the given one $\sim$; we have seen in Example 3.5.6 that it follows that $(A/\sim) \cong B$.

I will restate this fact as a corollary (of Theorem 3.5.1), to make it easier to reference it.

do not distribute

Corollary 3.5.7. *Let A be a set, and let $\sim$ be an equivalence relation defined on A. Assume that $f : A \to B$ is a surjective function such that $\forall a_1, a_2 \in A$*

$$f(a_1) = f(a_2) \iff a_1 \sim a_2 \quad .$$

Then $A/\sim \cong B$.

Example 3.5.8. Our first basic example of quotient was for the relation defined on $\mathbb{Z}$ by

$$a \sim b \quad \iff \quad b - a \text{ is an even number}$$

(cf. Example 2.3.2). Let $B = \{0, 1\}$, and define $f : \mathbb{Z} \to \{0, 1\}$ by declaring that $f(n) = 0$ if n is even, and $f(n) = 1$ if n is odd. (That is: $f(n) =$ the remainder of n after division by 2.) Then

$$\begin{aligned} f(a) = f(b) \quad &\iff \quad a, b \text{ are either both even or both odd} \\ &\iff \quad b - a \text{ is an even number} \\ &\iff \quad a \sim b \quad . \end{aligned}$$

Thus the equivalence relation associated with f agrees with the given relation $\sim$. Since f is surjective, this is all we need to do to conclude that the quotient $\mathbb{Z}/\sim$ is isomorphic to a set with two elements. ⌟

Example 3.5.9. Look back at Examples 2.3.5 and 2.4.3: the set is $A = \mathbb{R}^2$, and the equivalence relation is

$$(x_1, y_1) \sim (x_2, y_2) \quad \iff \quad x_1^2 + y_1^2 = x_2^2 + y_2^2 \quad .$$

Define $f : \mathbb{R}^2 \to \mathbb{R}^{\geq 0}$ by

$$f(x, y) = x^2 + y^2 \quad .$$

It is then clear that $(x_1, y_1) \sim (x_2, y_2)$ if and only if $f(x_1, y_1) = f(x_2, y_2)$. It is also clear that f is surjective (that's why I restricted the target to $\mathbb{R}^{\geq 0}$). I don't need to do any more work or any more thinking to conclude that $(\mathbb{R}^2/\sim) \cong \mathbb{R}^{\geq 0}$, which was the result of the discussion in Example 2.4.3. ⌟

Example 3.5.10. Let's contemplate once more the equivalence relation defined in Example 2.3.3: A is $\mathbb{R}$, and $\sim$ is defined by

$$a \sim b \quad \iff \quad b - a \in \mathbb{Z} \quad .$$

Define a function $f : \mathbb{R} \to S^1$ by sending r to the point obtained by rotating $2\pi r$ radians counterclockwise (cf. Example 3.4.6; I could simply write $f(r) = e^{2\pi i r}$ if I were allowed to use complex functions). If r is larger than 1, we will have to go around the circle more than once before finding $f(r)$.

Now $f(a) = f(b)$ if and only if rotating by $2\pi a$ determines the same point as rotating by $2\pi b$, if and only if $2\pi b - 2\pi a$ is an integer multiple of 2π, if and only if $b - a$ is integer. That is,

$$f(a) = f(b) \quad \iff \quad b - a \in \mathbb{Z} \quad .$$

This is precisely the equivalence relation I was given to begin with, and f is surjective, so I can conclude right away that

$$\mathbb{R}/\sim \cong S^1$$

matching the conclusion of Example 2.4.4 with considerably less effort. (Note the last picture in Example 2.4.4, which is simply a picture of the function f defined here!) ⌟

do not distribute

Example 3.5.11. The same type of discussion would apply to any other example of quotient that we have run across, or indeed to *any* quotient. In Example 2.4.2 we considered $\mathbb{Z} \times \mathbb{Z}^{\neq 0}$, with the relation

$$(p_1, q_1) \sim (p_2, q_2) \quad \Longleftrightarrow \quad p_1 q_2 = p_2 q_1 \quad .$$

Just define a function $f : \mathbb{Z} \times \mathbb{Z}^{\neq 0} \to \mathbb{Q}$ by

$$f(p, q) = \frac{p}{q} \quad .$$

This function is surjective, and

$$f(p_1, q_1) = f(p_2, q_2) \quad \Longleftrightarrow \quad \frac{p_1}{q_1} = \frac{p_2}{q_2} \quad \Longleftrightarrow \quad p_1 q_2 = p_2 q_1 \quad ,$$

reproducing the given equivalence relation. Nothing else is needed in order to conclude that

$$(\mathbb{Z} \times \mathbb{Z}^{\neq 0})/\sim \cong \mathbb{Q} \quad :$$

this is now an immediate consequence of Theorem 3.5.1.

I am cheating a little here—in fact, $\mathbb{Q}$ is *defined* by the process illustrated in Example 2.4.2, and using $\mathbb{Q}$ (as I do in this example) to define $\mathbb{Q}$ would be circular reasoning. The point I am making here is that *if* we already have a satisfactory notion of $\mathbb{Q}$, then arguing as above shows that the quotient construction of Example 2.4.2 *must* reproduce it. So we can indeed take this construction as a definition of $\mathbb{Q}$ if we are so inclined. ⌟

3.5.3 *Homework*

3.5.1: Let $f : A \to B$, $\pi : A \to A/\sim$, $\iota : \operatorname{im} f \to B$, $\widetilde{f} : A/\sim \to \operatorname{im} f$ be as in this section. Since ι is injective, it has a left inverse κ; since π is surjective, it has a right inverse ρ. Find a way to express $\widetilde{f}$ in terms of f, κ, ρ.

▷ **3.5.2:** Apply 'canonical decomposition' to the function $f : \mathbb{R} \times \mathbb{R} \to \mathbb{R}$ defined by $f(x, y) = x^2 + y^2$. That is: define carefully suitable sets S and T and a surjective function $\mathbb{R} \times \mathbb{R} \to S$, a bijective function $S \to T$, and an injective function $T \to \mathbb{R}$, so that the composition $\mathbb{R} \times \mathbb{R} \to S \to T \to \mathbb{R}$ is the function given above.

3.5.3: Revisit Exercise 2.3.6 using what we learned in this section: prove that the quotient of $\mathbb{Z}$ by the relation defined in Exercise 2.3.5 consists of exactly n elements.

▷ **3.5.4:** Let M_3 be the set of 3×3 matrices with integer entries. Consider the relation $\sim$ on M_3 defined by declaring that, $\forall A_1, A_2 \in M_3$, $A_1 \sim A_2$ if and only if $\det(A_1) = \det(A_2)$. Describe the quotient $M_3/\sim$, by using the technique explained in §3.5.2 (cf. Corollary 3.5.7).

3.5.5: Let P be the set of polynomials with real coefficients, and define a relation $\sim$ on P by declaring that if $f(x)$ and $g(x)$ are polynomials, then $f(x) \sim g(x)$ if and only if $f(0) = g(0)$. This relation is clearly an equivalence relation. Describe the quotient $P/\sim$, by finding a familiar set that is isomorphic to it. (This will be very easy if you think in terms of a canonical decomposition.)

3.5.6: Let S be the set of differentiable functions $\mathbb{R} \to \mathbb{R}$. Let $\sim$ be the relation defined on S by declaring that, $\forall f_1, f_2 \in S$, $f_1 \sim f_2$ if and only if the graphs of $y = f_1(x)$ and $y = f_2(x)$ have parallel tangent lines at $x = 1$. Describe the quotient $S/\sim$.

▷ **3.5.7:** Let U be any set containing $\mathbb{Z}$, and let $\mathcal{F}$ be the set of *finite* subsets of U. Consider the equivalence relation $\cong$ defined on $\mathcal{F}$ by the relation of isomorphism. (You checked that $\cong$ is an equivalence relation, in Exercise 3.4.1.) Describe the quotient $\mathcal{F}/\cong$.

do not distribute

▷ **3.5.8:** Consider the relation $\sim$ on $\mathbb{N} \times \mathbb{N}$ defined by setting

$$(a_1, b_1) \sim (a_2, b_2) \iff a_1 + b_2 = a_2 + b_1 \quad .$$

Describe the quotient $\mathbb{N} \times \mathbb{N}/\sim$.

Remark: The previous several exercises would have likely seemed somewhat challenging if I had assigned them back at our first encounter with quotients. If you find them the least bit challenging now, then you have not completely absorbed the content of this section. Look again at Examples 3.5.9–3.5.11, and at the discussion preceding them. Once you get the right insight, all these exercises will look childishly easy (maybe disappointingly so).

do not distribute

do not distribute

4
Interlude: constructing $\mathbb{R}$

4.1 Incompleteness of $\mathbb{Q}$.

4.1.1 Real numbers: general considerations

Our next major task is a better understanding of the notion of 'infinity'. The set $\mathbb{R}$ of real numbers will show up in several key examples, so it is now important that we gain a better grasp of what this set really is. So far, I have used $\mathbb{R}$ rather liberally, trusting that your experience with calculus makes you comfortable with it. We gain an intuitive idea of what a real number is in grade school, in terms of *decimal expansions:* for example, we know that $\pi = 3.141592653\ldots$, where the $\ldots$ stand for an infinite sequence of digits that one can presumably figure out if one needs to. We take expressions such as $3.14159\ldots$ as 'numbers', and note that *rational* numbers can be written in this way, e.g., $\frac{2}{5} = 0.4$, $\frac{1}{3} = 0.333\ldots$, and so on.

This is a good concrete realization of $\mathbb{R}$, but as it stands it is not an actual 'definition', is it? If we view the expansion $3.14159\ldots$ as a shorthand for the 'infinite sum'

$$3 + \frac{1}{10} + \frac{4}{100} + \frac{1}{1000} + \frac{5}{10000} + \frac{9}{100000} + \cdots$$

then we seemingly open a can of worms: we have never defined an 'infinite sum'; as you know from calculus, this is a delicate business, and it is something you covered by glibly assuming that real numbers were there to begin with. From the point of view taken in this text, where we are supposed to *really* know what things mean, this is unsatisfactory. Granting that we know what the numbers $0, 1, 2, 3, \ldots$ are (but do we really[1]?), that is, assuming that we have the set $\mathbb{N}$, then we have learned how to construct $\mathbb{Z}$ (Exercise 3.5.8); and once we have $\mathbb{Z}$, we know how to construct $\mathbb{Q}$ (§2.4.2). But we have not seen yet a construction of $\mathbb{R}$. My goal in this chapter is to convince you that the material we have developed is indeed enough to construct a working isomorphic copy of $\mathbb{R}$, and in the process we will justify the usual expression of real numbers as decimal expansions. In fact, we will obtain *two* constructions of $\mathbb{R}$, based on somewhat different ideas: the construction presented in §4.2 will use what we have learned about *order* relations, while the construction in §4.3 will use *equivalence* relations.

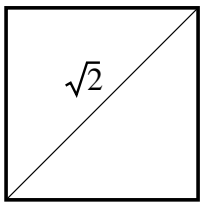
Actually, doing this in complete detail would take more room than I am willing to invest on this topic. The set of real numbers is officially required to satisfy a longish list of axioms[2]; I will just try to convince you that we can construct sets which satisfy several of these requirements, and that 'decimal expansions' are a good way to encode the information of elements of these sets. It would not be difficult to prove that the

[1] We are going to 'construct' $\mathbb{N}$ in our discussion of cardinality, in Chapter 5.
[2] You can see the whole list on the wikipedia page on 'Construction of the real numbers'.

do not distribute

constructions we will review satisfy all of the axioms; I will focus on one of the most striking requirements, which highlights the difference between $\mathbb{R}$ and the set $\mathbb{Q}$ of rational numbers.

In fact, before embarking on any new construction, we should pause a moment and ask whether we really need to do anything. Why don't we adopt $\mathbb{Q}$ as the set of 'all' numbers we care about? Answer: there are 'ordinary' numbers that are *not* rational numbers. For example, surely the length of the diagonal of a square of side 1 ought to be a number, right?



By Pythagoras' theorem, this 'ordinary' number is the square root of 2: $\sqrt{2}$.

4.1.2 $\sqrt{2}$ *is not rational*

I claim that the *square root of* 2 cannot be expressed as a rational number.

Proposition 4.1.1. *There are no integers a, b such that $(\frac{a}{b})^2 = 2$. That is, $\sqrt{2}$ is not a rational number.*

Proof. We argue by contradiction. Assume there exist integers a and b such that $(\frac{a}{b})^2 = 2$. After clearing common factors, we may assume that a and b are not both even. We will reach a contradiction by showing that the condition $(\frac{a}{b})^2 = 2$ implies that, on the contrary, both a and b are even.

If $(\frac{a}{b})^2 = 2$, then $a^2 = 2b^2$. This implies that a^2 is even, and it follows that a is itself even (you should have proven this in Exercise 1.3.7, by arguing that the square of an odd number is odd). Since p is even, there must be an integer c such that $a = 2c$. We then have

$$(2c)^2 = a^2 = 2b^2$$

therefore

$$4c^2 = 2b^2$$

and hence

$$b^2 = 2c^2 \quad .$$

But this implies that b^2 is even, and it would follow that b is also even. As promised, we have concluded that a and b are both even, a contradiction. □

We say that $\sqrt{2}$ is *irrational*. As it happens $\sqrt{2}$ is not all that special: it is easy to verify that many more numbers are irrational, such as $\sqrt{3}$, $\sqrt{5}$, or indeed $\sqrt[s]{p}$ for any prime number p and any integer $s > 1$ (Exercise 4.1.2). Further, important numbers such as π and e also turn out to be irrational. (You will verify that e is irrational in Exercise 4.4.3, which you could work out already if you are curious about this.) Later on (§5.2.4) we will understand that *most* real numbers are actually irrational, in a sense which we will make precise. Being rational is very special.

This is an important observation. When mathematics was in its infancy (it still is, but I am talking about thousands of years ago now) the naive perception of 'number' was that every number ought to be a quotient of integers, that is, rational. The Pythagoric school allegedly believed this very firmly. The story goes that

do not distribute

one of their own, Hippasus, suddenly realized during a boat trip that $\sqrt{2}$ was not rational, and the others promptly threw him overboard. Whether this is true or not, it is witness to the fact that the discovery that there are irrational numbers must have produced a paradigm shift in mathematics, whenever it happened.

You will certainly run again into Proposition 4.1.1 and generalizations of this statement. For example, in elementary analysis this fact is used to show that $\mathbb{Q}$ is not 'complete'; we will get a taste of that below. In algebra, it amounts to the statement that the polynomial $x^2 - 2$ is *irreducible* in the polynomial ring $\mathbb{Q}[x]$.

4.1.3 *Completeness*

Since we have not *yet* constructed the real numbers, in a sense we should not even be allowed to talk about $\sqrt{2}$: if $\mathbb{Q}$ is the only set of numbers we really know, then this number is missing (as we have verified), so even giving it a name should be questionable. Saying that $\mathbb{Q}$ is lacking something because 'it does not contain $\sqrt{2}$' presumes that we already know that this number exists in some form. (You may argue that it exists as the length of the diagonal of a square of side 1, but doing geometry in the plane $\mathbb{R}^2$ requires knowing already what $\mathbb{R}$ is!)

The formal way to express the fact that $\sqrt{2}$ is a 'hole' in $\mathbb{Q}$ without knowing anything (yet) about $\mathbb{R}$ is to observe that the set

$$\{x \in \mathbb{Q} \,|\, x^2 < 2\}$$

does not have a 'least upper bound' in $\mathbb{Q}$. The comments following Definition 2.2.8 may serve as a reminder of the terminology; but here is a stand-alone definition for this concept.

Definition 4.1.2. Let S be a set endowed with an ordering $\leq$, and let $A \subseteq S$ be a subset of S. An *upper bound* for A is any element $u \in S$ such that $a \leq u$ for all $a \in A$. A *least upper bound* $q \in S$ is an element such that

- q is an upper bound: for all $a \in A$, $a \leq q$;
- and q is a minimum element with this property: if q' is also an upper bound for A, then $q \leq q'$.

It is easy to see that the least upper bound for a set A, if it exists, is unique. (Exercise 4.1.4.) ⌟

Consider this definition for the set $S = \mathbb{Q}$ of rational number, with the usual ordering $\leq$.

Claim 4.1.3. *The set* $A = \{x \in \mathbb{Q} \,|\, x^2 < 2\}$ *does* not *have a least upper bound in* $\mathbb{Q}$.

Proof. The given set A certainly has upper bounds: any rational number $u > 0$ such that $u^2 \geq 2$ is an upper bound for A. Assume by contradiction that $q \in \mathbb{Q}$ is a least upper bound for this set. Then q is an upper bound, so $q^2 \geq 2$. Since q is rational, then $q^2 \neq 2$ (this is what we verified in §4.1.2); so $q^2 > 2$. Then $q^2 - 2 > 0$, hence (dividing by $2q$, which is certainly positive) $0 < \frac{q}{2} - \frac{1}{q}$. Let $t \in \mathbb{Q}$ be any rational number such that

$$0 < t < \frac{q}{2} - \frac{1}{q} \quad .$$

(There are arbitrarily small positive rational numbers, such as $\frac{1}{n}$ for large enough integers n, so this is not a problem.) Then $q - t \in \mathbb{Q}$, $q - t < q$, and yet

$$(q-t)^2 = q^2 - 2qt + t^2 > q^2 - 2qt > q^2 - 2q\left(\frac{q}{2} - \frac{1}{q}\right) = q^2 - q^2 + 2 = 2 \quad :$$

that is, $q - t$ is also an upper bound for A in $\mathbb{Q}$. This contradicts the assumption that q would be the *least* upper bound. □

do not distribute

Summarizing, we see that in $\mathbb{Q}$ there exist subsets which have upper bounds but have no *least* upper bound. Possibly the most important property of the set $\mathbb{R}$ of real numbers is that, on the contrary, *every* nonempty set of real numbers that has some upper bound *does have* a least upper bound in $\mathbb{R}$. You will see at your first encounter with analysis that all sort of wonderful facts follow from this one property. We say that $\mathbb{R}$ is 'complete', while $\mathbb{Q}$ is not 'complete'.

Definition 4.1.4. An ordered set S is *complete* if every nonempty subset $A \subseteq S$ that has upper bounds had *least* upper bounds. ⌟

Remark 4.1.5. We also say that a complete set 'satisfies the least upper bound property'. The reader should be aware that there other famous notions of 'completeness'; we will only be concerned with the one introduced in Definition 4.1.4. ⌟

Roughly speaking, a set S is a 'completion' of $\mathbb{Q}$ if it contains $\mathbb{Q}$, it is endowed with an ordering extending the ordering of $\mathbb{Q}$, and is complete; and further every element of S may be approximated to arbitrary precision by rational numbers. The set $\mathbb{R}$ of real numbers satisfies all these properties, but this is not the whole story: for example, these properties alone do not exclude the possibility that an element of s may be positive and yet less than every positive rational number. If this seems counterintuitive, it is because you have a good working knowledge of the real numbers, and no such silliness happens there. But if we are in the business of *constructing* the set of real numbers, we must make sure that we avoid such pathologies; so my actual definition of 'completion' (Definition 4.1.7 below) has to involve a few more ingredients. Here is one.

An order relation $\leq$ on a set S is a *total ordering* if $\forall a, b \in S$, either $a \leq b$ or $b \leq a$.

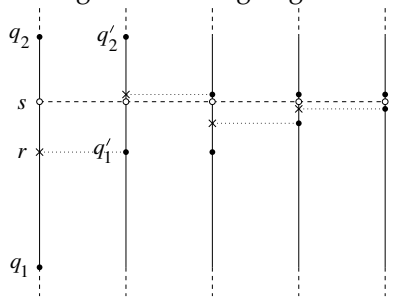
Lemma 4.1.6. *Let S be a set containing $\mathbb{Q}$ and endowed with a* total *ordering $\leq$ restricting to the usual ordering $\leq$ on $\mathbb{Q}$. Further, assume that for every element $s \in S$ there exist rational numbers $q_1, q_2 \in \mathbb{Q}$ such that $q_1 < s < q_2$.*

Then for every $\epsilon > 0$ we may choose these rational numbers $q_1 < q_2$ so that $q_2 - q_1 < \epsilon$.

Proof. The fact is true if $s \in \mathbb{Q}$, so we may assume that $s \notin \mathbb{Q}$. By assumption we may choose some $q_1, q_2 \in \mathbb{Q}$ such that $q_1 < s < q_2$, and we have to prove that for any given $\epsilon > 0$, we may choose these q_1, q_2 so that the difference $q_2 - q_1$ is less than ϵ. For this, it suffices to show that we can find $q_1', q_2' \in \mathbb{Q}$ such that $q_1' < s < q_2'$ and $q_2' - q_1' = \frac{1}{2}(q_2 - q_1)$: iterating this process, we can reduce the difference between the two rational numbers surrounding s to be below any given threshold.

Consider the 'average' $r := \frac{q_2 - q_1}{2}$ and note $q_1 < r < q_2$. Since the ordering on S is total and $s \notin \mathbb{Q}$, we have either $r < s$ or $r > s$. If $r < s$, let $q_1' = r$ and $q_2' = q_2$; if $r > s$, let $q_1' = q_1$ and $q_2' = r$. In either case we have $q_1' < s < q_2'$, and $q_2' - q_1' = q_2 - r = r - q_1 = \frac{1}{2}(q_2 - q_1)$ as needed, completing the proof. □

For what's worth, here is my mental image of what is going on in this proof:



At each step we halve the size of the interval surrounding s, by replacing one endpoint with the center of the previous interval. It is clear that we can make the interval as small as we wish by iterating this procedure.

do not distribute

Thus, under the hypotheses of Lemma 4.1.6, every element $s \in S$ can be squeezed between arbitrarily near rational numbers. This formalizes the requirement that one should be able to 'approximate to arbitrary precision' elements of a completion of $\mathbb{Q}$ by rational numbers. Borrowing a term from topology, we can say that $\mathbb{Q}$ is *dense* in S. With all this in mind, I will make the following (slightly nonconventional) definition.

Definition 4.1.7. A 'completion' of $\mathbb{Q}$ is a set $\mathbb{K}$ satisfying the following requirements.
—$\mathbb{K}$ contains a copy of $\mathbb{Q}$; so we can view $\mathbb{Q}$ as a subset of $\mathbb{K}$;
—$\mathbb{K}$ is endowed with a *total* ordering $\leq$ extending the usual order relation $\leq$ on $\mathbb{Q}$;
—For every $\alpha \in \mathbb{K}$, there are rational numbers q_1, q_2 such that $q_1 < \alpha < q_2$;
—For every $\alpha_1 < \alpha_2$ in $\mathbb{K}$ there is a rational number $q \in \mathbb{Q}$ such that $\alpha_1 < q < \alpha_2$;
—And $\mathbb{K}$ satisfies the least upper bound property with respect to $\leq$. ⌟

These are minimal requirements for a good set $\mathbb{R}$ of real numbers; but constructing a set $\mathbb{K}$ as in this definition, using the tools we have developed so far, seems highly non-obvious to me. In the next two subsections we will see two ways to do this. The resulting sets do in fact satisfy *all* official requirements of the set of real numbers, so they can be taken as concrete constructions of $\mathbb{R}$.

The constructions will be independent of each other, and they will produce isomorphic objects, so you can safely skip one of them if you wish. You can in fact skip both, if you are willing to trust me that completions exist; and in that case I'll see you again in §4.4. If you want a quick executive summary: In §4.2 we will take the standpoint that since $\mathbb{Q}$ fails the least upper bound property, we can complete $\mathbb{Q}$ by including least upper bounds for carefully chosen subsets of $\mathbb{Q}$. In §4.3 we will complete $\mathbb{Q}$ by assigning a limit to every sequence that ought to converge, such as the sequence determined by a decimal expansion. These are constructions with markedly different flavors; maybe surprisingly, they turn out to be equivalent.

4.1.4 *Homework*

4.1.1: You are probably familiar with the *rational root theorem*, which states that if $r = \frac{p}{q}$ is a rational root of a polynomial

$$P(x) = a_n x^n + a_{n-1} x^{n-1} + \cdots + a_1 x + a_0 \tag{4.1}$$

with integer coefficients, and p, q have no common factors, then necessarily p divides a_0 and q divides a_n. (The proof of this theorem is not difficult, but let's just take this statement for granted.)

Prove that if the polynomial with integer coefficients (4.1) is *monic*, i.e., $a_n = 1$, and $P(r) = 0$ for some $r \in \mathbb{Q}$, then in fact $r \in \mathbb{Z}$.

▷ **4.1.2:** Let $s > 1$ be an integer. Prove that if an integer n is not the s-th power of an integer, then $\sqrt[s]{n}$ is not a rational number. Deduce again that $\sqrt{2}$ is irrational, as well as e.g., the s-th roots $\sqrt[s]{p}$ of all prime integers p, for all $s > 1$. (Use Exercise 4.1.1.)

4.1.3: Let $A \subseteq \mathbb{Q}$ be the set of rational numbers of the form $1 - \frac{1}{n}$, as $n \in \mathbb{Z}^{>0}$. Does this set have a maximum? (Cf. Definition 2.2.8.) Does it have a least upper bound?

▷ **4.1.4:** Prove that if a set A has a least upper bound, then the least upper bound is unique.

4.1.5: Let $A \subseteq \mathbb{Z}$ be a nonempty set of integer. If A has an upper bound, does it necessarily have a *least* upper bound?

4.1.6: Let $\mathbb{K}$ be a completion of $\mathbb{Q}$ as defined in Definition 4.1.7, and let $\alpha \in \mathbb{K}$. Assume that $\alpha \geq 0$ and that $\alpha < q$ for every positive rational number $q \in \mathbb{Q}$. Prove that $\alpha = 0$.

do not distribute

4.2 Dedekind cuts

4.2.1 Dedekind cuts: definition

Our first construction will be based on 'Dedekind cuts'. I am going to give this definition precisely, and verify that the resulting set is a completion of $\mathbb{Q}$ in the sense explained in §4.1.3. It is in fact one of the standard models for the set $\mathbb{R}$ or real numbers, but I am not going to prove all that should be proved in order to establish this—for example, I will not study algebraic properties of reals in terms of Dedekind cuts. You should have no difficulties locating a more complete discussion if you so wish.

Definition 4.2.1. A *Dedekind cut* $(A|B)$ is a partition of $\mathbb{Q}$ into two nonempty subsets A, B such that $x < y$ for all $x \in A$ and all $y \in B$, and such that A has no maximum. I will denote by $\mathbb{D}$ the set of Dedekind cuts. ⌟

The notation $(A|B)$ is actually redundant: once we know A, B must be its complement in $\mathbb{Q}$ (since $\{A, B\}$ is a partition of $\mathbb{Q}$).

Chances are that you are not used to thinking of elements of $\mathbb{D}$ as 'numbers', and in particular of $\mathbb{D}$ as a useful replacement for $\mathbb{R}$, but hopefully this will seem reasonable at the end of the day. The very first remark that we need to make is that $\mathbb{Q}$ can be viewed as a subset of $\mathbb{D}$:

Proposition 4.2.2. *There is a natural injective function $\iota : \mathbb{Q} \to \mathbb{D}$, defined as follows: for $q \in \mathbb{Q}$,*

$$\iota(q) = (A|B) \quad \textit{where } A = \{x \in \mathbb{Q} \mid x < q\},\ B = \mathbb{Q} \smallsetminus A \quad .$$

Note that if $(A|B) = \iota(q)$, then q is in fact the minimum of the set B.

Proof. We have to verify that the pair $(A|B)$ specified in the statement is a Dedekind cut, so that ι does define a function $\mathbb{Q} \to \mathbb{D}$, and that this function ι is injective. In other words, we have to verify that if we set $A = \{x \in \mathbb{Q} \mid x < q\}$, then

- The complement B of A has the property that $x < y$ for all $x \in A$ and all $y \in B$;
- The set A has no maximum in $\mathbb{Q}$; and
- If $q_1 \neq q_2$, then the sets $\{x \in \mathbb{Q} \mid x < q_1\}$ and $\{x \in \mathbb{Q} \mid x < q_2\}$ are different.

The first two points verify that $(A|B)$ is indeed a Dedekind cut, and the third point verifies that this function is injective.

Each of these facts is very simple, but not 'obvious', and it is a good exercise to understand exactly why they are true. For example, the first one is true because of the transitivity of $\leq$ and the observation that, in the situation we are considering, the complement B of A is $\{y \in \mathbb{Q} \mid y \geq q\}$. Indeed, if $x < q$ (because $x \in A$) and $q \leq y$ (because $y \in B$), transitivity does imply that $x < y$.

To verify the second fact, note that if x and q are both rational, and $x < q$, then $\frac{x+q}{2}$ is still rational and $x < \frac{x+q}{2} < q$. This shows that x cannot be a maximum of A.

The third point is just as easy, and I will leave its proof to you (Exercise 4.2.1). □

We can use this injective function $\iota : \mathbb{Q} \hookrightarrow \mathbb{D}$ to identify $\mathbb{Q}$ with a subset of $\mathbb{D}$ (cf. Example 3.3.5). This will be done systematically in the considerations that follow. At the very least, this shows that we can indeed think about *some* Dedekind cuts as 'numbers'.

do not distribute

4.2.2 $\mathbb{Q} \subsetneq \mathbb{D}$

A simple but important remark is that $\mathbb{Q}$ *does not fill up* $\mathbb{D}$: not all Dedekind cuts are of the form $A = \{x \in \mathbb{Q} \,|\, x < q\}$, $B = \mathbb{Q} \smallsetminus A$. In order to verify this, it is helpful to observe the following.

Lemma 4.2.3. *Let $(A|B) = \iota(q)$ for a rational number q, as in Proposition 4.2.2. Then q is the least upper bound of A.*

Proof. Indeed, q is an upper bound for A since $q \in B$ and $x < y$ for all $x \in A$ and $y \in B$ (so $x < q$ for all $x \in A$). We have to verify that no rational number less than q is also an upper bound for A. Let then $q' \in \mathbb{Q}$, $q' < q$, and let $a = \frac{q'+q}{2}$. Then $a < q$, so $a \notin B$ (since q is the minimum in B), therefore $a \in A$. At the same time, $q' < a$: and this verifies that q' is not an upper bound for A, as needed. □

Example 4.2.4. Let

$$B = \{y \in \mathbb{Q}^{\geq 0} \,|\, y^2 \geq 2\}$$

and let A be the complement $\mathbb{Q} \smallsetminus B$. Thus,

$$A = \mathbb{Q}^{<0} \cup \{x \in \mathbb{Q} \,|\, x^2 < 2\} \quad .$$

Then $(A|B)$ is clearly a Dedekind cut. Arguing by contradiction, assume that the cut $(A|B)$ equals the image $\iota(q)$ of a rational number q. Then by Lemma 4.2.3 the rational number q would be a least upper bound for A. This would show that A has a least upper bound in $\mathbb{Q}$, contradicting Claim 4.1.3.

This contradiction proves that $(A|B)$ is *not* equal to $\iota(q)$ for any $q \in \mathbb{Q}$. ⌟

Morally, what we are doing is thinking of $\mathbb{Q}$ as usual, as the rational numbers on a line. A Dedekind cut is a way to split $\mathbb{Q}$ in two parts, one to the left of the other:

A cut belongs to $\mathbb{Q} \cong \operatorname{im} \iota$ when B has a minimum $q \in \mathbb{Q}$:

A cut does not belong to $\operatorname{im} \iota$ when B has no minimum in $\mathbb{Q}$, that is, where the minimum of B falls at a 'hole' in $\mathbb{Q}$ (such as $\sqrt{2}$). Here is the cut in Example 4.2.4:

Of course these comments (such as $\mathbb{Q}$ being part of a 'line', or some numbers such as $\sqrt{2}$ being 'holes' in $\mathbb{Q}$) are only after-the-fact interpretations. We do not *need* this extra information to define and manipulate Dedekind cuts, and that is the whole point: Dedekind cuts can be used to make rigorous sense of these notions.

4.2.3 *Order relation on* $\mathbb{D}$

The key property of $\mathbb{D}$ that I want to highlight is that it is complete with respect to a natural order relation extending the ordering $\leq$ on $\mathbb{Q}$. If you think in terms of the pictures I drew above, then you should find the following definition very reasonable.

Definition 4.2.5. We define a relation $\leq$ on $\mathbb{D}$ as follows: for two Dedekind cuts $(A_1|B_1)$ and $(A_2|B_2)$,

$$(A_1|B_1) \leq (A_2|B_2) \quad \iff \quad A_1 \subseteq A_2 \quad .$$

This is clearly an order relation. ⌟

do not distribute

Just to make sure that you agree with the last line in this definition, why don't you go ahead and prove that $\leq$ is indeed antisymmetric (Exercise 4.2.2). Also, you should check that if $q_1, q_2 \in \mathbb{Q}$ and $\iota(q_1) = (A_1|B_1)$, $\iota(q_2) = (A_2|B_2)$ are the corresponding Dedekind cuts, then

$$q_1 \leq q_2 \quad \Longleftrightarrow \quad (A_1|B_1) \leq (A_2|B_2) \quad .$$

Thus, the order relation $\leq$ on $\mathbb{D}$ extends the ordinary ordering $\leq$ on $\mathbb{Q}$, as expected.

Next, let's verify that $\leq$ is a *total* ordering in $\mathbb{D}$.

Proposition 4.2.6. *For any two Dedekind cuts $(A_1|B_1)$ and $(A_2|B_2)$, either $(A_1|B_1) \leq (A_2|B_2)$ or $(A_2|B_2) \leq (A_1|B_1)$.*

Proof. If $A_1 = A_2$, then in fact $(A_1|B_1) = (A_2|B_2)$; so we may assume $A_1 \neq A_2$. Thus either there is an $a \in A_2$ such that $a \notin A_1$, or there is an $a \in A_1$ such that $a \notin A_2$.

In the first case we have $a \in B_1$, and by definition of Dedekind cut we have that $x < a$ for all $x \in A_1$. Also by definition of Dedekind cut we have $a < y$ for all $y \in B_2$; hence $x < a < y$, and in particular $x \neq y$. So $x \notin B_2$, for all $x \in A_1$. Since A_2 is the complement of B_2, this implies that $x \in A_2$ for all $x \in A_1$. Therefore $A_1 \subseteq A_2$, and hence $(A_1|B_1) \leq (A_2|B_2)$ by Definition 4.2.5.

By the same argument, if there exists $a \in A_1$ such that $a \notin A_2$, then it follows that $(A_2|B_2) \leq (A_1|B_1)$. □

It is clear that every Dedekind cut $(A|B)$ is within an interval with rational endpoints: indeed, both A and B are nonempty by definition; and if $q_1 \in A$ and $q_2 \in B$, then (identifying $\mathbb{Q}$ with its copy in $\mathbb{D}$) we have $q_1 < (A|B) \leq q_2$ (and hence $q_1 < (A|B) < q_2 + 1$ if we insist on having $(A|B)$ in the interior of the interval).

It is also clear that in between any two cuts $(A_1|B_1) < (A_2|B_2)$ we can find a rational number. Prove this on your own! (Exercise 4.2.3.)

4.2.4 *Least upper bounds in* $\mathbb{D}$

At this point we have checked all properties of completions of $\mathbb{Q}$ listed in Definition 4.1.7, with the exception of the least upper bound property. Let's discuss upper bounds.

With the given definition of $\leq$, a Dedekind cut $(L|M)$ is an upper bound for a subset $S \subseteq \mathbb{D}$ if and only if

$$\forall (A|B) \in S : \quad A \subseteq L \quad .$$

Equivalently, this is the case if and only if $\forall (A|B) \in S$, $B \supseteq M$.

Of course not all subsets of $\mathbb{D}$ have upper bounds. For example, the 'copy' $\iota(\mathbb{Q})$ of $\mathbb{Q}$ in $\mathbb{D}$ does not have an upper bound: if $(A|B)$ were an upper bound for $\iota(\mathbb{Q})$, then we would have that $\forall q \in \mathbb{Q}$

$$\{x \in \mathbb{Q} | x < q\} \subseteq A \quad ;$$

but then A would contain the *union*

$$\cup_{q \in \mathbb{Q}} \{x \in \mathbb{Q} | x < q\} \quad ,$$

and this union is the whole of $\mathbb{Q}$ (check this!). So we would have $A = \mathbb{Q}$, $B = \mathbb{Q} \setminus B = \emptyset$, while for a Dedekind cut $(A|B)$ both A and B are supposed to be nonempty.

Finally, we are ready to verify that $\mathbb{D}$ satisfies the 'least upper bound property':

Theorem 4.2.7. *Let S be a nonempty subset of $\mathbb{D}$ such that S has an upper bound. Then S has a least upper bound in $\mathbb{D}$.*

do not distribute

Proof. Let $S = \{(A_i|B_i)\}_{i\in I}$, where I is any indexing set. Since S has an upper bound, say $(A|B)$, then we have that $A_i \subseteq A$ for all i and $B \subseteq B_i$ for all i. In particular, the intersection

$$M := \bigcap_{i\in I} B_i$$

is not empty. By the 'family' version of part (7) of Theorem 2.1.11 (see Exercise 2.1.7), the union of the sets A_i equals the complement of M:

$$L := \bigcup_{i\in I} A_i = \bigcup_{i\in I} (\mathbb{Q} \smallsetminus B_i) = \mathbb{Q} \smallsetminus \bigcap_{i\in I} B_i = \mathbb{Q} \smallsetminus M \quad .$$

This union is clearly nonempty (since it contains the A_i's). It is also clear that $\forall x \in L$, $\forall y \in M$, we have $x < y$. (Prove this! Exercise 4.2.4.) Finally, L has no maximum: if q were a maximum for L, then for some $j \in I$ we would have $q \in A_j$, and q would be a maximum for A_j, contradicting the fact that $(A_j|B_j)$ is a Dedekind cut.

Therefore $(L|M)$ is a Dedekind cut. I claim that it is the least upper bound of S.

$(L|M)$ is an upper bound for S: Let $(A_j|B_j) \in S$. Since L is the union of all the sets A_i, L contains all these sets, and in particular it contains A_j. By Definition 4.2.5, $(A_j|B_j) \leq (L|M)$.

$(L|M)$ precedes any other upper bound for S: If $(L'|M')$ is an upper bound for S, then L' contains all the sets A_i. It follows that L' contains their union, and this implies $L \subseteq L'$. By Definition 4.2.5, this means that $(L|M) \leq (L'|M')$, concluding the proof. □

Summarizing: $\mathbb{D}$ also satisfies the least upper bound property. We can use this property to define specific elements of $\mathbb{D}$. For example, if we care to *define* the (positive) square root of 2 in $\mathbb{D}$, we can do so by prescribing it to be the least upper bound of the set $\{q \in \mathbb{Q} \subseteq \mathbb{D} \,|\, q^2 < 2\}$: and then we will have 'filled the hole' in $\mathbb{Q}$ corresponding to the irrational number $\sqrt{2}$.

The set $\mathbb{R}$ of real numbers is characterized by several axioms, as mentioned in §4.1.1; we have verified that $\mathbb{D}$ satisfies a few of the most important ones. It does in fact satisfy all of them, although I do not want to take the room necessary to complete this discussion. (You will enjoy exploring some of the algebra of $\mathbb{D}$ in the homework.) The upshot is that Dedekind cuts do provide a simple framework for the rigorous definition of real numbers.

4.2.5 *Homework*

▷ **4.2.1:** Prove that the function $\mathbb{Q} \to \mathbb{D}$ defined in §4.2.1 is injective: that is, prove that if $q_1 \neq q_2$, then $\{x \in \mathbb{Q} \,|\, x < q_1\} \neq \{x \in \mathbb{Q} \,|\, x < q_2\}$.

▷ **4.2.2:** Prove that the relation $\leq$ on $\mathbb{D}$ defined in Definition 4.2.5 is antisymmetric.

▷ **4.2.3:** Let $(A_1|B_1) < (A_2|B_2)$ be two Dedekind cuts. Prove that there exists a rational number q such that $(A_1|B_1) < q < (A_2|B_2)$.

▷ **4.2.4:** With notation as in the proof of Theorem 4.2.7, prove that $x < y$ for all $x \in L$ and all $y \in M$.

▷ **4.2.5:** Provide a sensible definition of the 'sum' $(A_1|B_1) + (A_2|B_2)$ of two arbitrary Dedekind cuts $(A_1|B_1)$, $(A_2|B_2)$. This should be done in such a way that if $q_1, q_2 \in \mathbb{Q}$ and $\iota(q_1) = (A_1|B_1)$, $\iota(q_2) = (A_2|B_2)$, then $(A_1|B_1) + (A_2|B_2) = \iota(q_1 + q_2)$. (You don't need to prove in detail that your proposed definition is a Dedekind cut.)

4.2.6: Using the definition you provided in Exercise 4.2.5, prove that for every Dedekind cut $(A|B)$ there exists a Dedekind cut $(A'|B')$ such that $(A|B) + (A'|B') = 0$.

do not distribute

4.3 Cauchy sequences

4.3.1 Decimal expansions and Cauchy sequences

I am ready to bet that your intuitive idea of a 'real number' does not really match the technical definition of 'Dedekind cut'. As I mentioned in §4.1.1, we are used to thinking of real numbers in terms of their 'decimal expansion'. For the purpose of this discussion, a *decimal expansion*

$$b.a_1a_2a_3a_4\cdots \tag{4.2}$$

of a nonnegative real number is a sequence of rational numbers

$$\begin{aligned} q_0 = b \quad &= \quad b \\ q_1 = b.a_1 \quad &= \quad b + \frac{a_1}{10} \\ q_2 = b.a_1a_2 \quad &= \quad b + \frac{a_1}{10} + \frac{a_2}{10^2} \\ q_3 = b.a_1a_2a_3 \quad &= \quad b + \frac{a_1}{10} + \frac{a_2}{10^2} + \frac{a_3}{10^3} \\ &\cdots \end{aligned}$$

where $b \in \mathbb{Z}^{\geq 0}$, and the digits a_i are integers between 0 and 9. Expansions of negative real numbers may be defined analogously (and I'll leave the details of that to the reader). Of course there is nothing special about 10; we could consider 'binary' expansions, in which the digits a_i are chosen to be either 0 or 1.

These sequences are clearly 'special': they are increasing, and they don't go by leap and bounds; each new digit increments the number by a tiny bit, and this gets tinier as the expansion goes on. To formalize this observation, we may note that if $m \geq n \geq 0$, then $q_m - q_n < 1 = 10^0$; if $m \geq n \geq 1$, then $q_m - q_n < 0.1 = 10^{-1}$; if $m \geq n \geq 2$, then $q_m - q_n < 0.01 = 10^{-2}$; and so on. Otherwise put: For any N, $q_m - q_n < 10^N$ for all large enough m, n (for example $m \geq n \geq N$). This makes these sequences *Cauchy sequences.*

Definition 4.3.1. A sequence[3] $r_0, r_1, r_2, \ldots$ is a *Cauchy sequence* if

$$(\forall \epsilon > 0)\,(\exists N)\,(\forall m, n > N)\ :\quad |r_m - r_n| < \epsilon,$$

that is: for all $\epsilon > 0$, $|r_m - r_n| < \epsilon$ for all large enough integers m, n. ⌟

Example 4.3.2. Let $q \in Q$ be any rational number, and define a sequence by setting $r_n = q$ for all n; this could be called a 'constant' sequence. Then $\{r_n\}_{n\in N}$ is a Cauchy sequence. Indeed, for *all* m, n we have $r_m - r_n = q - q = 0$, so indeed $|r_m - r_n| < \epsilon$, for all $\epsilon > 0$.

In this (atypical) case, it does not matter how you choose N. ⌟

Example 4.3.3. As observed above, a decimal expansion is just a notation for a certain type of Cauchy sequence. With notation as above, given $\epsilon > 0$ we may find an integer N such that $10^{-N} < \epsilon$, and then we have noted that for all $m, n \geq N$, $|q_m - q_n| < 10^{-N} < \epsilon$.

In this (more typical) case, N has to be chosen carefully in order to satisfy the stated requirement for all $m, n > N$. ⌟

[3] These $r_0, r_1, \ldots$ can be, for example, rational numbers. The definition makes sense (and is useful) in every circumstance when we have a good notion of 'distance' between two elements. For $\mathbb{Q}$, I am adopting 'distance' to mean 'absolute value of the difference'.

do not distribute

Remark 4.3.4. Don't let the ϵ confuse or intimidate you. It's just the name of a quantity; $\forall \epsilon > 0$ tells that it can be arbitrarily large or small. If ϵ is smaller, presumably the corresponding N will have to be larger in order to satisfy the requirement in the definition; the definition requires that such an N may be found no matter how small ϵ is. (Look again at the case of decimal expansions, where this is very transparent.) Since ϵ is just a name, I could write 1000ϵ or $\epsilon/2$ in the definition without changing the requirement in the least. You have hopefully processed this fact already when you worked out Exercise 1.3.4. ⌟

We are going to construct $\mathbb{R}$ as the quotient of the set of *all* Cauchy sequences of rational numbers modulo a certain equivalence relation. I will use $\mathcal{C}$ to denote the set of Cauchy sequences of rational numbers.

Definition 4.3.5. Two Cauchy sequences $\{a_n\}_{n\in\mathbb{N}}$, $\{b_n\}_{n\in\mathbb{N}}$ are *equivalent*, and we write $\{a_n\}_{n\in\mathbb{N}} \sim \{b_n\}_{n\in\mathbb{N}}$, if

$$(\forall \epsilon > 0)\,(\exists N)\,(\forall n > N): \quad |a_n - b_n| < \epsilon \quad , \tag{4.3}$$

that is, if for all $\epsilon > 0$, $|a_n - b_n| < \epsilon$ for all large enough integers n. ⌟

We could say that $\{a_n\}_{n\in\mathbb{N}}$, $\{b_n\}_{n\in\mathbb{N}}$ are equivalent if $a_n - b_n$ 'converges to 0 as $n \to \infty$'. Thus, we are in a sense requiring the sequences a_n and b_n to 'have the same limit'. However, we are not (yet) in the position to think about it in these terms. A sequence such as the sequence encoded by the decimal expansion of $\sqrt{2}$:

$$1,\ 1.4,\ 1.41,\ 1.414,\ 1.4142,\ 1.41421,\ 1.414213,\ \ldots$$

does not have a limit in $\mathbb{Q}$—that is the moral of what we learned in §4.1.2.

The whole point of the construction is to create a number system in which this sequence, as well as every other Cauchy sequence, does have a limit.

Claim 4.3.6. *The relation defined in Definition 4.3.5 is an equivalence relation.*

Proof. Reflexivity and symmetry are completely straightforward and left to the reader. Transitivity requires a bit of work.

Assume $\{a_n\}_{n\in\mathbb{N}} \sim \{b_n\}_{n\in\mathbb{N}}$ and $\{b_n\}_{n\in\mathbb{N}} \sim \{c_n\}_{n\in\mathbb{N}}$. For a given ϵ, there exist N', N'' such that for all $n' > N'$ and all $n'' > N''$ we have $|a_{n'} - b_{n'}| < \frac{\epsilon}{2}$, $|b_{n''} - c_{n''}| < \frac{\epsilon}{2}$. (Here I just applied Definition 4.3.5. Look at Remark 4.3.4 if the switch from ϵ to $\epsilon/2$ seems confusing.) Let $N = \max(N', N'')$. Then for all $n > N$ we have $n > N'$ and $n > N''$, therefore

$$|a_n - c_n| = |a_n - b_n + b_n - c_n| \leq |a_n - b_n| + |b_n - c_n| < \frac{\epsilon}{2} + \frac{\epsilon}{2} = \epsilon \quad .$$

(Here I have used the 'triangle inequality', Exercise 4.3.1.)

Therefore $\{a_n\}_{n\in\mathbb{N}} \sim \{c_n\}_{n\in\mathbb{N}}$ according to Definition 4.3.5, and this proves transitivity. □

Now that we have an equivalence relation $\sim$ on our set $\mathcal{C}$, we can consider the quotient $\mathcal{C}/\!\sim$.

Definition 4.3.7. I will denote by $\mathbb{S}$ the quotient $\mathcal{C}/\!\sim$ of the set of Cauchy sequences of rational numbers by the equivalence relation defined in Definition 4.3.5. ⌟

The claim is that this new set $\mathbb{S}$ of classes of Cauchy sequences is a good model for the set $\mathbb{R}$ of real numbers. Again (as in §4.2) I will refrain from providing a complete proof of this claim, that is, a proof that $\mathbb{S}$ satisfies all the axioms required of $\mathbb{R}$. I will concentrate on showing that $\mathbb{S}$, like $\mathbb{D}$ in §4.2, is a 'completion' of $\mathbb{Q}$ in the sense of Definition 4.1.7.

do not distribute

4.3.2 $\mathbb{Q} \subsetneq \mathbb{S}$

In order to view $\mathbb{Q}$ as a subset of $\mathbb{S}$, we have to define an injective function $\iota : \mathbb{Q} \to \mathbb{S}$. This is actually very straightforward. If q is a rational number, the *constant* sequence defined by setting $a_n = q$ for all n is a Cauchy sequence (Example 4.3.2); we can let $\iota(q)$ be the equivalence class of this sequence:

$$\forall q \in \mathbb{Q} : \quad \iota(q) = [\{q, q, q, q, \dots\}]_\sim \quad .$$

Claim 4.3.8. *The function ι is injective.*

Proof. To verify that ι is an *injective* function, let q, q' be rational number and assume that $\iota(q) = \iota(q')$. Then the equivalence classes of the corresponding constant sequences $\{a_n\}_{n\in\mathbb{N}}$, $\{a'_n\}_{n\in\mathbb{N}}$ are equal, therefore (by Lemma 2.3.7!)

$$\{a_n\}_{n\in\mathbb{N}} \sim \{a'_n\}_{n\in\mathbb{N}} \quad .$$

By definition of $\sim$, this means that

$$(\forall \epsilon > 0)\,(\exists N)\,(\forall n > N) : \quad |a_n - a'_n| < \epsilon \quad .$$

In this case $a_n = q$ and $a'_n = q'$ for all n, so N and n are actually irrelevant: we obtain

$$\forall \epsilon > 0 : \quad |q - q'| < \epsilon \quad .$$

Since the only number that is less then any number > 0 is 0 itself, this says

$$|q - q'| = 0 \quad .$$

Thus $\iota(q) = \iota(q')$ implies $q = q'$; this proves the injectivity of ι. □

Remark 4.3.9. We could also associate with q the class of the Cauchy sequence determined by its decimal expansion: for example, $\frac{1}{7}$ would be sent to the class of the sequence

$$0,\ 0.1,\ 0.14,\ 0.142,\ 0.1428,\ 0.14285,\ 0.142857,\ \dots$$

This would agree with ι (do you see why? Exercise 4.3.3), but hopefully you agree with me that this is substantially more complicated than the definition in terms of constant sequences. ⌟

I will use ι to identify $\mathbb{Q}$ with a subset of $\mathbb{S}$, and therefore often omit ι from the notation.

The next natural observation is that this copy of $\mathbb{Q}$ does *not* fill up $\mathbb{S}$: $\mathbb{Q} \subsetneq \mathbb{S}$, as the header to this subsection announces. That is, there are Cauchy sequences of rational numbers which are *not* equivalent to $\iota(q)$ for any $q \in \mathbb{Q}$. This should strike you as intuitively clear at this point: If the decimal expansion of $\sqrt{2}$ were equivalent to a constant sequence q for $q \in \mathbb{Q}$, then the square of this particular q would be 2, contradicting Proposition 4.1.1. This is not a proof—to make it into a proof, I would need to discuss how to do algebra with Cauchy sequences, and I don't want to take the distracting tour that this would require. The fact that $\mathbb{Q} \cong \iota(\mathbb{Q})$ does not equal $\mathbb{S}$ will be (rigorously) clear once we establish that $\mathbb{S}$ is complete, since $\mathbb{Q}$ is not.

4.3.3 *Order relation on* $\mathbb{S}$

Next, we define an ordering $\leq$ on $\mathbb{S}$, extending the usual ordering $\leq$ on $\mathbb{Q}$.

Definition 4.3.10. Let $\alpha = [\{a_n\}_{n\in\mathbb{N}}]_\sim$ and $\beta = [\{b_n\}_{n\in Nbb}]_\sim$ be two elements of $\mathbb{S}$. We say that $\alpha \leq \beta$ if either $\alpha = \beta$ or $\exists N$ such that $a_n < b_n$ for all $n > N$. ⌟

do not distribute

That is: $\alpha \leq \beta$ if either $\alpha = \beta$ or $a_n < b_n$ for all large enough integers n.

Of course we have some work to do. We have to prove that this notion is well-defined (i.e., it does not depend on the Cauchy sequences chosen to represent α and β), and that it is indeed an order relation. The following lemma establishes the first point.

Lemma 4.3.11. *Let $\{a_n\}_{n\in\mathbb{N}}$, $\{a'_n\}_{n\in\mathbb{N}}$, $\{b_n\}_{n\in\mathbb{N}}$, $\{b'_n\}_{n\in\mathbb{N}}$ be Cauchy sequences. Assume that $\{a_n\}_{n\in\mathbb{N}} \not\sim \{b_n\}_{n\in\mathbb{N}}$, and that $\{a_n\}_{n\in\mathbb{N}} \sim \{a'_n\}_{n\in\mathbb{N}}$, $\{b_n\}_{n\in\mathbb{N}} \sim \{b'_n\}_{n\in\mathbb{N}}$.*

Then if $\exists N$ such that $a_n < b_n$ for all $n > N$, then $\exists N'$ such that $a'_n < b'_n$ for all $n > N'$.

The proof of this lemma is not difficult, but requires a steady hand; it is good practice for the type of manipulations needed in handling expressions involving several quantifiers.

Proof. Since $\{a_n\}_{n\in\mathbb{N}} \not\sim \{b_n\}_{n\in\mathbb{N}}$, there exists an $\epsilon > 0$ such that

$$(\forall M)\,(\exists m > M)\ :\quad |a_m - b_m| \geq \epsilon \quad . \tag{4.4}$$

(This conclusion is obtained by negating (4.3).)

Since $\{a_n\}_{n\in\mathbb{N}}$ and $\{b_n\}_{n\in\mathbb{N}}$ are Cauchy sequences, $\{a_n\}_{n\in\mathbb{N}} \sim \{a'_n\}_{n\in\mathbb{N}}$, and $\{b_n\}_{n\in\mathbb{N}} \sim \{b'_n\}_{n\in\mathbb{N}}$, we can find an N' such that $\forall m, n > N'$ we have

$$|a_m - a_n| < \frac{\epsilon}{4} \quad , \quad |b_m - b_n| < \frac{\epsilon}{4} \quad , \quad |a_n - a'_n| < \frac{\epsilon}{4} \quad , \quad |b_n - b'_n| < \frac{\epsilon}{4} \quad .$$

We may also assume that $N' \geq N$, with N given as in the statement; therefore, $a_n < b_n$ for all $n > N'$ by hypothesis.

Let $M = N'$ in (4.4) to obtain an $m > N'$ such that $|a_m - b_m| \geq \epsilon$. Since $m > N'$, we have $a_m < b_m$; so this says that $b_m \geq a_m + \epsilon$. I will record this inequality as

$$a_m + \frac{\epsilon}{2} \leq b_m - \frac{\epsilon}{2} \quad . \tag{4.5}$$

Using the triangle inequality (Exercise 4.3.1) we have

$$|a_m - a'_n| = |a_m - a_n + a_n - a'_n| \leq |a_m - a_n| + |a_n - a'_n| < \frac{\epsilon}{4} + \frac{\epsilon}{4} = \frac{\epsilon}{2} \quad ,$$

therefore $a'_n < a_m + \frac{\epsilon}{2}$; and

$$|b_m - b'_n| = |b_m - b_n + b_n - b'_n| \leq |b_m - b_n| + |b_n - b'_n| < \frac{\epsilon}{4} + \frac{\epsilon}{4} = \frac{\epsilon}{2} \quad ,$$

therefore $b'_n > b_m - \frac{\epsilon}{2}$. Putting these observations together with (4.5), we can conclude

$$(\forall n > N')\ :\quad a'_n < a_m + \frac{\epsilon}{2} \leq b_m - \frac{\epsilon}{2} < b'_n \quad ,$$

and this gives the statement. □

Verifying that the relation $\leq$ we have now securely defined on $\mathbb{S}$ is an ordering relation is actually rather straightforward.

- The definition implies in particular that for every α, $\alpha \leq \alpha$; so $\leq$ is reflexive.
- If $\alpha \leq \beta$ and $\beta \leq \gamma$, then clearly $\alpha \leq \gamma$ if $\alpha = \beta$ or $\beta = \gamma$. If $\alpha \neq \beta$ and $\beta \neq \gamma$, choose representatives $\{a_n\}_{n\in\mathbb{N}}$, $\{b_n\}_{n\in\mathbb{N}}$, $\{c_n\}_{n\in\mathbb{N}}$ for these three classes. Since $\alpha \leq \beta$ and $\beta \leq \gamma$, we can find an N such that $a_n < b_n$ and $b_n < c_n$ for all $n > N$. It follows then that $a_n < c_n$ for all $n > N$, and this implies that $\alpha \leq \gamma$. So $\leq$ is transitive.

do not distribute

- If $\alpha \leq \beta$ and $\beta \leq \alpha$ and $\alpha \neq \beta$, then we have a contradiction. Indeed, for representatives $\{a_n\}_{n\in\mathbb{N}}$, $\{b_n\}_{n\in\mathbb{N}}$ we would have that there is an N such that, for all $n > N$, $a_n < b_n$ and $b_n < a_n$. This contradiction shows that if $\alpha \leq \beta$ and $\beta \leq \alpha$, then necessarily $\alpha = \beta$. So $\leq$ is antisymmetric.

Now we have an ordering relation on $\mathbb{S}$. It does extend the usual ordering relation on $\mathbb{Q}$:

Claim 4.3.12. *Let $q, q' \in \mathbb{Q}$ and assume that $\iota(q) \leq \iota(q')$ in $\mathbb{S}$. Then $q \leq q'$ in $\mathbb{Q}$.*

Proof. If $\iota(q) = \iota(q')$, then $q = q'$ (since ι is injective) and therefore $q \leq q'$.

If $\iota(q) \leq \iota(q')$ and $\iota(q) \neq \iota(q')$, then we may choose the standard representative constant sequences $\{q, q, q, \dots\}$ and $\{q', q', q', \dots\}$ for $\iota(q)$ and $\iota(q')$, and Definition 4.3.10 implies directly that $q < q'$. So $q \leq q'$ in this case as well. □

We can also verify that $\leq$ is a *total* ordering on $\mathbb{S}$, as Definition 4.1.7 demands.

Proposition 4.3.13. *Let $\alpha, \beta \in \mathbb{S}$. Then either $\alpha \leq \beta$ or $\beta \leq \alpha$.*

Proof. If $\alpha = \beta$, there is nothing to prove. So let's assume that $\alpha \neq \beta$; we can then choose representatives $\{a_n\}_{n\in\mathbb{N}}$ and $\{b_n\}_{n\in\mathbb{N}}$ for α and β, respectively, such that $\{a_n\}_{n\in\mathbb{N}} \not\sim \{b_n\}_{n\in\mathbb{N}}$. This means that there exists an $\epsilon > 0$ for which

$$(\forall M)\,(\exists m > M) : |a_m - b_m| \geq \epsilon \quad . \tag{4.6}$$

Since $\{a_n\}_{n\in\mathbb{N}}$ and $\{b_n\}_{n\in\mathbb{N}}$ are Cauchy sequences, using this ϵ we can find an N such that $\forall m, n > N$ we have

$$|a_m - a_n| < \frac{\epsilon}{2} \quad \text{and} \quad |b_m - b_n| < \frac{\epsilon}{2} \quad .$$

Note that this implies

$$a_n < a_m + \frac{\epsilon}{2} \quad \text{and} \quad b_n > b_m - \frac{\epsilon}{2} \tag{4.7}$$

for all $m, n > N$. By setting $M = N$ in (4.6), we find a specific $m > N$ such that $|a_m - b_m| \geq \epsilon$. Without loss of generality we may assume that $a_m < b_m$; this will lead to the conclusion $\alpha \leq \beta$. The assumption $b_m < a_m$ would lead to the conclusion $\beta \leq \alpha$ by precisely the same argument.

Assuming that $a_m < b_m$, we have $b_m - a_m \geq \epsilon$ or, equivalently,

$$a_m + \frac{\epsilon}{2} \leq b_m - \frac{\epsilon}{2} \quad .$$

Putting this together with (4.7) we obtain that for all $n > N$

$$a_n < a_m + \frac{\epsilon}{2} \leq b_m - \frac{\epsilon}{2} < b_n \quad .$$

This shows that there exists an N such that $a_n < b_n$ for all $n > N$, establishing $\alpha \leq \beta$ as needed. □

Definition 4.1.7 also demands that we verify that for every $\alpha \in \mathbb{S}$ there exist rational numbers q_1, q_2 such that $q_1 < \alpha < q_2$, and that for all $\alpha_1, \alpha_2 \in \mathbb{S}$ there exists a rational $q \in \mathbb{Q}$ such that $\alpha_1 < q < \alpha_2$. Both claims are good exercises, so I'll leave them to you (Exercise 4.3.4 and 4.3.5).

do not distribute

4.3.4 *Least upper bounds in* $\mathbb{S}$

Summarizing, at this point we have defined $\mathbb{S}$ as a totally ordered set containing $\mathbb{Q}$ and verified most of the requirements in Definition 4.1.7. All that is missing in order to verify that $\mathbb{S}$ is a completion of $\mathbb{Q}$ is the least upper bound property. The proof is unfortunately rather technical, and I won't blame you if you simply want to trust me on this one. If you do go carefully through the argument, you will practice skills that should become second nature to you when you begin studying subjects such as real analysis.

Theorem 4.3.14. *Let D be a nonempty subset of $\mathbb{S}$ such that D has an upper bound. Then D has a least upper bound in $\mathbb{S}$.*

Proof. The main construction in this argument will be reminiscent of an analogous construction we used in the proof of Lemma 4.1.6.

Let $u \in \mathbb{S}$ be an upper bound for D. Every element of $\mathbb{S}$ is $\leq$ some rational number (Exercise 4.3.4), so we may assume $u \in \mathbb{Q}$. Since D is not empty, it contains some element $\delta \in \mathbb{S}$. Every element of $\mathbb{S}$ is $>$ some rational number (Exercise 4.3.4 again), so there exists a rational number $\ell \in \mathbb{Q}$ which is *not* an upper bound for D (since $\ell < \delta \in D$).

Now let $u_0 = u$, $\ell_0 = \ell$, and consider the following inductive procedure constructing sequences $\{\ell_n\}_{n\in\mathbb{N}}$, $\{u_n\}_{n\in\mathbb{N}}$ of rational numbers. Assuming that we have constructed ℓ_n and u_n, let $r \in \mathbb{Q}$ be the average $(\ell_n + u_n)/2$.

—If r is an upper bound for D, then let $\ell_{n+1} = \ell_n$ and $u_{n+1} = r$;

—If r is *not* an upper bound for D, then let $\ell_{n+1} = r$ and $u_{n+1} = u_n$.

Note that *every* u_n is an upper bound for D and *no* ℓ_n is an upper bound for D.

At each iteration the size of the interval $[\ell_n, u_n]$ is reduced in half. It follows that $\{\ell_n\}_{n\in\mathbb{N}}$, $\{u_n\}_{n\in\mathbb{N}}$ are *Cauchy sequences.* Indeed, let $\epsilon > 0$ be given; then there exists an N such that $u_N - \ell_N < \epsilon$; and for all $m, n > N$, the rationals ℓ_m, ℓ_n belong to the interval $[\ell_N, u_N]$ by construction, so $|\ell_m - \ell_n| < \epsilon$. By the same token, $|u_m - u_n| < \epsilon$ for $m, n > N$. In fact, again by the same reason, we have $|u_n - \ell_n| < \epsilon$ for $n > N$. This says that the Cauchy sequences $\{\ell_n\}_{n\in\mathbb{N}}$, $\{u_n\}_{n\in\mathbb{N}}$ are *equivalent* according to Definition 4.3.5.

Therefore these sequences determine the same element $\mu \in \mathbb{S}$. We will be done if we can show that μ *is the least upper bound of* D.

Claim 4.3.15. *With notation as above, μ is an upper bound for D.*

Proof. Let $\delta \in D$; we have to show that $\delta \leq \mu$. This is the case if $\delta = \mu$, so we may assume $\delta \neq \mu$.

Let $\{d_n\}_{n\in\mathbb{N}}$ be a Cauchy sequence representing δ; therefore $\{d_n\}_{n\in\mathbb{N}} \not\sim \{u_n\}_{n\in\mathbb{N}}$. By definition of equivalence (Definition 4.3.5), there exists an $\epsilon > 0$ such that

$$(\forall M)\,(\exists m > M)\,:|u_m - d_m| \geq \epsilon \quad , \tag{4.8}$$

that is, either $d_m \leq u_m - \epsilon$ or $d_m \geq u_m + \epsilon$ for such m.

Since both $\{d_n\}_{n\in\mathbb{N}}$ and $\{u_n\}_{n\in\mathbb{N}}$ are Cauchy sequences, we can find an N such that for $m, n > N$ we have $|d_m - d_n| < \frac{\epsilon}{2}$ and $|u_m - u_n| < \frac{\epsilon}{2}$, that is

$$d_m - \frac{\epsilon}{2} < d_n < d_m + \frac{\epsilon}{2} \quad \text{and} \quad u_m - \frac{\epsilon}{2} < u_n < u_m + \frac{\epsilon}{2} \quad . \tag{4.9}$$

Applying (4.8) to $M = N$, we obtain a specific $m > N$ for which $|u_m - d_m| \geq \epsilon$. First I note that necessarily $d_m \leq u_m - \epsilon$: indeed, otherwise we would have $d_m \geq u_m + \epsilon$, implying that for all $n > N$

$$d_n > d_m - \frac{\epsilon}{2} \geq u_m + \frac{\epsilon}{2} > u_m \quad ;$$

do not distribute

this would give $\delta > u_m$, contradicting the fact that u_m is an upper bound for D. Therefore $d_m \leq u_m - \epsilon$. Putting this together with (4.9), we see that

$$(\forall n > N) : \quad d_n < d_m + \frac{\epsilon}{2} \leq u_m - \frac{\epsilon}{2} < u_n \quad ,$$

and this yields $\delta \leq \mu$ as needed. □

Claim 4.3.16. *If $\alpha < \mu$, then α is not an upper bound for D.*

Proof. Let $\{a_n\}_{n\in\mathbb{N}}$ be a Cauchy sequence representing α. Since $\mu = [\{\ell_n\}_{n\in\mathbb{N}}]_\sim$, $\alpha < \mu$ implies that $\alpha \neq \mu$ and $a_n < \ell_n$ for all large enough n (Definition 4.3.10). Since $\alpha \neq \mu$, $\{a_n\}_{n\in\mathbb{N}}$ and $\{\ell_n\}_{n\in\mathbb{N}}$ are not equivalent. Therefore, there exists an $\epsilon > 0$ such that

$$(\forall M)\,(\exists m > M) : |\ell_m - a_m| \geq \epsilon \quad , \tag{4.10}$$

that is, either $a_m \leq \ell_m - \epsilon$ or $a_m \geq \ell_m + \epsilon$ for this specific index m. Since $a_n < \ell_n$ for all large enough n, we may in fact assume that for this m we have $a_m \leq \ell_m - \epsilon$.

Since $\{a_n\}_{n\in\mathbb{N}}$ is a Cauchy sequence, there is some N such that for all $m, n > N$ we have $|a_m - a_n| < \epsilon$. Applying (4.10) to $M = N$ gives a specific $m > N$ for which

$$a_m \leq \ell_m - \epsilon \quad \text{and} \quad (\forall n > N) : \quad |a_m - a_n| < \epsilon \quad .$$

This implies

$$(\forall n > N) : \quad a_n < a_m + \epsilon \leq \ell_m \quad .$$

Therefore $\alpha \leq \ell_m$ for this specific m. Since ℓ_m is not an upper bound (by construction), then α is not an upper bound, and this proves the claim. □

Claims 4.3.15 and 4.3.16 verify the conditions defining least upper bounds as in Definition 4.1.2, so we can conclude that μ is the least upper bound of D. □

Phew! That looks rather involved, certainly more than the corresponding fact for Dedekind cuts (Theorem 4.2.7). You may wonder why we should go through this extra work, considering that the discussion of Dedekind cuts is somewhat leaner and essentially accomplishes the same goal (of producing a concrete construction of a completion of $\mathbb{Q}$). The main advantage of Cauchy sequences is that they can be used in a more general setting: one can construct 'completions' (of a suitable kind, not necessarily according to Definition 4.1.7) of much more general 'metric spaces' by employing Cauchy sequences, with little more work than what we have done here. One instance is the construction of the 'field of p-adic numbers', which is a *different* completion of $\mathbb{Q}$, based on a different notion of what it means to be 'small'. There is such a notion for every prime p, and these different completions are very important in number theory.

But this is not our concern. At this point we have verified that the set $\mathbb{S}$ is a completion of $\mathbb{Q}$ in the sense specified in Definition 4.1.7, and that was my goal. Again, the set $\mathbb{R}$ is required to satisfy several more axioms, and one can verify that the set $\mathbb{S}$ we studied in this subsection does satisfy all of them. This set, like $\mathbb{D}$, may be taken as a rigorous construction of the set $\mathbb{R}$ of real numbers.

4.3.5 *Homework*

▷ **4.3.1:** Let x, y be rational (or even real) numbers. Prove that $|x + y| \leq |x| + |y|$. This is called the 'triangle inequality'.

do not distribute

4.3.2: If you know the technical definition of 'limit' of a sequence: Prove that every convergent sequence is a Cauchy sequence.

▷ **4.3.3:** For $q \in \mathbb{Q}$ define $\iota(q)$ as in §4.3.2, that is, as the equivalence class of the constant sequence $q, q, q, \ldots$. Also define $\iota'(q)$ to the the class of the Cauchy sequence corresponding to the decimal expansion of q. Prove that $\iota(q) = \iota'(q)$.

▷ **4.3.4:** Let $\alpha \in \mathbb{S}$. Prove that there exist rational numbers $q_1, q_2 \in \mathbb{Q}$ such that $q_1 < \alpha < q_2$.

▷ **4.3.5:** Let $\alpha, \beta \in \mathbb{S}$ and assume $\alpha < \beta$. Prove that there exists a rational number $q \in \mathbb{Q}$ such that $\alpha < q < \beta$.

4.3.6: For $\alpha, \beta \in \mathbb{S}$, we define $\alpha + \beta \in \mathbb{S}$ as follows. Choose representative Cauchy sequences $\{a_n\}_{n\in\mathbb{N}}$, $\{b_n\}_{n\in\mathbb{N}}$ for α, β, and let $\alpha + \beta$ be the class $[\{a_n + b_n\}_{n\in\mathbb{N}}]_\sim$.

- Prove that $\{a_n + b_n\}_{n\in\mathbb{N}}$ is a Cauchy sequence.
- Prove that the definition of $\alpha + \beta$ is independent of the chosen representatives.

A similar definition can be given for the product of two elements of $\mathbb{S}$. The usual properties of addition and multiplication hold for these operations (but don't feel compelled to work all of this out!).

4.4 *Decimal expansions, again*

4.4.1 *Back to decimal expansions*

Now that we know that completions of $\mathbb{Q}$ exist, we may try to understand elements of these completions more concretely. This is what decimal expansions do for us. As in §4.3.1, a decimal expansion

$$b.a_1a_2a_3a_4\cdots \tag{4.11}$$

of a nonnegative real number is shorthand for a sequence of rational numbers

$$\begin{aligned}
q_0 = b &= b \\
q_1 = b.a_1 &= b + \frac{a_1}{10} \\
q_2 = b.a_1a_2 &= b + \frac{a_1}{10} + \frac{a_2}{10^2} \\
q_3 = b.a_1a_2a_3 &= b + \frac{a_1}{10} + \frac{a_2}{10^2} + \frac{a_3}{10^3} \\
&\cdots
\end{aligned}$$

where $b \in \mathbb{Z}^{\geq 0}$, and the digits a_i are integers between 0 and 9. (As in §4.3.1, I will leave you the pleasure to modify the definitions to deal with *negative* numbers.) Of course we could equally well work with expansions over any base n, where $n > 1$ is an integer: then $b.a_1a_2a_3\cdots$ would stand for the sequence whose k-th term is

$$b + \frac{a_1}{n} + \frac{a_2}{n^2} + \cdots + \frac{a_k}{n^k}$$

and the 'digits' a_i are restricted to be between 0 and $n-1$. A popular choice is $n = 2$, giving 'binary expansions' such as $0.010101010\cdots$. (What is this binary number, in base 10?)

I will stick to *decimal* expansions (i.e., in base 10) out of habit, but everything that follows applies expansions in any base.

do not distribute

Let $\mathbb{K}$ be *any* completion of $\mathbb{Q}$, in the sense we agreed upon in Definition 4.1.7.

—*From decimal expansions to* $\mathbb{K}$*:* Given a decimal expansion $b.a_1a_2a_3a_4\cdots$ as above, consider the set of rational numbers obtained by taking more and more digits

$$\{b\,,\ b.a_1\,,\ b.a_1a_2\,,\ b.a_1a_2a_3\,,\ \dots\}\quad .$$

This may be viewed as a subset of $\mathbb{K}$, since $\mathbb{K}$ contains (an isomorphic copy of) $\mathbb{Q}$. This set has an upper bound, since all its elements are $< b+1$; and then it has a *least upper bound* in $\mathbb{K}$ since by assumption $\mathbb{K}$ satisfies the least upper bound property. The least upper bound is unique (Exercise 4.1.4), so this is a well-defined element of $\mathbb{K}$ determined by $b.a_1a_2a_3a_4\cdots$. ⌟

—*From* $\mathbb{K}$ *to decimal expansions.* Conversely, we can associate a decimal expansion with every element $\alpha\in\mathbb{K}$. Since $\mathbb{K}$ satisfies the requirements listed in Definition 4.1.7, $\alpha<q$ for some rational number $q\in\mathbb{Q}$. In particular, the set of integers $<\alpha$ is bounded above (by q), so it must contain a maximum integer b. For this integer b we have

$$b<\alpha\quad\text{and}\quad b+1\geq\alpha\quad .$$

To keep the notation simple I am going to assume that $b\in\mathbb{Z}^{\geq 0}$ (so that expansions will be as above; you can easily adapt the discussion to the negative case if you wish). Next, consider the numbers

$$b\quad,\quad b+\frac{1}{10}\quad,\quad b+\frac{2}{10}\quad,\dots,\quad b+\frac{8}{10}\quad,\quad b+\frac{9}{10}\quad,\quad b+1\quad .$$

The leftmost is $<\alpha$ and the rightmost is $\geq\alpha$, and $\mathbb{K}$ is totally ordered, so every one of these numbers can be compared to α. For one $a_1\in\{0,\dots,9\}$ we must have

$$b+\frac{a_1}{10}<\alpha\quad\text{and}\quad b+\frac{a_1+1}{10}\geq\alpha\quad .$$

This selects the first digit a_1 in the sought-for expansion. The other digits are obtained by the same token. Consider the numbers

$$b+\frac{a_1}{10}\quad,\quad b+\frac{a_1}{10}+\frac{1}{100}\quad,\quad b+\frac{a_1}{10}+\frac{2}{100}\quad,\dots,\quad b+\frac{a_1}{10}+\frac{9}{100}\quad,\quad b+\frac{a_1+1}{10}\quad .$$

For some uniquely determined $a_2\in\{0,\dots,9\}$ we must have

$$b+\frac{a_1}{10}+\frac{a_2}{100}<\alpha\quad\text{and}\quad b+\frac{a_1}{10}+\frac{a_2+1}{100}\geq\alpha\quad .$$

This selects the second digit a_2. And so on. ⌟

Thus, every decimal expansion determines an element of $\mathbb{K}$, and every element of $\mathbb{K}$ has a decimal expansion. Further:

Proposition 4.4.1. *Different elements of* $\mathbb{K}$ *have different decimal expansions.*

Proof. I will prove that if $\alpha\in\mathbb{K}$ has decimal expansion $b.a_1a_2\cdots$, then α equals the least upper bound in $\mathbb{K}$ of the sequence $\{b, b.a_1, b.a_1a_2, \dots\}$. This will show that elements of $\mathbb{K}$ are *uniquely* determined by their decimal expansions; equivalently, different elements have different decimal expansions.

To prove that α equals the least upper bound of $\{b, b.a_1, b.a_1a_2, \dots\}$, argue by contradiction: let β be the least upper bound of this sequence, and assume $\beta\neq\alpha$.

By construction, $b.a_1\cdots a_k<\alpha$ for all k; therefore $\beta\leq\alpha$, hence $\beta<\alpha$. By one of the properties listed in Definition 4.1.7, there exists a rational number q such that $\beta<q<\alpha$; and by the same property we can find a

do not distribute

second rational number q' such that $\beta < q' < q$. Let k be large enough such that $10^{-k} < q - q'$, that is, such that $q' + 10^{-k} < q$.

As $\beta < q' < q$, necessarily $b.a_1 \cdots a_k < q' < q$. In particular, $b.a_1 \cdots a_k + 10^{-k} < q' + 10^{-k} < q$.

On the other hand, by construction $\alpha \leq b.a_1 \cdots a_k + 10^{-k}$, the number obtained from $b.a_1 \cdots a_k$ by increasing the last digit by 1. Since $q < \alpha$, we obtain $q < b.a_1 \cdots a_k + 10^{-k}$.

The statements $b.a_1 \cdots a_k + 10^{-k} < q$ and $q < b.a_1 \cdots a_k + 10^{-k}$ are incompatible, so we have reached a contradiction. This contradiction follows from the assumption $\alpha \neq \beta$, so we have proved that $\alpha = \beta$ as needed. □

Do different decimal expansions determine different elements of $\mathbb{K}$? This is actually not the case: for example, the decimal expansions $0.99999\cdots$ and $1.00000\cdots$ represent *the same* rational number 1. (If you were not already familiar with this bizarre fact, it will become clearer once we emerge from the next subsection, cf. (4.12).) Therefore, these expansions determine the *same* element of any completion $\mathbb{K}$ of $\mathbb{Q}$.

As it happens, trailing $99999\cdots$ are the only exceptions.

Summarizing: Elements of any completion $\mathbb{K}$ satisfying the requirements of Definition 4.1.7 can be effectively described by decimal expansions. Thus, we are indeed allowed to think of real numbers in terms of decimal expansions. This is good news! My main goal in this chapter was to provide firm(er) foundations for this fact. And note again that working in base 10 is an arbitrary choice. We can equivalently handle real numbers by viewing them as binary expansions, or expansions in any other base.

The previous considerations apply to every completion of $\mathbb{Q}$, so they must apply in particular to the sets $\mathbb{D}$ and $\mathbb{S}$ constructed in §4.2 and §4.3. It is probably clear that $\mathbb{S}$ has to do with decimal expansions, since (as observed in §4.3.1) decimal expansions are our prototype of what a 'Cauchy sequence' should be. The connection with Dedekind cuts may seem more obscure, so we'll take a moment to look into it.

Extracting the decimal expansion of a Dedekind cut $(A|B)$ amounts to providing better and better approximations for the cut between A and B: at every stage in the process, the number $b.a_1 \cdots a_k$ we construct is in A, while the number obtained by increasing the last digit is in B.

Example 4.4.2. For instance, see what happens for the cut in Example 4.2.4, where

$$A = \mathbb{Q}^{<0} \cup \{x \in \mathbb{Q} \,|\, x^2 < 2\} \quad \text{and} \quad B = \{y \in \mathbb{Q}^{\geq 0} \,|\, y^2 \geq 2\} \quad .$$

The largest integer b in A is 1, since $1^2 = 1 < 2$ while $(1+1)^2 \geq 2$. So we have

$$1 \in A \quad \text{and} \quad 2 \in B \quad .$$

Next, among the numbers

$$1 \quad , \quad 1.1 \quad , \quad 1.2 \quad , \quad 1.3 \quad , \quad 1.4 \quad , \quad 1.5 \quad , \quad 1.6 \quad , \quad 1.7 \quad , \quad 1.8 \quad , \quad 1.9 \quad , \quad 2$$

we see that

$$1.4 \in A \quad \text{and} \quad 1.5 \in B \quad ,$$

since $(1.4)^2 = 1.96 < 2$ while $(1.5)^2 = 2.25 \geq 2$. Among the numbers

$$1.4 \quad , \quad 1.41 \quad , \quad 1.42 \quad , \ldots , \quad 1.48 \quad , \quad 1.49 \quad , \quad 1.5$$

we see that

$$1.41 \in A \quad \text{and} \quad 1.42 \in B \quad ,$$

since $(1.41)^2 = 1.9881 < 2$ while $(1.42)^2 = 2.0164 \geq 2$. And so on: it is hopefully clear that this process will generate all the digits of the decimal expansion of $\sqrt{2}$. ⌟

do not distribute

4.4.2 *Decimal expansions of rational/irrational numbers*

Assuming that we can manipulate decimal expansions (or expansions in any base) as the notation suggests, we can verify that rational numbers are characterized as follows:

Proposition 4.4.3. *Let*

$$b.a_1a_2a_3a_4\cdots$$

be a decimal expansion of a nonnegative real number r. Then r is rational if and only if the digits are eventually periodic: for some N, k

$$a_{n+k} = a_n$$

for all $n \geq N$.

The decimal expansion is said to have 'repeating decimals' if the condition expressed in this proposition holds. For example,

$$1.7465231231231231231231231\cdots$$

has repeating decimals 231. (What are N and k in this case, to use notation as in Proposition 4.4.3?)

Expansions of *negative* real numbers satisfy the same criterion for rationality; this only requires minor adjustments to the notation. And once again, note that working in base 10 is an arbitrary choice: the same statement holds in any base (for example for 'binary' expansions), with the same proof.

Proof. First assume that r is a rational number: hence, $r = \frac{p}{q}$ for two integers p and q. To obtain the decimal expansion of r, perform 'long division' of q into p. The process involves nonnegative remainders $< q$; since these are finitely many, at some stage in the long division a remainder has to repeat, and this will produce a periodic decimal expansion.

Conversely, assume that a decimal expansion is eventually periodic, as in the statement. After multiplying by a power of 10 and removing the integer part, I may assume that the number is

$$r = 0.a_1\cdots a_ka_1\cdots a_ka_1\cdots a_k\cdots$$

and it suffices to show that this number is rational. Observe that

$$10^k r - r = a_1\cdots a_k.a_1\cdots a_ka_1\cdots a_ka_1\cdots a_k - 0.a_1\cdots a_k = a_1\cdots a_k$$

is an integer; call this integer p (with decimal expansion $a_1\cdots a_k$). Then

$$(10^k - 1)r = 10^k r - r = p$$

from which

$$r = \frac{p}{10^k - 1} \quad ,$$

showing that this number is rational and concluding the proof. □

The notation in this proof is somewhat unpleasant, so you should work out simple examples to really see what is going on. For instance, follow the second part of the argument on the expansion

$$1.7465231231231231231231231\cdots$$

Multiplying by 10000 and subtracting 17465 leaves

$$r = 0.231231231231231231231\cdots$$

do not distribute

Working as in the proof,

$$1000r - r = 231.231231231231\cdots - .231231231231\cdots = 231 \quad ,$$

from which

$$r = \frac{231}{1000-1} = \frac{231}{999} \quad .$$

Reversing the first operations, the given number is shown to equal

$$\frac{1}{10000}\left(\frac{231}{999} + 17465\right) = \frac{2907961}{1665000}$$

after mindless arithmetic manipulations.

Applying the same procedure to r=$0.99999\cdots$ should clarify a claim made in passing in §4.4. In this case $k = 1$, and we have

$$(10-1)r = 9.999\cdots - 0.999\cdots = 9 \quad ,$$

i.e., $9r = 9$. It follows that

$$0.99999\cdots = r = \frac{9}{9} = 1 \quad . \tag{4.12}$$

And what about the *binary* number $0.01010101\cdots$? Here the base is 2, k is 2, and the repeating digits 01 determine the integer 1. Therefore

$$\text{(binary) } 0.01010101\cdots = \frac{1}{2^2-1} = \frac{1}{3} = \text{(decimal) } 0.33333\cdots \quad .$$

And indeed, multiplying the binary $0.01010101\cdots$ by the binary 11 (decimal 3) gives binary $0.111111\cdots$, which equals 1 by the same argument dealing with trailing $999\cdots$ in base 10.

A little practice with binary arithmetic may be useful. Again, there is nothing special about base 10; base 2 may be more natural in some contexts. As we have established, we can think of and manipulate real numbers as binary expansions. This point of view will be useful in a little while, for example when we first encounter 'uncountable' sets in §5.2.3 (see the comments following Corollary 5.2.7).

As a consequence of Proposition 4.4.3, the decimals in numbers such as $\sqrt{2}$, e, or π are guaranteed to *not* repeat (in any base). It is conjectured that *every* sequence of digits appears in the decimal expansion of π, but no one (to my knowledge) has a clue as to how one would go about proving something of this sort.

4.4.3 *Homework*

4.4.1: Apply to the number 1 the procedure explained in §4.3.1 to find the decimal expansion of an element of any completion $\mathbb{K}$ of $\mathbb{Q}$. Do you get $1.000\cdots$?

4.4.2: Put on your calculus hat if you wish, and consider the real numbers defined by the following series:

$$r_1 = \sum_{n\geq 1} \frac{1}{2^{2n}} \quad , \quad r_2 = \sum_{n\geq 1} \frac{1}{2^{n^2}} \quad .$$

Is r_1 a rational number? Why/why not? What about r_2? (Hint: Do not put on your calculus hat; rather, view these numbers as 'binary' expansions, i.e., work in base 2.)

do not distribute

▷ **4.4.3:** As you surely remember from calculus, the number e can be expressed as an infinite series:

$$e = \sum_{n=0}^{\infty} \frac{1}{n!} = 1 + \frac{1}{1!} + \frac{1}{2!} + \frac{1}{3!} + \frac{1}{4!} + \cdots \quad .$$

Is e a rational number?

- Assume e is rational: then $e = \frac{p}{q}$, where p, q are positive integers with no common factor. Show that if this is the case, then the number

$$r = q! \left(\frac{1}{(q+1)!} + \frac{1}{(q+2)!} + \frac{1}{(q+3)!} + \cdots \right) = \frac{1}{q+1} + \frac{1}{(q+1)(q+2)} + \frac{1}{(q+1)(q+2)(q+3)} + \cdots$$

  would necessarily be a positive integer.

- Show that $(q+1)r - 1 < r$, and deduce a contradiction.

The conclusion is that e is an *irrational* number.

This pretty argument is due to A. R. G. Macdivitt and Y. Yanagisawa. The first proof of this fact was obtained by Leonhard Euler in 1737, using 'continued fractions'.

I do not know of a similarly straightforward way to verify that π is irrational.

do not distribute

5
Cardinality

5.1 What is a 'number'?

5.1.1 The concept of 'number'

After this short excursion into the definition of $\mathbb{R}$ we go back to set theory, and try to approach an even more fundamental question. After all, we defined *real* numbers by clever constructions only involving *rational* numbers (the constructions presented in §4.2 and §4.3); we defined *rational* numbers by a clever construction only involving *integers* (this was one of our example of quotients by equivalence relations, Example 2.4.2); and we defined *integers* by a clever construction only involving *natural numbers* (Exercise 3.5.8). But how do we define natural numbers?

We have developed the main concepts of naive set theory without using numbers in any substantial way; numbers have only served as a good source of examples. For instance, the constructions of more and more refined types of numbers listed above above have conveniently illustrated set-theoretic operations such as quotients modulo equivalence relations. Assuming all along that we had a set of 'natural numbers' was handy, but not necessary. In fact, once we have developed a little set theory (as we have), we can take the standpoint of Exercise 3.5.7 and use set theory itself to *define* a notion of 'natural number': we can declare that the number 0 is the equivalence class of $\emptyset$ modulo isomorphisms of sets; the number 1 is the isomorphism class of $\{*\}$ (or of any other set with one element; for example, the class of $\{\emptyset\}$); the number 2 is the isomorphism class of a set such as $\{*, \circ\}$, and so on. To be able to do this I must be working in a family $\mathcal{F}$ of finite sets that contains at least one set with n elements for every number n; a standard way to do this is to require $\mathcal{F}$ to contain

$$\emptyset, \quad \{\emptyset\}, \quad \{\emptyset, \{\emptyset\}\}, \quad \{\emptyset, \{\emptyset\}, \{\{\emptyset, \{\emptyset\}\}\}\}, \quad \dots \tag{5.1}$$

(Every element in this list is the set of elements preceding it.) Once I have such a set $\mathcal{F}$, I can *define* $\mathbb{N}$ to be $\mathcal{F}/\cong$, associating with every finite set its equivalence class modulo the isomorphism relation. In fact, the sets listed in (5.1) may be taken as specific representatives for each equivalence class no matter how I choose $\mathcal{F}$; so I could just as well define $\mathbb{N}$ as the set of sets appearing in this list, and that would dispense me from having to worry about the family $\mathcal{F}$. Every finite set is isomorphic to exactly one set in this list, so this definition efficiently captures the essence of what isomorphic finite sets have in common. *That* is what a natural number ought to be.

But why stop there? The same approach allows us to define a notion of 'number' for *every* set.

Definition 5.1.1. Two sets determine the same *cardinal number* if they are isomorphic. The cardinal number

do not distribute

determined by a set S, or *cardinality of* S, is denoted $|S|$. ⌟

Since 'isomorphism of sets' (a.k.a. 'bijectivity') is an equivalence relation on any family of sets (Exercise 3.4.1) we can indeed define a cardinal number as an equivalence class modulo isomorphism, at least within a chosen family of sets. We also say that isomorphic sets 'have the same cardinality'.

You may wonder why I don't just define cardinality as an equivalence class with respect to the isomorphism relation on the 'set of all sets'. The problem is that there *isn't* a 'set of all sets'. I pointed this out already towards the end of §2.1.4. Later on in this chapter (§5.4.1) we will understand this fact (even) more thoroughly. The appropriate way to address this problem would again be to produce a specific representative for every possible isomorphism class, as (5.1) does for natural numbers. However, doing this would require discussing *ordinal numbers;* for the sake of brevity, I will just make do with the main message of Definition 5.1.1, that is, that cardinality is what isomorphic sets have in common.

If a set is finite, its cardinality is the same information as the number of elements that it contains. Indeed, two finite sets are isomorphic if and only if they have the same number of elements (Example 3.4.2, and cf. Definition 2.1.1).

Summarizing, numbers are not really 'primitive' entities: we have just seen how they may be defined in terms of more basic concepts (sets, equivalence relations…), and how this point of view leads us to appreciate a notion which generalizes the concept of number to possibly infinite sets. There is nothing particularly advanced or strange about the point of view I just described. The *names* of the numbers ('*one*', '*two*', and so on) must be words that people invented to refer to situations in which they had collections of objects that they could match one-by-one. The 'threeness' in a sack with three potatoes and in a bag with three coins boils down to the fact that if I put aside one coin and one potato at a time, then I will run out of coins precisely when I run out of potatoes. Defining a number as an isomorphism class of sets is nothing other than a formal way to do just this.

One advantage of doing this formally is that intuition may fail when one attempts to do the same with *infinite* sets: a naive matching protocol will not succeed in deciding whether two infinite sets have the same size, because it will never complete its task. That is probably the reason why we have just one word for 'infinite'! The approach through set theory bypasses these difficulties, in the sense that we can certainly consider and deal precisely with isomorphisms (i.e., bijections, 'matching') of infinite sets.

The results of this exploration may be somewhat counterintuitive at first; here are a few probing questions to contemplate before we start. Your intuition may or may not give you the 'right' answers, or any feeling of solid certainty about them. After we develop a bit of material, these questions will have a completely satisfactory, rigorous answer.

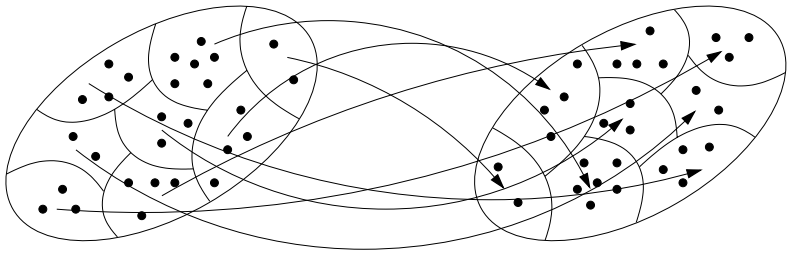
- Are there different types of infinite cardinal numbers, or just one?
- In what sense can we say that a cardinal number is 'smaller' than another?
- Once we agree that there are different orders of infinity, should $\mathbb{N}$ be smaller than $\mathbb{Z}$, or should they determine the same cardinal number? What about $\mathbb{N}$ and $\mathbb{Q}$? $\mathbb{N}$ and $\mathbb{R}$?

5.1.2 *A simple-minded lemma on bijections*

In a proof down the road I will need to invoke a very simple fact concerning bijections, which I could have listed as an exercise in Chapter 3. Informally, suppose you can break up two sets into parts that you can match, so that each part from one set is isomorphic to the corresponding part from the other. Then I want to conclude that the two sets are isomorphic to begin with. For example, the messy figure that follows depicts

do not distribute

two sets, each endowed with a partition splitting it into chunks:



These two sets have the same numbers of elements, because I can match the two partitions (as I have done with arrows) in such a way that matching subsets have the same numbers of elements.

The formal version of this statement tells us that this holds for all sets, in terms of cardinality:

Lemma 5.1.2. *Let A, B be sets, and let $\mathcal{P}_A$, $\mathcal{P}_B$ be partitions of A, B respectively. Let $\alpha : \mathcal{P}_A \to \mathcal{P}_B$ be a bijection such that*

$$(\forall S \in \mathcal{P}_A) : \quad S \text{ and } \alpha(S) \text{ have the same cardinality}$$

(that is, $S \cong \alpha(S)$). Then A and B have the same cardinality: $|A| = |B|$.

Don't let the notation distract you. A partition is just a set of sets, and $\alpha : \mathcal{P}_A \to \mathcal{P}_B$ just denotes a function from the set $\mathcal{P}_A$ to the set $\mathcal{P}_B$. If S is an element of $\mathcal{P}_A$, then $\alpha(S)$ is the corresponding element of $\mathcal{P}_B$. Thus, S is a certain subset of A and $\alpha(S)$ is a certain subset of B; it makes sense to compare these two sets in terms of the isomorphism relation, and the function α in the statement is assumed to have the property that $S \cong \alpha(S)$ for every S in the partition $\mathcal{P}_A$. The statement is that if there is an isomorphism α from $\mathcal{P}_A$ to $\mathcal{P}_B$ with this property, then $A \cong B$. Please look back at the picture and realize that it represents the situation considered in general in this lemma.

Proof of Lemma 5.1.2. By hypothesis, for each S in $\mathcal{P}_A$ there exists a bijection $f_S : S \to \alpha(S)$. Choose one such bijection f_S for each $S \in \mathcal{P}_A$, and denote by $g_S : \alpha(S) \to S$ the inverse of f_S. Define a function $f : A \to B$ as follows:

- Let $a \in A$; then a belongs to exactly one set $S \in \mathcal{P}_A$ (by definition of partition);
- Since α is a function $\mathcal{P}_A \to \mathcal{P}_B$, S determines a unique $\alpha(S) \in \mathcal{P}_B$, and we have a function $f_S : S \to \alpha(S)$;
- Define $f(a) \in B$ to be the element $f_S(a)$ of $\alpha(S) \subseteq B$.

Likewise, define a function $g : B \to A$ by essentially the same token:

- Let $b \in B$; then b belongs to exactly one set $T \in \mathcal{P}_B$ (by definition of partition);
- Since $\alpha : \mathcal{P}_A \to \mathcal{P}_B$ is a bijection, there exists a unique $S \in \mathcal{P}_A$ such that $T = \alpha(S)$, and we have a function $g_S : T = \alpha(S) \to S$;
- Define $g(b) \in A$ to be the element $g_S(b)$ of $S \subseteq A$.

It is easy to verify that $g \circ f = 1_A$ and $f \circ g = 1_B$ (essentially because for each $S \in \mathcal{P}_A$, $g_S \circ f_S = 1_S$ and $f_S \circ g_S = 1_{\alpha(S)}$). This shows that $f : A \to B$ is an isomorphism, and therefore $A \cong B$ as claimed. □

If you prefer, you can construct a proof of Lemma 5.1.2 by defining f as I did, and then proving explicitly that f is injective and surjective. Be my guest (Exercise 5.1.2).

do not distribute

5.1.3 *Another (not so simple-minded) lemma on bijections*

A second fact that we will need in the general discussion of cardinality will be a direct comparison between a set S and its power set $\mathcal{P}(S) = 2^S$ of subsets of S (cf. Definition 2.1.12).

If S is finite, then it is clear that 2^S is 'larger' than S: if S has n elements, then 2^S has 2^n elements (Exercise 2.1.15), and 2^n is larger than n.

But what if S is *infinite?* We may still retain the intuition that $|2^S|$ ought to be 'larger' than $|S|$, but we have don't yet have a way to compare cardinal numbers in this generality, and (as we will soon find out) intuition can only get us so far in such comparisons. We need something more precise. So here is the deceptively simple, but in reality not-so-simple lemma:

Lemma 5.1.3. *For every set S, S and the power set 2^S determine* different *cardinal numbers.*

How are we going to prove this? The fact that Lemma 5.1.3 is (hopefully!) so clear for finite sets is a psychological obstacle to understanding it well in the general case. Forget for a moment that we know from the start that 3 is strictly smaller than $2^3 = 8$. How would you prove that there are no bijections $S \to 2^S$, if $S = \{*, \circ, \bullet\}$? Think about this a moment.

Answer: Argue by contradiction. Imagine that I told you that I have found a function $f : S \to 2^S$ which is bijective. You can convince me that I must be wrong by *constructing an element in 2^S that is not in the image of f.* This would contradict the claim that f is bijective, since bijective functions are surjective. If you can do this for every possible function $f : S \to 2^S$, then you will have proven that there is no bijective function $f : S \to 2^S$.

Thus, the key to your argument will be some kind of procedure that starts from a given function $f : S \to 2^S$, and constructs an element of 2^S that is not in the image of f. The difficulty is that you don't know what f may be—your procedure must work no matter what f is. On the other hand, your procedure can (and in fact will have to) use f in some way.

Can we do this? Given f, we have to construct a subset of S that is not in the image of f. Well, here it is:

$$T := \{s \in S \,|\, s \notin f(s)\} \quad .$$

I claim that T does not equal $f(s)$, for any $s \in S$! I will illustrate this in a couple of examples, and then give a formal proof for the general case. This is a famous proof, called a 'diagonal argument', and going back to Georg Cantor, one of the founders of set theory.

Example 5.1.4. Let $S = \{*, \circ, \bullet\}$, and let $f : S \to 2^S$ be a function. We have to construct a subset of S that is not in the image of f. The claim is that

$$T := \{s \in S \,|\, s \notin f(s)\} \quad .$$

does not equal $f(s)$, for any $s \in S$. You could conceivably verify this claim by listing all functions $S \to 2^S$ and patiently checking that in each case the set T is not in the target of f. This is not really feasible, as it would involve hundreds of individual verifications even in this small case, but you can get a feel for what is going on by trying a few examples. Here is one:

$$\begin{cases} f(*) = \{*, \circ\} \\ f(\circ) = \emptyset \\ f(\bullet) = \{\bullet\} \end{cases}$$

do not distribute

What is the corresponding T? Since $*$ is in $f(*)$, $*$ won't be in T; since $\circ$ is *not* in $f(\circ) = \emptyset$, $\circ$ should be in T; and since $\bullet$ is in $f(\bullet)$, $\bullet$ won't be in T. Thus,

$$T = \{\circ\}$$

for this f. And look: $\{\circ\}$ is not listed on the right-hand side of the definition of f, as promised.

Another example? Here it is:

$$\begin{cases} f(*) = \{\circ\} \\ f(\circ) = \{\bullet\} \\ f(\bullet) = \{*, \circ, \bullet\} \end{cases}$$

For this f, $T = \{*, \circ\}$; this is not in the list on the right-hand-side.

Another one?

$$\begin{cases} f(*) = \{*, \circ\} \\ f(\circ) = \{\circ, \bullet\} \\ f(\bullet) = \{*, \bullet\} \end{cases}$$

This time $T = \emptyset$, since all of $*$, $\circ$, $\bullet$ happen to belong to the corresponding element of 2^S. Again, T is not in the list.

Please work this out for five more examples of your choice, to see that it always works. Then try to understand *why* it must always work. ⌟

Even in the ridiculously small case of Example 5.1.4 there are way too many functions $S \to 2^S$ ($8^3 = 512$ to be precise) to realistically list them all and verify that in each case the subset T you construct is not in the list you begin with. What we need is a general argument, proving the statement of Lemma 5.1.3 without reference to a specifically chosen S or f. Here is that general argument.

Proof of Lemma 5.1.3. We have to prove that there are no isomorphisms $S \to 2^S$. The general strategy is explained above: we are going to prove that no function $f : S \to 2^S$ is surjective, and in particular this shows that no such function can be a bijection. Given any function $f : S \to 2^S$, I define the subset

$$T := \{s \in S \,|\, s \notin f(s)\} \quad .$$

I claim that T is not in the image of f; that is, I claim that there is no $s \in S$ such that $f(s) = T$. To prove this claim, let $s \in S$. Then either $s \in T$, or $s \notin T$.

- If $s \in T$, then by definition of T we have that $s \notin f(s)$. Since $s \in T$ and $s \notin f(s)$, this shows $f(s) \neq T$ in this case.
- If $s \notin T$, then by definition of T we have that $s \in f(s)$. Since $s \notin T$ and $s \in f(s)$, this shows $f(s) \neq T$ in this case as well.

This shows that $f(s) \neq T$ for all $s \in S$, verifying my claim and proving that f is not surjective. This implies that f is not bijective, and concludes the proof of Lemma 5.1.3. □

Lemma 5.1.3 settles the first probing question I listed in §5.1.1:

Theorem 5.1.5. *There are different infinite cardinals.*

Proof. Indeed, $\mathbb{N}$ is an infinite set; so is $2^{\mathbb{N}}$; and by Lemma 5.1.3, $|\mathbb{N}| \neq |2^{\mathbb{N}}|$. □

do not distribute

In fact, it will turn out that

$$|\mathbb{N}| \quad , \quad \left|2^{\mathbb{N}}\right| \quad , \quad \left|2^{(2^{\mathbb{N}})}\right| \quad , \quad \left|2^{\left(2^{(2^{\mathbb{N}})}\right)}\right| \quad , \quad \cdots$$

are all different cardinal numbers. (This is not obvious, to me at least! It will follow from the 'Cantor-Bernstein-Schröder theorem'; see Corollary 5.4.2.) Once we have a more refined way to compare (infinite) cardinalities, we will in fact understand that this sequence is 'increasing'. This will be one good way to come to terms with the fact that there does not exist a 'set of all sets': such a set would have to have the 'largest' cardinality, but taking the power set always produces a 'larger' cardinality, so this would lead to a contradiction. What we are missing to draw such conclusions is an *order* relation for cardinalities. Defining such a relation will be our first order of business in the next section.

5.1.4 *Homework*

5.1.1: Given two finite sets A, B, construct another finite set M such that $|M| = |A| \cdot |B|$ by only invoking set-theory operations and without mentioning 'numbers' in any way. Then do the same for $+$: construct a set S such that $|S| = |A| + |B|$, 'without using numbers'.

Here, $|A|$, $|B|$, etc. mean the (conventional) numbers of elements in the set A, B, etc. Your construction will show how to extend operations such as $+$ and $\cdot$ to arbitrary cardinal numbers.

▷ **5.1.2:** Give an alternative proof of Lemma 5.1.2 by proving directly that the function $f : A \to B$ defined in the proof given in §5.1.2 is both injective and surjective (without using the inverse functions g_S).

5.1.3: Verify Lemma 5.1.3 for $S = \{*, \circ\}$ by working as in Example 5.1.4: list all 16 functions $S \to 2^S$, and verify that in each case the subset T corresponding to a function is not in the image of that function.

5.2 *Shades of infinity*

5.2.1 *Comparing cardinalities*

At this point we have verified that $|S|$ and $|2^S|$ are *different* cardinalitites (Lemma 5.1.3). I have informally claimed that in fact $|S|$ is 'smaller' than $|2^S|$. How should we compare cardinalities?

Definition 5.2.1. Let A and B be sets.

- We write $|A| \leq |B|$ if *there exists an injective function $A \to B$;*
- We write $|A| < |B|$ if $|A| \leq |B|$ and $|A| \neq |B|$, that is, *there is an injective function $A \to B$, but there does not exist a* bijective *function $A \to B$.* ⌟

Note that this system of notation is compatible with the situation for finite sets: if A and B are finite, and there is an injective function $A \to B$, then indeed the number of elements in A is less-than-or-equal-to the number of elements in B, that is, $|A| \leq |B|$ (Exercise 3.3.1); and if the sets are not isomorphic, then the inequality is strict.

So the proposed notation extends to arbitrary sets the familiar situation for finite set. But wait...I just gave a definition involving cardinalities, that is, equivalence classes modulo isomorphisms, by saying something about representatives. Don't we have to prove that this definition is independent of the choice of the sets A, B representing the cardinal numbers $|A|$, $|B|$? (We have run already into this type of problems: look at

do not distribute

the discussion in §3.5.1, p. 85.) Yes, we need to check this—more precisely, *you* need to check this: work out Exercise 5.2.1.

Now we have answered the second probing question listed at the end of §5.1.1: we have a precise sense in which we can state that a cardinal number is 'smaller' than another cardinal number.

With this in mind, we can capitalize on our work in §5.1.3 and state a more precise version of Lemma 5.1.3.

Theorem 5.2.2. *For all sets S, $|S| < |2^S|$.*

Proof. It is easy to see that $|S| \leq |2^S|$, that is, that there are injective functions $S \to 2^S$: for example, one such function is obtained by sending every $s \in S$ to the corresponding subset $\{s\} \in 2^S$. The fact that $|S| \neq |2^S|$ is subtler, but that's precisely what we proved in Lemma 5.1.3. □

Caveat. I have used the symbol $\leq$, suggesting that this comparison of cardinalities may be an *order* relation. It will indeed turn out to be an order relation (in the appropriate sense), but this is a comparatively difficult result: it will follow from the 'Cantor-Bernstein-Schröder theorem', which we will prove in a while. (Reflexivity and transitivity are very straightforward, and you should verify them yourself—Exercise 5.2.2.) For the time being, you can take $\leq$ among cardinals just as a convenient shorthand; but keep in mind that it will turn out to be an order relation.

5.2.2 *The 'diagonal' argument*

What is 'diagonal' about the proof of Lemma 5.1.3? Remember that a while back we had understood something interesting about the power set 2^S, for every set S:

Remark 5.2.3. For every set S, the power set of S is isomorphic to the set of functions $S \to \{0,1\}$. ⌟

Do you recognize this? It is a restatement of Exercise 3.1.4, where the two-element set was called $\{F,T\}$, and the statement was not using the language of isomorphisms since we had not yet developed that notion. Now we have, so we understand that Exercise 3.1.4 amounts to the statement that there is a bijection between the set of subsets of S and the set of functions from S to a set with 2 elements. *If* S is finite, this again implies that the number of elements in the power set 2^S is $2^{|S|}$.

In Example 5.1.4 and in the proof of Lemma 5.1.3 I chose to view 2^S as the set of subsets of S, but since it may be viewed as the set of functions $S \to \{0,1\}$, we can rethink through the argument from this different viewpoint.

The simplest way to associate a function $\varphi : S \to \{0,1\}$ to a subset A of S (and conversely) is to let $\varphi(s)$ be 1 if $s \in A$, and 0 if $s \notin A$. (Re-do Exercise 3.1.4 if you feel uncertain about this!) From this point of view, to prove Lemma 5.1.3 we would have to show that if you arbitrarily choose a function $\varphi_s : S \to \{0,1\}$ for every element $s \in S$, then you are necessarily leaving out some function. So the task would be to *construct* a function $S \to \{0,1\}$ that is necessarily missed by your choice—no matter how cleverly you have arranged φ_s to work.

Example 5.2.4. As usual, let's look at $S = \{*, \circ, \bullet\}$ to understand (even) better what is going on. Let's agree for convenience of notation that we list the elements of S in the order $*$, $\circ$, $\bullet$. Then we can encode a function $\varphi : S \to \{0,1\}$ very concisely, by just listing $\varphi(*)$, $\varphi(\circ)$, $\varphi(\bullet)$, in this order. With this convention, if I write

$$1 \quad 1 \quad 0$$

then I mean the function $\varphi : S \to \{0,1\}$ for which $\varphi(*) = 1$, $\varphi(\circ) = 1$, $\varphi(\bullet) = 0$. The subset corresponding to this function would then be

$$\{*, \circ\} \quad :$$

do not distribute

$*$ and $\circ$ are in because the corresponding values are 1, and $\bullet$ is out because the corresponding value is 0.

In Example 5.1.4 I was considering functions $f\colon S \to 2^S$. The first function mentioned in that example was

$$\begin{cases} * \mapsto \{*, \circ\} \\ \circ \mapsto \emptyset \\ \bullet \mapsto \bullet \end{cases}$$

Using the notation introduced above, I can now write precisely the same information as:

$$\begin{cases} * \mapsto 1 & 1 & 0 \\ \circ \mapsto 0 & 0 & 0 \\ \bullet \mapsto 0 & 0 & 1 \end{cases}$$

The other two functions shown in Example 5.1.4 were

$$\begin{cases} * \mapsto \{\circ\} \\ \circ \mapsto \{\bullet\} \\ \bullet \mapsto \{*, \circ, \bullet\} \end{cases} \quad , \quad \begin{cases} * \mapsto \{*, \circ\} \\ \circ \mapsto \{\circ, \bullet\} \\ \bullet \mapsto \{*, \bullet\} \end{cases}$$

and they would be written respectively

$$\begin{cases} * \mapsto 0 & 1 & 0 \\ \circ \mapsto 0 & 0 & 1 \\ \bullet \mapsto 1 & 1 & 1 \end{cases} \quad , \quad \begin{cases} * \mapsto 1 & 1 & 0 \\ \circ \mapsto 0 & 1 & 1 \\ \bullet \mapsto 1 & 0 & 1 \end{cases}$$

Now look at the *diagonal entries* (from top-left to bottom-right) in the arrays of 0's and 1's in these three examples. They are

$$1 \quad 0 \quad 1 \quad , \quad 0 \quad 0 \quad 1 \quad , \quad 1 \quad 1 \quad 1 \quad .$$

The information conveyed by the first, for instance, is that $*$ belongs to $f(*)$, $\circ$ does not belong to $f(\circ)$, and $\bullet$ belongs to $f(\bullet)$, if f is the first function considered above. The key point now is that this is *precisely* the prescription *opposite* to the one defining the subset T used in Example 5.1.4 (and in the proof of Lemma 5.1.3):

$$T := \{s \in S \,|\, s \notin f(s)\} \quad .$$

This says that the function $S \to \{0, 1\}$ corresponding to T is obtained by precisely switching 0 and 1 in the diagonal entries of the array. Doing this to the three diagonals extracted above gives

$$0 \quad 1 \quad 0 \quad , \quad 1 \quad 1 \quad 0 \quad , \quad 0 \quad 0 \quad 0$$

and now you should go and verify that the subsets corresponding to these functions are precisely the subsets we found in Example 5.1.4.

I could have written the proof of Lemma 5.1.3 using this 'diagonal' idea. Given any set of functions $\varphi_s : S \to \{0, 1\}$, one for each $s \in S$, define $\psi : S \to \{0, 1\}$ by 'switching 0 and 1 in the diagonal entries': that is,

$$\psi(s) = \begin{cases} 1 & \text{if } \varphi_s(s) = 0 \\ 0 & \text{if } \varphi_s(s) = 1 \end{cases}$$

I'll leave you the pleasure of verifying that ψ does *not* equal φ_s for any $s \in S$. (Exercise 5.2.3.) This will give you an alternative proof of Lemma 5.1.3.

do not distribute

5.2.3 *Countable and uncountable sets*

Theorem 5.1.5 (really a corollary of Lemma 5.1.3) already settles the almost philosophical issue of whether or not there is 'only one' infinite cardinality: this turns out *not* to be the case. We certainly[1] have infinite sets: the simplest one is arguably $\mathbb{N}$:

$$\mathbb{N} = \{0, 1, 2, 3, 4, \dots\}$$

This set appears very naturally in set theory, as you verified in Exercise 3.5.7: it is the set of cardinalities of finite sets.

You can convince yourself that $|\mathbb{N}|$ is the 'smallest' infinite cardinality, in the sense that if S is any infinite set, then $|\mathbb{N}| \leq |S|$, that is, there is an injective function $\mathbb{N} \to S$. Without being too rigorous about this, we can start with an infinite set S and remove one element from it per second: at second 0 we remove an element s_0, at second 1 we remove an element s_1, etc. If this process ever reaches the empty set, then S was finite to begin with. If the process *never* ends, then S is infinite, and the function $f : \mathbb{N} \to S$ defined by $f(n) = s_n$ is injective (because once we have thrown an element away, we do not have the option of discarding it again).

Since $\mathbb{N}$ is so special, we associate with its cardinality and the cardinality of its subsets a special name.

Definition 5.2.5. A set S is *countable* if $|S| \leq |\mathbb{N}|$, that is, if there exists an injective function $S \to \mathbb{N}$. It is *uncountable* if there is no such function. ⌟

Recall (Example 3.5.6) that, up to isomorphism, an injective function is just like the inclusion of a subset. So, a set is countable if and only if it may be identified with a subset of $\mathbb{N}$. The first part of the following result gives us an alternative way to view countable sets.

Proposition 5.2.6. *A nonempty set S is countable if and only if there is a* surjective *function $\mathbb{N} \twoheadrightarrow S$. An* infinite *set is countable if and only if it is isomorphic to $\mathbb{N}$.*

Proof. Assume S is nonempty. If $f : S \to \mathbb{N}$ is injective, then it has a left-inverse $g : \mathbb{N} \to S$ (Theorem 3.3.7); as $g \circ f = \mathrm{id}_S$, g has a right inverse, hence it is surjective. Thus, if S is countable, so that there is an injective function $S \to \mathbb{N}$, then there is a surjective function $\mathbb{N} \twoheadrightarrow S$.

Conversely, if $g : \mathbb{N} \twoheadrightarrow S$ is surjective, then it has a right-inverse $f : S \to \mathbb{N}$. As $g \circ f = \mathrm{id}_S$, f has a left inverse, hence it is injective. This shows that if there is a surjective function $\mathbb{N} \to S$, then there is an injective function $S \to \mathbb{N}$, and therefore S is countable.

These two observations prove the first assertion.

For the second assertion, assume that S is infinite and countable, so that there is an injective function $f : S \to \mathbb{N}$. Let $T \subseteq \mathbb{N}$ be the image of f; then S is isomorphic to T, so it suffices to show that $\mathbb{N} \cong T$. I am going to define an increasing sequence of elements of T, $t_0 < t_1 < t_2 < \cdots$ *inductively*, as follows:

—Since T is infinite, and in particular nonempty, then it has a minimum element t_0 ($\mathbb{N}$ is well-ordered!);

—Assume we have constructed $t_0 < \cdots < t_n$, all in T; since T is infinite, so is $T \smallsetminus \{t_0, \dots, t_n\}$, and in particular this set is not empty, and it has a minimum. I let t_{n+1} be this minimum[2].

[1] 'Certainly': actually, in more rigorous foundations of set theory the existence of an infinite set has to be required as an axiom.

[2] Here are the first few steps of this inductive procedure: Since T is nonempty, then the set T has a minimum element t_0; since T is infinite, $T - \{t_0\}$ is nonempty, so it has a minimum element t_1; and keep going:

—Since T is infinite, $T - \{t_0, t_1\} \neq \emptyset$, so it has a minimum element t_2;

—Since T is infinite, $T - \{t_0, t_1, t_2\} \neq \emptyset$, so it has a minimum element t_3;

—Since T is infinite, $T - \{t_0, t_1, t_2, t_3\} \neq \emptyset$, so it has a minimum element t_4;

and so on. One frequently uses induction to formalize the *'and so on'* in such recipes.

do not distribute

This process generates an increasing sequence of integers $t_0 < t_1 < t_2 < \cdots$, and every element of T eventually appears on this list: indeed, by construction $t_n \geq n$, so if $n \in T$ then n must have been reached by the process by the time we construct t_n. In other words, $T = \{t_0, t_1, t_2, \dots\}$. We can then define a function $f : \mathbb{N} \to T$ by setting $f(m) = t_m$ for all $m \in \mathbb{N}$. The function f is injective (because if $i < j$, then $t_i < t_j$) and surjective (because every element of T is t_m for some m), so it is an isomorphism. This shows $\mathbb{N} \cong T$, as needed. □

The name 'countable' is chosen well: S is countable if we can 'count' its elements, one by one, as in the non-rigorous procedure imagined at the beginning of this subsection, and detailed a little more precisely in the proof of Proposition 5.2.6. We will never exhaust an infinite countable set S by throwing away one element at a time, counting them off $\mathbb{N}$ as we do so; but if we could reach 'the end of time' (which of course we can't), a time at which we have accounted for all natural numbers, at that hypothetical time we would also have accounted for all elements of S. The intuition I have for a countable set consists of being able to 'count' *all* of its elements, one by one. Thus I like to include finite sets in Definition 5.2.5. You should be aware that some reserve the term 'countable' for infinite (countable) sets. Others use 'denumerable' for infinite countable sets.

What we have verified at this point (Theorem 5.1.5) is that *not all infinite sets are countable:* there are sets for which no matter how you decide to count elements out, then at the end of time, when you have accounted for all elements of $\mathbb{N}$, *there still are elements of S left.* If you find this counterintuitive, you are in excellent company—the realization that there are different 'levels' of infinity, essentially due to Cantor (around 1870) initially met with substantial resistance among some influential mathematicians such as Kronecker and Poincaré. Eminent philosophers such as Wittgenstein vehemently objected to this fact, even decades after Cantor's work. Still, we have proved everything that needs to be proved to draw this conclusion. Let me rephrase this particular case of Theorem 5.2.2, using the terminology we have introduced.

Corollary 5.2.7. *The set $2^{\mathbb{N}}$ is uncountable.*

This may not feel too satisfying, because after all you may never have thought about the set $2^{\mathbb{N}}$ before. But in fact I claim that you are familiar with an isomorphic copy of $2^{\mathbb{N}}$: *the set of real numbers $\mathbb{R}$ is isomorphic to $2^{\mathbb{N}}$.* You should try to think about this yourself, at least at some intuitive level (Exercise 5.2.5). Once you hit upon the right idea you should be able to convey it convincingly, although dotting all *i*s and crossing all *t*s may be a bit technical. (Hint: As we have established in §4.4, one can represent real numbers by binary expansions, i.e., 'decimal expansions in base 2'.) Once this is understood, Corollary 5.2.7 may be restated as follows.

Theorem 5.2.8. *The set $\mathbb{R}$ is uncountable.*

In fact it is easy to provide a direct proof of this statement that does not explicitly involve Lemma 5.1.3: we can simply adapt the diagonal argument to $\mathbb{R}$.

Alternative proof of Theorem 5.2.8. We verified a while back (Example 3.4.5) that $\mathbb{R}$ is isomorphic to the open interval $(0,1)$, so in order to prove Theorem 5.2.8 we can just as well prove that $(0,1)$ is uncountable. It is also notationally convenient to replace $\mathbb{N}$ by the set $\mathbb{N}^* = \mathbb{Z}^{>0} = \{1, 2, 3, \dots\}$ of nonzero natural numbers. This changes nothing, since clearly $\mathbb{N} \cong \mathbb{N}^*$ (the function $n \mapsto n + 1$ is bijective).

Thus, we just need to prove that $|\mathbb{N}^*| < |(0,1)|$. We will use the same strategy that was so successful earlier: we will show that every function $f : \mathbb{N}^* \to (0,1)$ misses some real number $r \in (0,1)$, by explicitly constructing such a number.

do not distribute

Let $f : \mathbb{N}^* \to (0,1)$ be any function. We can write the images $f(1)$, $f(2)$, etc. in terms of their decimal expansions:

$$\begin{aligned} f(1) &= 0.a_{11}a_{12}a_{13}a_{14}\cdots \\ f(2) &= 0.a_{21}a_{22}a_{23}a_{24}\cdots \\ f(3) &= 0.a_{31}a_{32}a_{33}a_{34}\cdots \\ f(4) &= 0.a_{41}a_{42}a_{43}a_{44}\cdots \\ &\cdots \end{aligned}$$

For example, if $f(1) = .6680339\cdots$, then $a_{11} = 6$, $a_{12} = 6$, $a_{13} = 8$, $a_{14} = 0$, and so on. Now take the 'diagonal' in this array:

$$a_{11}a_{22}a_{33}a_{44}\cdots$$

and *change every one of these digits.* It does not matter much how you do this: we can agree that if a digit is not equal to 1, then we will replace it by 1, and that if it is equal to 1 we will replace it by 2. Call $b_1b_2b_3b_4\cdots$ the digits obtained in this fashion, and let

$$r := 0.b_1b_2b_3b_4\cdots$$

be the corresponding real number. Then $r \in (0,1)$, and I claim that for all $n \in \mathbb{N}^*$, $f(n) \neq r$. This will conclude the proof, because it will show that f was not surjective to begin with, so it cannot be a bijection; as f was arbitrary, this shows that there are *no* bijections $\mathbb{N}^* \to (0,1)$.

Why is it that $f(n) \neq r$, no matter what n is? Because the n-th digit of $f(n)$ is a_{nn}, while the n-th digit of r is b_n, and we have arranged things so that $b_n \neq a_{nn}$. The decimal expansions of $f(n)$ and r differ, so $f(n) \neq r$. □

5.2.4 *Cardinality of $\mathbb{Z}$ and $\mathbb{Q}$*

You know several sets between the countable $\mathbb{N}$ and the uncountable $\mathbb{R}$: for example, the set $\mathbb{Q}$ of rational numbers. Is it countable or uncountable? What about the set $\mathbb{Z}$ of integers? And what about sets such as $\mathbb{Z} \times \mathbb{Z}$, or some kind of infinite product $\mathbb{Q} \times \mathbb{Q} \times \mathbb{Q} \times \cdots$? At the other end of the scale, we have seen that the open interval $(0,1)$ is uncountable. What about a *closed* interval $[0,1]$? would it be 'larger' than $(0,1)$, from this point of view?

It is actually very easy to analyze all these examples. One general theorem we are about to prove (the 'Cantor-Bernstein-Schröder theorem', Theorem 5.4.1) will make it even easier to compute the cardinality of most of them, but it is fun to work them out by hand first.

Example 5.2.9. Recall (Proposition 5.2.6) that infinite subsets of $\mathbb{N}$ are automatically isomorphic to $\mathbb{N}$, even if they may look 'smaller'. For instance, in the proof of Theorem 5.2.8 we used the fact that the set $\mathbb{N}^* = \mathbb{Z}^{>0}$ is isomorphic to $\mathbb{N}$: it is true that the inclusion function

$$\mathbb{N}^* \to \mathbb{N} \quad : \quad n \mapsto n$$

is injective but not surjective (it misses 0); *however,* there are bijective functions $\mathbb{N}^* \to \mathbb{N}$. For example, $n \mapsto n+1$ is such a function[3]. This is "Hilbert's Grand Hotel": if you have a hotel with infinitely many rooms

[3] Make sure this point sheds any residual mystery. Contrast this situation with the result of Theorem 5.2.8, which tells us that there are *no* bijective functions $\mathbb{N} \to \mathbb{R}$.

do not distribute

0, 1, 2, etc., your rooms are all full, and a new guest arrives, that's no big problem: just ask every guest to move from room n to room $n+1$, and that will free room 0 for the new guest.

Similarly, the set $2\mathbb{N}$ of *even* natural numbers is isomorphic to $\mathbb{N}$: the function $n \mapsto 2n$ is an isomorphism $\mathbb{N} \to 2\mathbb{N}$. By a very similar token, the set of odd natural numbers is also isomorphic to $\mathbb{N}$. If countably infinitely many new guests show up at your hotel, that's also not a big problem even if all rooms are taken already: just have every guest move from room n to room $2n$, and that will free up a countable infinity of rooms. (If uncountably many new guests show up, most of them will be out of luck.)

You can concoct many similar examples, even for sets that are not presented as subsets of $\mathbb{N}$. The set $\mathbb{Z}$ of integers is countable. This is clear, because we can 'list' all integers, for example as follows:

$$0, -1, 1, -2, 2, -3, 3, -4, 4, -5, 5, \ldots \tag{5.2}$$

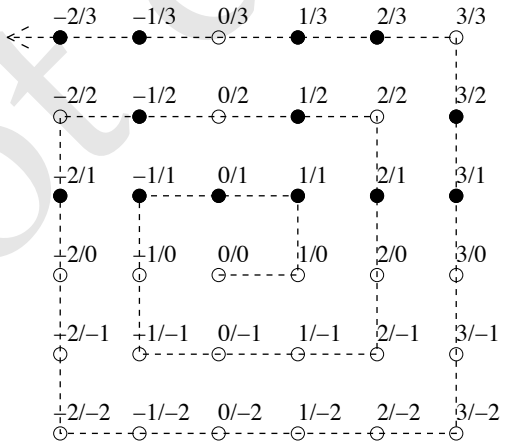
If you want a formula for the corresponding function $\mathbb{N} \to \mathbb{Z}$, that would be

$$n \mapsto f(n) = \begin{cases} \frac{n}{2} & \text{if } n \text{ is even} \\ -\frac{n+1}{2} & \text{if } n \text{ is odd} \end{cases} \tag{5.3}$$

and it is easy to show that this defines a bijection $\mathbb{N} \to \mathbb{Z}$. ⌟

In general, if you can list elements of an infinite set S—meaning that you can device a way to write them one after the other so that every element of the set is reached at some point—then the set is countable: this process is equivalent to defining a *surjective* function $\mathbb{N} \to S$, so S is countable by Proposition 5.2.6. It is not really necessary to come up with an explicit 'formula' for such a function. The display (5.2) already convinces me that $\mathbb{Z}$ is countable.

Example 5.2.10. The set $\mathbb{Q}$ of rational numbers is countable(!), even if(it looks almost as crowded as $\mathbb{R}$. Again, the point is that we can list all rational numbers 'one by one'. Indeed, we can list all fractions, for example as follows:



As I spiral out from $0/0$, I collect the a/b which correspond to fractions (so $b \neq 0$) and avoid repetitions. The black dots are the fractions that I pick up, in this order:

$$\frac{1}{1}, \frac{0}{1}, \frac{-1}{1}, \frac{2}{1}, \frac{1}{2}, -\frac{1}{2}, -\frac{2}{1}, \frac{3}{1}, \frac{3}{2}, \frac{2}{3}, \frac{1}{3}, \ldots$$

Of course there are many other ways to achieve the same effect. ⌟

do not distribute

With this, we have answered in full the third probing question from the list at the end of §5.1.1: $|\mathbb{N}|$ is 'strictly smaller' than $|\mathbb{R}|$ (Theorem 5.2.8); while $|\mathbb{N}| = |\mathbb{Z}| = |\mathbb{Q}|$, despite appearances. The set $\mathbb{Q}$ of rational number looks 'larger' (or at least 'thicker') than $\mathbb{Z}$ or $\mathbb{N}$ due to an elaborate optical illusion: it is actually just as large or small as $\mathbb{Z}$. We have a different perception of these countable sets exclusively because we are used to taking them *along with specific order relations*, and the natural orderings on these sets are very different. For example: $\mathbb{N}$ is well-ordered, while neither $\mathbb{Z}$ or $\mathbb{Q}$ are; and within any two elements of $\mathbb{N}$ or $\mathbb{Z}$ there are only finitely many other elements, while within any two rational numbers there are *infinitely many* rational numbers. The extra information due to the ordering is what makes the difference, not the abstract 'sets' themselves.

By contrast, there is an insurmountable gap between $\mathbb{Q}$ and $\mathbb{R}$: never mind that they 'look' the same; the first is countable, and the second is not. We now realize that there are more 'holes' in $\mathbb{Q}$, such as $\sqrt{2}$ (§4.1.2) or e (Exercise 4.4.3) than there are elements of $\mathbb{Q}$. The set $\mathbb{Q}$ looks less like a colander than like a fishing net made of very fine wire: most of it is holes.

Again, if you find this somewhat counterintuitive, you are in large company; many people find it unsettling at first, but they get used to it. After all, the math is laid out in front of you, and it is comparatively straightforward.

5.2.5 *Homework*

▷ **5.2.1:** Prove that the notation $|A| < |B|$ introduced in §5.2.1 is well-defined. (Note: This will be easy once you understand what it is asking for. Look back at §3.5.1 for a reminder on what 'well-defined' means.)

▷ **5.2.2:** Prove that the 'relation' $\leq$ satisfies reflexivity and transitivity.

▷ **5.2.3:** Complete the alternative proof of Lemma 5.1.3 sketched in §5.2.2. That is: with notation as in the end of §5.2.2, prove that ψ does *not* equal φ_s for any $s \in S$.

5.2.4: Let A, B be sets. Prove that $|A| \leq |B|$ if and only if there exists a *surjective* function $B \to A$. (Argue as in the beginning of the proof of Proposition 5.2.6.)

▷ **5.2.5:** Explain why $\mathbb{R} \cong 2^{\mathbb{N}}$. (You don't need to be too rigorous, but you should provide a convincing sketch of an argument.)

5.2.6: Verify explicitly that the function defined by (5.3) is a bijection.

5.2.7: We have seen that the set of *all* functions $\mathbb{N} \to \{0,1\}$ is uncountable. Now consider the set S of 'eventually periodic' functions $f : \mathbb{N} \to \{0,1\}$: that is, functions such that there exist N, k for which

$$f(n+k) = f(n)$$

for all $n \geq N$. Prove that S is countable.

5.2.8: We have seen that the set $2^{\mathbb{N}}$ of subsets of $\mathbb{N}$ is uncountable (Corollary 5.2.7). Since $\mathbb{Q} \cong \mathbb{N}$, it follows that the set $2^{\mathbb{Q}}$ of *all* subsets of $\mathbb{Q}$ is uncountable.

Let's throw away 'most' of these subsets: consider the set of subsets $A \subsetneq \mathbb{Q}$ with the property that $x < y$ for all $x \in A$ and all $y \notin A$. Is this set countable or uncountable?

do not distribute

5.3 *Je le vois, mais je ne le crois pas!*

5.3.1 *Cardinality and set-theoretic operations*

We already know that the cardinality of the power set of a set S is 'larger' than the cardinality of S. What about more simple-minded set-theory operations? Rather than attempting a thorough treatment of all possibilities, I will just review a few standard situations.

Example 5.3.1. The union $A \cup B$ of two countable sets is countable. Indeed, if A and B are countable, then we can list all elements of both sets:

$$A = \{a_0, a_1, a_2, a_3, \dots\}$$
$$B = \{b_0, b_1, b_2, b_3, \dots\}$$

To list all elements of $A \cup B$, I just need to zig-zag between the two lists:

$$a_0\,,\ b_0\,,\ a_1\,,\ b_1\,,\ a_2\,,\ b_2\,,\ \dots \quad .$$

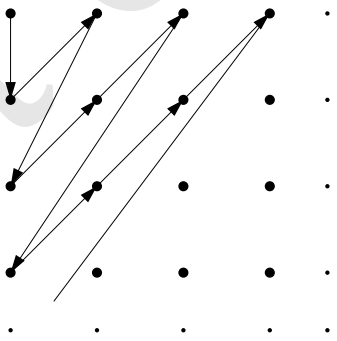
Of course such a list amounts to a surjection from a countable set to $A \cup B$, so $A \cup B$ is countable (Proposition 5.2.6).

It follows by a simple induction that the union of finitely many countable sets is countable. ⌟

Example 5.3.2. In fact, the same idea shows that a *countable union of countable sets* is countable: if

$$A_0 = \{a_{00}, a_{01}, a_{02}, a_{03}, \dots\}$$
$$A_1 = \{a_{10}, a_{11}, a_{12}, a_{13}, \dots\}$$
$$A_2 = \{a_{20}, a_{21}, a_{22}, a_{23}, \dots\}$$
$$\dots$$

then I just need to collect the elements along diagonals, as follows:



That is,

$$a_{00}\,,\ a_{10}\,,\ a_{01}\,,\ a_{20}\,,\ a_{11}\,,\ a_{02}\,,\ a_{30}\,,\ a_{21}\,,\ a_{12}\,,\ a_{03}\,,\ \dots \quad .$$

It is clear that every element will be reached by this procedure. ⌟

Example 5.3.3. The same token gives you that if A and B are countable, then so is $A \times B$: you can put elements of $A \times B$, that is, pairs (a_i, b_j), in a grid, just as above, and then zig-zag your way through it.

Incidentally, this gives another explanation of why $\mathbb{Q}$ is countable: recall from §2.4.2 that $\mathbb{Q}$ may be realized as a quotient of $\mathbb{Z} \times \mathbb{Z}^{\neq 0}$; thus, there is a surjective function $\mathbb{Z} \times \mathbb{Z}^{\neq 0} \twoheadrightarrow \mathbb{Q}$. Since $\mathbb{Z} \times \mathbb{Z}^{\neq 0}$ is infinite

do not distribute

countable (as product of two countable sets, as we have just seen) this gives a surjective function $\mathbb{N} \twoheadrightarrow \mathbb{Q}$. It follows that $\mathbb{Q}$ is countable, by Proposition 5.2.6.

Again, a simple induction shows that the product of any finite collection of countable sets is countable (Exercise 5.3.1). ⌟

5.3.2 *Algebraic and transcendental numbers*

Now that we have more tools at our disposal, we can give a more careful look at the 'gap' between the cardinalities of $\mathbb{Q}$ and $\mathbb{R}$.

Definition 5.3.4. A real number is *algebraic* if it is a root of a nonzero polynomial equation with integer coefficients. I will denote by $\mathbb{A}$ the set of algebraic real numbers. ⌟

Rational numbers are algebraic, since a/b is a solution of the equation $bx - a = 0$; thus $\mathbb{A}$ is sandwiched between $\mathbb{Q}$ and $\mathbb{R}$: $\mathbb{Q} \subseteq \mathbb{A} \subseteq \mathbb{R}$.

Note however that 'most' algebraic numbers are not rational, because the roots of a polynomial, even with integer coefficients, are not usually rational. For example, $\sqrt{2}$ is not rational, as we have verified (Example 4.1.2), but it is algebraic, because it is a root of the equation $x^2 - 2 = 0$.

Thus, $\mathbb{Q} \subsetneq \mathbb{A}$: there seems to be a gap between $\mathbb{Q}$ and $\mathbb{A}$. Is it as unsurmountable as the gap from $\mathbb{Q}$ to $\mathbb{R}$? Give yourself a moment to think about this before reading on.

Proposition 5.3.5. *The set $\mathbb{A}$ of algebraic real numbers is countable.*

Proof. The claim is that the set of all real number roots of all polynomials with integer coefficients is countable. Let R_d be the set of real roots of all polynomials with integer coefficients and of degree d; since a countable union of countable sets is countable (as we verified in Example 5.3.2), it is enough to show that each R_d is countable.

Construct a set R'_d of pairs $(p(x), r)$, where $p(x)$ is a polynomial of degree d with integer coefficients and r is a root of $p(x)$. We have a surjective function $R'_d \to R_d$, defined by sending $(p(x), r)$ to r. By Proposition 5.2.6, it suffices to show that R'_d is countable: indeed, it will then follow that R_d is countable.

To deal with R'_d, consider the set P_d of nonzero polynomials of degree d with integer coefficients. I claim that in order to prove that R'_d is countable, it suffices to show that P_d is countable. Indeed, we have a function $f : R'_d \to P_d$, defined by sending $(p(x), r)$ to $p(x)$; so we may view R'_d as the union of the inverse images $f^{-1}(p(x))$ of all $p(x) \in P_d$. Every nonzero polynomial of degree d has finitely many roots—in fact, it has at most d real roots[4]; this means that $f^{-1}(p(x))$ has at most d elements. Thus, if we show that P_d is countable, we will be able to conclude that R'_d is countable, as a countable union of finite sets.

Finally, P_d *is* countable. Indeed, every polynomial of degree d with integer coefficients may be written as

$$a_0 + a_1 x + \cdots + a_d x^d \quad ,$$

where $a_i \in \mathbb{Z}$. Thus, P_d is isomorphic to a subset of the product $\underbrace{\mathbb{Z} \times \cdots \times \mathbb{Z}}_{d \text{ times}}$; and finite products of countable sets are countable (Example 5.3.3 again). □

This argument sounds trickier than it really is. The point is really that the polynomial equations of a given degree and with integer coefficients can be listed; and each equation of degree d has at most d roots, so the

[4] You are probably familiar with this fact from elementary algebra; if not, you will have to trust me.

do not distribute

set of all roots of all polynomial equations of a given degree with integer coefficients is countable. And then $\mathbb{A}$ is a countable union of such sets, so it too is countable.

Remark 5.3.6. A *complex* number may be algebraic: for example, the imaginary number i is a root of the polynomial x^2+1, which has integer coefficients. The proof of Proposition 5.3.5 really shows that the set of *all* algebraic numbers (not just the real ones) is countable. This is a very important set; it is denoted $\overline{\mathbb{Q}}$, and it is called the 'algebraic closure of $\mathbb{Q}$'. There is an amazingly subtle and clever way to simultaneously permute the elements of $\overline{\mathbb{Q}}$ while leaving $\mathbb{Q}$ untouched. These permutations form the *Galois group of* $\overline{\mathbb{Q}}$, arguably one of the most interesting objects in mathematics(!). Much of number theory is captured by features of this group. ⌟

By Proposition 5.3.5, $\mathbb{A}$ is another copy of the set $\mathbb{N}$: it is arranged in yet another way to make it look even thicker than $\mathbb{Q}$, but it is not 'larger' than $\mathbb{Q}$ or $\mathbb{N}$ in an abstract sense. By contrast, there must be an unsurmountable gap between $\mathbb{A}$ and $\mathbb{R}$: not only most real numbers are irrational; they are not even algebraic.

Definition 5.3.7. A real number is *transcendental* if it is not algebraic. I will denote by $\mathbb{T} = \mathbb{R} \smallsetminus \mathbb{A}$ the set of transcendental real numbers. ⌟

Corollary 5.3.8. *The set $\mathbb{T}$ of transcendental real numbers is uncountable.*

Proof. If $\mathbb{T}$ were countable, then $\mathbb{R} = \mathbb{A} \cup \mathbb{T}$ would be countable, contradicting Theorem 5.2.8. □

In fact, the set of transcendental numbers has the same cardinality as $\mathbb{R}$; you will be able to prove this yourself (Exercise 5.3.3).

Corollary 5.3.8 is particularly striking because it is usually very hard to prove that any given number is transcendental. For example, π and e are known to be transcendental, but it is surprisingly hard to prove that they are. No one knows whether $\pi + e$ is transcendental, and progress on such questions would be considered very valuable. On the other hand, we have just proved for free that there *are* transcendental numbers, and that in fact there are uncountably many. A 'random' real number is 100% likely to be transcendental.

5.3.3 *The power of the continuum*

Most of the examples we have reviewed so far turned out to be countable; the exception is the set of transcendental numbers, which happens to be uncountable (Corollary 5.3.8). The argument that proved this fact may be used to obtain many other examples of uncountable sets.

Example 5.3.9. The reasoning behind Corollary 5.3.8 is that removing a countable subset from an uncountable set S leaves an uncountable set: indeed, if $C \subseteq S$ is countable and $S \smallsetminus C$ is also countable, then $S = C \cup (S \smallsetminus C)$ is countable (Example 5.3.1). This simple observation, or applications of Lemma 5.1.2, often makes it easy to show that a set is uncountable, or to prove that two sets have the same cardinality.

For example, consider the set C of fractions $\frac{1}{n}$, $n \geq 2$, inside the open interval $U = (0,1)$. Since U is isomorphic to $\mathbb{R}$ (Example 3.4.5), it is uncountable; on the other hand C is countable (for example, $n \mapsto \frac{1}{n+2}$ gives an isomorphism $\mathbb{N} \to C$); so $U \smallsetminus C$ is uncountable. Now consider the set C' consisting of C and of the number 1. Clearly $\mathbb{N} \cong C'$: send n to $\frac{1}{n+1}$ to get a bijection. By the 'simple-minded' Lemma 5.1.2, the sets

$$U = C \cup (U \smallsetminus C) \quad \text{and} \quad U' = C' \cup (U \smallsetminus C)$$

are isomorphic: the partitions $\{C, U \smallsetminus C\}$ of U and $\{C', U \smallsetminus C\}$ of U' match.

It follows that $(0,1] = U'$ is isomorphic to $(0,1) = U$, so $(0,1] \cong \mathbb{R}$ as well.

do not distribute

Arguing similarly you can conclude that many subsets of $\mathbb{R}$—every nonempty interval, closed or open on either side, and unions of such intervals, etc.—are all isomorphic to $\mathbb{R}$. Once you work out Exercise 5.3.2, it will be clear that the set of transcendental numbers is also isomorphic to $\mathbb{R}$. ⌟

Once more, this may strike you as counterintuitive. The intervals $(0,1)$ and $(0,1]$ look 'qualitatively' different: in the latter, the element 1 is clearly 'special' in some sense. Well, this element is *not* special from the set-theoretic point of view; it is special from the 'ordering' point of view (1 is the maximum of $(0,1]$, while $(0,1)$ does not have a maximum), but the order relation is extra information, and we have already encountered situations in which the choice of ordering can make isomorphic sets look very different (cf. the considerations towards the end of §5.2).

The moral here is that the notion of 'isomorphism *of sets*' is rather loose. From the point of view of topology, the closed interval $[0,1]$ and the open interval $(0,1)$ are *not* isomorphic: one is 'compact', and the other is not. These two sets are not isomorphic *as topological spaces* (they are not *homeomorphic*); while they are isomorphic *as sets*. The difference is even more dramatic in the comparison between $\mathbb{R}$ and the set of transcendental numbers.

This will of course become clearer once you learn some topology; you will learn more than enough when you begin studying real analysis. In the last chapter of these notes I will review very basic concepts in topology, but not quite enough to fully appreciate these subtleties.

All the sets we have considered here are subsets of $\mathbb{R}$, and one result we are about to prove will simplify further the computation of their cardinality. You may be a little surprised that all the examples turn out to be either countable (isomorphic to $\mathbb{N}$) or to have the same cardinality of $\mathbb{R}$ (isomorphic to $2^{\mathbb{N}}$). The cardinality of $\mathbb{R}$ is called the *cardinality (or* power*) of the continuum.* The *continuum hypothesis* states that there is no cardinality strictly between $|\mathbb{N}|$ and $|2^{\mathbb{N}}|$. This was formulated by Cantor in 1877, and attempts to prove it were fruitless. Much later (around 1960, through work of Paul Cohen) it was finally understood that the hypothesis *cannot* be proved or disproved using standard ('Zermelo-Fränkel') set theory: accepting the hypothesis or rejecting it are both consistent with the Zermelo-Fränkel axioms (provided these are themselves consistent).

5.3.4 $\mathbb{R} \cong \mathbb{R} \times \mathbb{R}$

By Lemma 5.1.3, $2^{(2^{\mathbb{N}})}$ has a 'larger' cardinality than the power of the continuum. It is easy to see that less drastic operations do not produce cardinalities larger than $|2^{\mathbb{N}}|$. Here is a remarkable example.

Proposition 5.3.10. $|2^{\mathbb{N}} \times 2^{\mathbb{N}}| = |2^{\mathbb{N}}|$.

Proof. The set $2^{\mathbb{N}} \times 2^{\mathbb{N}}$ consists of pairs (α, β), where both α and β are functions $\mathbb{N} \to \{0,1\}$. Given such a pair, define a function $\gamma : \mathbb{N} \to \{0,1\}$ as follows: $(\forall n \in \mathbb{N})$

$$\gamma(n) = \begin{cases} \alpha\left(\frac{n}{2}\right) & \text{if } n \text{ is even} \\ \beta\left(\frac{n-1}{2}\right) & \text{if } n \text{ is odd} \end{cases}$$

The assignment $(\alpha, \beta) \mapsto \gamma$ defines a function $f : 2^{\mathbb{N}} \times 2^{\mathbb{N}} \to 2^{\mathbb{N}}$. Conversely, given a function $\gamma : \mathbb{N} \to \{0,1\}$, define two functions $\alpha, \beta : \mathbb{N} \to \{0,1\}$ by

$$\alpha(n) := \gamma(2n) \quad , \quad \beta(n) := \gamma(2n+1) \quad .$$

The assignment $\gamma \to (\alpha, \beta)$ defines a function $g : 2^{\mathbb{N}} \to 2^{\mathbb{N}} \times 2^{\mathbb{N}}$.

do not distribute

The two functions f, g are inverses of each other, so they are both isomorphisms. This proves $2^{\mathbb{N}} \times 2^{\mathbb{N}} \cong 2^{\mathbb{N}}$, as needed. □

It is just as easy to prove that any countable (even infinite) product $2^{\mathbb{N}} \times 2^{\mathbb{N}} \times \cdots$ still has 'just' the cardinality of the continuum. It follows that $(\mathbb{R})^{\mathbb{N}}$ is certifiably smaller than $2^{\mathbb{R}}$.

Oh, but wait—Doesn't Proposition 5.3.10 imply that the cardinality of the 'real line' $\mathbb{R}$ equals the cardinality of the 'plane' $\mathbb{R}^2$? Yes. And once again, if you find this counterintuitive, you are in excellent company. Cantor himself, who first understood this fact, wrote about it to Dedekind, on June 29, 1877, adding (in French in the middle of a letter in German, for whatever reason) *'Je le vois, mais je ne le crois pas!'* ('I see it, but I don't believe it!')

It is in fact the case that if S is *any* infinite set, then $|SxS| = |S|$.

5.3.5 *Homework*

▷ **5.3.1:** Prove that any finite product $A_1 \times \cdots \times A_r$ of countable sets is countable.

Question: is a *countable* product of countable sets, $\prod_{n\in\mathbb{N}} A_n$, necessarily countable?

▷ **5.3.2:** Prove that if S is uncountable, and $C \subseteq S$ is countable, then $|S \smallsetminus C| = |S|$. Hint: Let $N \subseteq (S \smallsetminus C)$ be an infinite countable set, so $S = (S \smallsetminus (C \cup N)) \cup (C \cup N)$, while $S \smallsetminus C = (S \smallsetminus (C \cup N)) \cup N$.

▷ **5.3.3:** Prove that the set $\mathbb{T}$ of transcendental number has the cardinality of the continuum. (Do not use the continuum hypothesis.)

5.3.4: Prove that the set $\mathbb{C}$ of complex numbers has the cardinality of the continuum.

5.4 *Cantor-Bernstein-Schröder*

5.4.1 *The Cantor-Bernstein-Schröder theorem*

We already know how to compare cardinal numbers: we have agreed (§5.1) that we write $|A| = |B|$ to mean that the sets A and B are isomorphic, and (§5.2) we write $|A| \leq |B|$ to mean that there is an injective function $A \hookrightarrow B$.

You have already proven (Exercises 5.2.1 and 5.2.2) that this is a well-defined notion, and that the 'relation' $\leq$ among cardinalities satisfies reflexivity and transitivity. Why the quotes? Relations are defined on *sets,* and I don't want to wade into the question of whether there is a 'set of cardinalities'. I will gloss over this point: we can still make sense of $\leq$ as a notation, and ask whether this notation satisfies the usual properties we consider for relations. (In case you are wondering: there actually is no such set.)

As such, the choice of the symbol $\leq$ of course hints that $\leq$ should be an *order* relation. You have already checked reflexivity and transitivity, but what about antisymmetry? This is actually a big deal: it is by no means trivial, and is often stated as a 'named theorem' to emphasize its importance.

Theorem 5.4.1 (Cantor-Bernstein-Schröder theorem)**.** *Let A and B be sets, and assume that there is an injective function $f : A \hookrightarrow B$ and an injective function $g : B \hookrightarrow A$. Then A and B are isomorphic.*

Of course what this theorem says is that if $|A| \leq |B|$ and $|B| \leq |A|$, then $|A| = |B|$. This verifies the antisymmetric property for the 'relation' $\leq$ among cardinal numbers.

do not distribute

We will prove this theorem in short order. It is worth noting first that the fact that $\leq$ is an order relation simplifies some of the considerations we have used in order to analyze the examples in previous sections. For instance, let $\mathbb{A}$ denote the set of algebraic numbers, and let S be any set sandwiched between $\mathbb{N}$ and $\mathbb{A}$:

$$\mathbb{N} \subseteq S \subseteq \mathbb{A} \quad .$$

Then

$$|\mathbb{N}| \leq |S| \leq |\mathbb{A}| \quad .$$

But $|\mathbb{A}| = |\mathbb{N}|$: this is Proposition 5.3.5. Therefore we discover that $|\mathbb{N}| \leq |S| \leq |\mathbb{N}|$, and invoking Theorem 5.4.1 we can conclude that $|S| = |\mathbb{N}|$. Therefore, no *ad hoc* argument is needed in order to prove that $\mathbb{Z}$ or $\mathbb{Q}$ (or any other subset of $\mathbb{A}$) is countable.

Similarly, since we know that the open interval $(0, 1)$, and in fact any open interval (a, b) with $a < b$, is isomorphic to $\mathbb{R}$ (Example 3.4.5), it follows that *any* subset of $\mathbb{R}$ containing an open interval has the cardinality of the continuum. The special considerations we used in Example 5.3.9 in order to establish that $(0, 1] \cong \mathbb{R}$ are not necessary, once we can use Theorem 5.4.1.

Theorem 5.4.1 also allows us to formalize the ideas we mentioned at the end of §5.1, following Theorem 5.1.5. All the cardinalities

$$|\mathbb{N}| < \left|2^{\mathbb{N}}\right| < \left|2^{(2^{\mathbb{N}})}\right| < \left|2^{\left(2^{(2^{\mathbb{N}})}\right)}\right| < \cdots \tag{5.4}$$

are different: indeed, if there were repetitions, then we would have $\alpha < \beta < \alpha$ for some cardinalities α, β in this list. But this would imply $\alpha \leq \beta \leq \alpha$, and by Theorem 5.4.1 it would follow that $\alpha = \beta$, contradicting the assumption $\alpha < \beta$. Thus, Theorem 5.4.1 has the following consequence.

Corollary 5.4.2. *There exist infinitely many different infinite cardinal numbers.*

This fact could be proven more directly, but it is nice to see it as a straightforward consequence of the material we have developed.

Further, a 'set of all sets' S would have to contain 2^S as a subset (every element of 2^S would be a set, so it would be an element of S), and then we would have $|2^S| \leq |S|$ and $|S| \leq |2^S|$. Theorem 5.4.1 would then imply that $|S| = |2^S|$. This again would contradict Lemma 5.1.3. So there really cannot be a 'set of all sets'. Hopefully this clarifies the issues raised by Russell's paradox, about which you learned when you worked on Exercise 2.1.12. You may find the standard explanation of Russell's paradox simpler than this argument, but if you think about it a little, you will recognize that the proof of Lemma 5.1.3 is really nothing more than a formal version of that simple idea.

Incidentally, there are 'many more' cardinalities than the ones appearing in the countable collection (5.4). In fact, there are so many cardinals that they don't form a set, just as there are too many sets to form a 'set of all sets'.

You may enjoy the article on the notion of 'infinity' in the April 1995 issue of the Scientific American. It reviews for the layperson several topics that we have discussed in the past few sections.

5.4.2 *Proof of the Cantor-Bernstein-Schröder theorem*

The first proofs found for Theorem 5.4.1 were rather difficult. In time, easier proofs have been discovered, and the one we will see is quite easy, but possibly more involved than most other proofs in these notes.

Recall what the theorem states: if A and B are sets, and we have two injective functions $f : A \hookrightarrow B$ and $g : B \hookrightarrow A$, then A and B are in fact isomorphic. This is what we have to prove.

do not distribute

This fact is completely clear if A and B are finite, both because of considerations based on counting elements (which is an option that is not available in the infinite case) and because if f and g are as above, then $g \circ f$ is an injective map *from A to itself.* Back in Exercise 3.3.1 you verified that *if A is finite,* this forces $g \circ f$ to be a bijective map. The same would apply to $f \circ g$, and it would follow easily that f and g are necessarily isomorphisms. This argument does *not* work if A and B are infinite: for example, take $A = B = \mathbb{Z}$ and let both f and g be defined by $n \mapsto 2n$. Then $g \circ f$ and $f \circ g$ are both given by $n \mapsto 4n$, and this function is not an isomorphism.

This tells us that an argument working in the general case must be substantially more involved. We will use the 'simple-minded' Lemma 5.1.2: given two sets A and B and injective functions f and g as in the statement, we will break up A and B in three pieces each: A', A'', A''' (forming a partition of A) and B', B'', B''' (forming a partition of B). We will engineer the construction so that it comes with bijections $A' \to B'$, $A'' \to B''$ and (this is the clever part) $B''' \to A'''$. It will follow that $|A'| = |B'|$, $|A''| = |B''|$, $|A'''| = |B'''|$, and by Lemma 5.1.2 we will be able to conclude that $|A| = |B|$.

Here is the complete argument.

Proof. We are going to use Lemma 5.1.2: we will find matching partitions of A and B, and by Lemma 5.1.2 this will prove that A and B are isomorphic.

Given an element $a = a_0 \in A$, either $a_0 \in \operatorname{im} g$ or $a_0 \notin \operatorname{im} g$. In the first case, there is a unique $b_0 \in B$ such that $g(b_0) = a_0$: unique because g is injective. Again, either $b_0 \in \operatorname{im} f$ or $b_0 \notin \operatorname{im} f$; in the first case, there is a unique $a_1 \in A$ such that $f(a_1) = b_0$. Continuing this process, we construct a sequence

$$a_0 \quad , \quad b_0 \quad , \quad a_1 \quad , \quad b_1 \quad , \quad \cdots$$

such that $a_i \in A$, $b_i \in B$, and $a_i = g(b_i)$, $b_i = f(a_{i+1})$:

$$a = a_0 \xleftarrow{\ g\ } b_0 \xleftarrow{\ f\ } a_1 \xleftarrow{\ g\ } \cdots \xleftarrow{\ g\ } b_{i-1} \xleftarrow{\ f\ } a_i \xleftarrow{\ g\ } b_i \xleftarrow{\ f\ } a_{i+1} \xleftarrow{\ g\ } \cdots$$

One of three possibilities has to occur:

- Either this sequence never stops;
- or it stops at an element $a_N \in A$;
- or it stops at an element $b_M \in B$.

The sequence will stop at an element of A when that element is not in the image of g, and it will stop at an element of B when that element is not in the image of f. Let A' be the subset of elements $a \in A$ for which the sequence never stops, A'' the subset of elements for which it stops at A, and A''' the subset of elements for which it stops at B. We have obtained a partition[5] $\{A', A'', A'''\}$ of A.

In exactly the same fashion we can obtain a partition $\{B', B'', B'''\}$ for B: B' is the set of elements $b \in B$ for which the sequence obtained by chasing inverse images does not stop, B'' is the set of elements for which it stops at A, and B''' is the set of elements for which it stops at B.

By Lemma 5.1.2, in order to prove that $A \cong B$ it suffices now to prove that

(i) $A' \cong B'$;

(ii) $A'' \cong B''$;

[5] It may actually occur that some of these sets are empty; in that case, simply omit them from this list.

do not distribute

(*iii*) $A''' \cong B'''$.

We conclude the proof by proving these three facts.

(*i*) If $a \in A'$, then $b = f(a) \in B'$. Indeed, starting the process with $b = f(a) \in B$ first produces a, and the sequence originating from a does not stop. Therefore the sequence originating from b does not stop, and this tells us $b \in B'$. Therefore, f restricts to a function $f' : A' \to B'$. Since f is injective, so is f'. To prove that f' is surjective, let $b \in B'$; then the sequence originating from b does not stop, and in particular it begins by producing an $a \in A$ such that $f(a) = b$. As the sequence does not stop, a is in fact an element of A', and $f'(a) = b$. This proves that f' is surjective. It follows that f' is both injective and surjective, proving that $A' \cong B'$.

(*ii*) Let $a \in A''$, so that the sequence originating from a stops at A. It follows that the sequence originating from $b = f(a)$ also stops at A, showing that $b = f(a) \in B''$. Therefore, f restricts to a function $f'' : A'' \to B''$. Since f is injective, so is f''. To see that f'' is surjective, let $b \in B''$; then the sequence originating from b stops at A, and in particular it does *not* stop at b: therefore there exists an $a \in A$ such that $f(a) = b$. As the sequence stops at A, a is in fact an element of A'', and $f''(a) = b$. This proves that f'' is surjective. It follows that f'' is both injective and surjective, proving that $A'' \cong B''$.

(*iii*) Repeat precisely the same argument as in case (*ii*), but starting from B''': prove that g restricts to a function $g''' : B''' \to A'''$, and then argue that g''' is injective (because g is injective) and surjective (arguing as in the previous steps). This shows that $B''' \cong A'''$, and hence $A''' \cong B'''$ since the relation of isomorphism is symmetric. □

5.4.3 *An example*

To understand better a proof such as the one we have just gone through, it is often very helpful to work it out on a simple example. Take the one indicated in the comments preceding the proof: $A = B = \mathbb{Z}$, and

$$f : A \to B \quad \text{defined by } f(n) := 2n \quad , \quad g : B \to A \quad \text{defined by } g(n) := 2n \quad .$$

What are the sets A', A'', etc. in this case? Can we 'see' that they indeed match as the proof shows?

The main tool in the proof is the 'sequence' determined by an element of A or B. One obtains this sequence by taking as many preimages as one can via f and g. Now note that $\operatorname{im} f$ and $\operatorname{im} g$ both consist of *even* numbers; if a number $a \in A$ is odd, it has no preimages by g, and the sequence it generates stops immediately. If it is even, then its preimage is $a/2$. Taking preimages amounts to dividing by 2, and any nonzero number can be divided by 2 only finitely many times, so we see that no nonzero number belongs to the set A'. On the other hand, 0 *can* be divided by 2 as many times as we want, so $0 \in A'$. Therefore:

$$A' = \{0\}$$

and $B' = \{0\}$ by exactly the same reason. Next, note that every nonzero integer $a \in A$ can be written as a product of a power of 2, say 2^r, by an odd number. Since taking preimages amounts to dividing by 2, we see that the sequence starting from $a = 2^r \cdot \text{odd}$ will be r step long. For example, take $a = 2^5 \cdot 3 = 96 \in A$. Then the sequence is

$$96 \xleftarrow{g} 48 \xleftarrow{f} 24 \xleftarrow{g} 12 \xleftarrow{f} 6 \xleftarrow{g} 3$$

and has 5 steps as expected. With notation as in §5.4.2, we have $a_0 = 96$, $b_0 = 48$, $a_1 = 24$, $b_1 = 12$, $a_2 = 6$, and finally $b_2 = 3$. The sequence stops at B, so $96 \in A'''$.

do not distribute

In general, at this point it should be clear that if r is even, the sequence will stop at A; if r is odd, it stops at B. So

$$A'' = \{a \mid a = 2^r \cdot m \quad \text{with } m \text{ odd and } r \text{ even}\}$$

and

$$A''' = \{a \mid a = 2^r \cdot m \quad \text{with } m \text{ odd and } r \text{ odd}\}$$

For example, all odd numbers belong to A'', while the even numbers that are not multiples of 4 (such as 2, 6, 10, 14, etc.) belong to A''', and so do the multiples of 8 that are not multiples of 16, etc.

The situation with B is analogous, but 'reversed':

$$B'' = \{a \mid a = 2^s \cdot m \quad \text{with } m \text{ odd and } s \text{ odd}\}$$

and

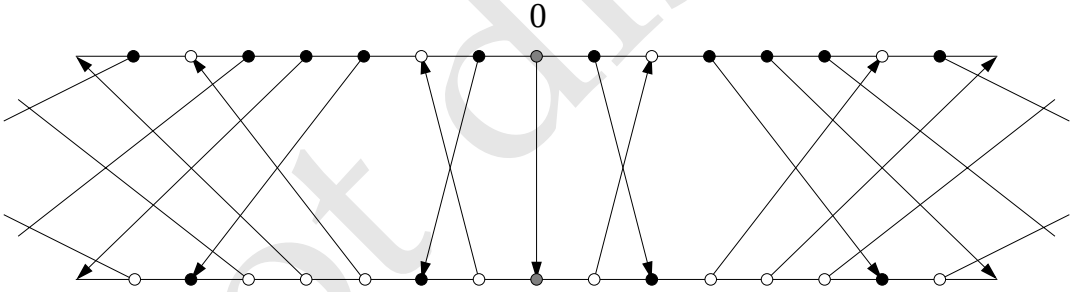
$$B''' = \{a \mid a = 2^s \cdot m \quad \text{with } m \text{ odd and } s \text{ even}\}$$

For example, all odd numbers belong to B'''.

It is clear that $A' \cong B'$. As for A'' and B'', note that multiplying by 2 a number changes the parity of the exponent of its 2^r factor: if $a = 2^r m$, with m odd and r even, then $f(a) = 2^{r+1}m$ has $r+1$ odd. Further, if $b = 2^s m$ with m odd and s odd, then in particular $s \geq 1$, so we can divide b by 2; so $b = f(a)$ with $a = 2^{s-1}m$. This confirms that f induces a function $A'' \to B''$, and that this function is surjective as well as injective.

Exactly the same reasoning works for the restriction of g to $B''' \to A'''$.

Here is a picture of the resulting isomorphisms $A' \to B'$, $A'' \to B''$, $B''' \to A'''$. The black dots stand for the elements of A'', B''; the white ones for A''', B'''. Note the pleasant symmetry, due to the fact that I have chosen the same function for f and g:



5.4.4 *Homework*

5.4.1: Let A, B be sets. Assume that there is a surjective function $A \to B$, and a surjective function $B \to A$. Prove that $|A| = |B|$, or find a counterexample to this statement.

▷ **5.4.2:** What is the cardinality of the set $\mathbb{N}^{\mathbb{N}}$ of all functions $\mathbb{N} \to \mathbb{N}$? What about the set of all relations on $\mathbb{N}$? (Hint: What is a relation? what is a function?)

5.4.3: Go through the proof of Theorem 5.4.1 for $A = B = \mathbb{N}$, and

$$f : A \to B \quad \text{defined by } f(n) := n+1 \quad , \quad g : B \to A \quad \text{defined by } g(n) := n+2 \quad ,$$

as I did for the example in §5.4.3: find out what A', A'', etc. are, and describe the isomorphisms $A' \to B'$, etc.

do not distribute

6
Very basic topology

6.1 Continuity, XIX-century version.

6.1.1 Expressions depending on quantifiers, again

One moral we should draw from our little excursion through cardinality is that 'sets' by themselves have very little structure: $\mathbb{N}$, $\mathbb{Z}$, $\mathbb{Q}$ are just different manifestations of the 'same' set, even if they look quite different (§5.2); even $\mathbb{R}$ and $\mathbb{R}^2$ turn out to be isomorphic sets (§5.3.4), even if they are completely different from the point of view of geometry. The distinction I made in (for instance) Example 2.4.4 between describing a quotient as a circle rather than as an interval is not set-theoretic: circles and intervals are isomorphic as sets, and they are just isomorphic to $\mathbb{R}$. If the set were all I care about, I could represent the same quotient by the set $\mathbb{T}$ of transcendental numbers: that, too, is in bijection with $\mathbb{R}$ (as we have learned in §5.3.2, at least if we did our homework). But doing so would definitely lose sight of the information that one is trying to convey.

What distinguishes a circle from an interval is *topology*, not set theory. On the other hand, topology itself is built by using the language of set theory; we have learned enough that we can understand the most basic definitions in topology, and this would illustrate well how the language of naive set theory may be used to build from scratch other fields of mathematics. Even if we do not have the room to develop enough material to really give justice to the examples we have encountered, it is useful to see how the subject begins.

In fact, my main goal will be to show that taking a rather abstract topological standpoint leads to an enhanced understanding (and in a sense a substantial simplification) of notions with which the reader is already familiar from calculus. For this, we go back to our very first encounter with quantifiers in §1.1.5. Back then we were considering expressions such as (1.2), which I will reproduce here:

$$\forall \epsilon > 0\,,\, \exists N\,,\, \forall n\,:\quad n > N \implies |s_n| < \epsilon \tag{6.1}$$

(where implicitly $\epsilon \in \mathbb{R}$, $n, N \in \mathbb{N}$, $s_n \in \mathbb{R}$) and we practiced 'understanding' what they mean. This particular expression means that s_n is as close as we please to 0, for all n larger than some N which will depend on the approximation we need[1]. This may or may not be true for a given s_n. For example, (6.1) is *not* true if $s_n = 1$ for all n: if $s_n = 1$, then $|s_n| \not< \epsilon$ as soon as I choose ϵ smaller than 1, and no choice of N will change this. On the other hand, (6.1) is true if (for example) $s_n = 1/n$: given $\epsilon > 0$, let N be any integer larger than $1/\epsilon$; then

$$n > N \implies n > \frac{1}{\epsilon} \implies \frac{1}{n} < \epsilon \implies |s_n| < \epsilon \quad .$$

[1] If you worked your way through §4.3, then you have handled more complicated expressions of essentially the same type.

do not distribute

Writing long expressions such as (6.1) is cumbersome. The official shorthand is

$$\lim_{n\to\infty} s_n = 0$$

and we say that *'the limit of the sequence s_n as n goes to infinity is 0'*.

Of course you know this! Every self-respecting calculus book gives this definition, and/or the version for functions that we will encounter below. Chances are, however, that you did not spend much time contemplating and understanding these definitions back then; you get a second chance now.

It is important to understand what the shorthand $\lim_{n\to\infty} s_n = 0$ does *not* say, at face value: despite appearances, it does not say that there is a number $\lim_{n\to\infty} s_n$, and that number is 0. What it says is (6.1): $\forall \epsilon$ $\exists N \ldots$, and anything else, including the interpretation as a 'computation', must be proven and justified on the basis of (6.1) alone. In fact, the study of this type of expressions (and of their version for functions) is a staple of elementary courses in analysis. The end-result of this study is a number of key facts that lead to efficient ways to 'compute' limits; those are the facts that you learned in a good calculus course.

6.1.2 *Limits and continuity, calculus-style*

The definitions of limits and continuity for functions that you (may) have encountered in your calculus courses are tailored to the case of functions from $\mathbb{R}$ to $\mathbb{R}$; I will recall these definitions below. In the rest of the section I will introduce a seemingly very different notion of 'continuity', based on abstract point-set topology. In §6.3 I will show that these notions are fully compatible. What may be surprising about this story is that the topological notion, which is completely general, is in a sense much easier to grasp than the very particular case of functions $\mathbb{R} \to \mathbb{R}$ that you have encountered in calculus. This is a good example of a common phenomenon in mathematics: the general, more powerful notions often end up being *simpler,* more transparent, than the particular cases that arise in more restricted contexts.

The version of (6.1) for functions $f : \mathbb{R} \to \mathbb{R}$, for a 'limit' of L at c, is:

$$\forall \epsilon > 0\,,\ \exists \delta > 0 \,:\ 0 < |x - c| < \delta \implies |f(x) - L| < \epsilon \quad . \tag{6.2}$$

The shorthand for this expression is

$$\lim_{x\to c} f(x) = L \quad :$$

if (6.2) is satisfied, we say that *'the limit of $f(x)$ as x goes to c is L'*.

Parsing (6.2) requires exactly the same type of thinking that we practiced at the beginning of these notes, and that we also engaged when dealing with Cauchy sequences in §4.3. In working out concrete examples, it is useful to keep in mind that

$$|x - c| < \delta$$

means that $c - \delta < x < c + \delta$, i.e., $x \in (c - \delta, c + \delta)$.

Example 6.1.1. Let's verify that

$$\lim_{x\to 2} x^2 = 4$$

according to (6.2). For any given arbitrary ϵ, we should find a δ such that

$$0 < |x - 2| < \delta \implies |x^2 - 4| < \epsilon \quad .$$

I claim that this works if I choose $\delta = \min(\frac{\epsilon}{5}, 1)$. Indeed, if

$$|x - 2| < \min\left(\frac{\epsilon}{5}, 1\right)$$

do not distribute

then in particular $|x-2|<1$, so $1<x<3$, and it follows that $3<x+2<5$: so

$$|x+2|<5 \quad .$$

But then we have $|x-2|<\frac{\epsilon}{5}$ and $|x+2|<5$, and hence

$$|x^2-4| = |x-2|\cdot|x+2| < \frac{\epsilon}{5}\cdot 5 = \epsilon,$$

as needed. ⌟

The first time one sees this type of manipulations, they look very mysterious. How did I know that I had to choose δ as I did, in order to make sure that $|x^2-4|<\epsilon$? The answer is that I did *not* know it. I had a template in mind of how the verification would go, and as I went through it I tweaked that template to get this particular choice. This looks haphazard, and maybe it is; other possibilities for δ may work just as well. But we need just one that works, and this one does (as I verified above).

Truth be told, no one wants to 'compute limits' this way, if one can avoid it. Early on in a course in analysis one proves theorems that make handling limits much more straightforward in certain situations. For example, one proves that if $p(x)$ is a polynomial, then $\lim_{x\to c} p(x) = p(c)$. Once you know this, then applying it to $c=2$ and $p(x)=x^2$ verifies the limit in Example 6.1.1 without any work at all.

In calculus you learned several such statements, and this allowed you to compute simple limits.

Remark 6.1.2. Even the very idea that one can 'compute' limits requires a justification. Writing $\lim_{x\to c} f(x) = L$ suggests that there is a number $\lim_{x\to c} f(x)$, and that you can compute this number and verify that it equals L. This is not what (6.2) says. What (6.2) says is $\forall\epsilon\ \exists\delta\ldots$. The notation $\lim_{x\to c} f(x) = L$ (and the idea of 'computing' limits) is justified by the fact that if one L works in (6.2), then that L is *unique;* this is what allows us to think of $\lim_{x\to c} f(x)$ as a number, when the limit exists according to the definition. ⌟

How do we know that 'L is unique'? Here is what uniqueness means:

Lemma 6.1.3. *Let $f:\mathbb{R}\to\mathbb{R}$ be a function, $c\in\mathbb{R}$, and assume that*

$$\forall\epsilon>0\,,\ \exists\delta>0\ :\ 0<|x-c|<\delta \implies |f(x)-L|<\epsilon$$

and

$$\forall\epsilon>0\,,\ \exists\delta>0\ :\ 0<|x-c|<\delta \implies |f(x)-L'|<\epsilon$$

for two numbers L, L'. Then $L=L'$.

Proving this lemma is a good exercise (Exercise 6.1.2).

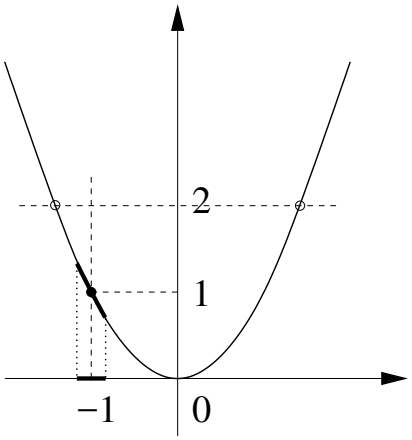
Remark 6.1.4. Consider the following argument: The first expression means $\lim_{x\to c} f(x) = L$; the second one means $\lim_{x\to c} f(x) = L'$; therefore

$$L = \lim_{x\to c} f(x) = L' \quad .$$

If you think that this is a proof of Lemma 6.1.3, then you are missing the point. Reread Remark 6.1.2 and the preceding several paragraphs several times if necessary, until you see what is going on. This is an important issue. ⌟

In any case, what (6.2) is conveying is that '*$f(x)$ will be as close as you please to L for all $x\neq c$ in a small enough neighborhood of c*'. For example, it is *not* true that $\lim_{x\to -1} x^2 = 2$:

do not distribute

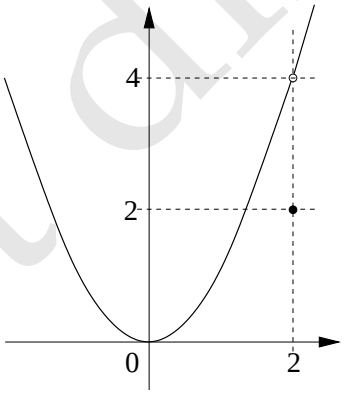


to keep $f(x)$ close to 2, x should be close to $\sqrt{2} = 1.414\cdots$, or to $-\sqrt{2}$, and if x is close to $\pm\sqrt{2}$ then it is not close to -1. For this function, if you are close to $x = -1$, then $f(x)$ is close to 1; in fact, you can get $f(x)$ to be *as close as you want to* 1 by keeping x close to -1. You can verify this formally from (6.2), and conclude that $\lim_{x\to -1} x^2 = 1$. In fact, this is another case that is very easy to handle using the result mentioned above, to the effect that $\lim_{x\to c} p(x) = p(c)$ if $p(x)$ is a polynomial.

It is of course not necessarily the case that $\lim_{x\to c} f(x) = f(c)$. In fact, this can fail to happen in many ways; here is one.

Example 6.1.5. Define

$$f(x) = \begin{cases} x^2 & \text{if } x \neq 2 \\ 2 & \text{if } x = 2 \end{cases}$$



This looks silly, but you can't stop me or anyone else from defining a function in this way. What can we say about $\lim_{x\to 2} f(x)$?

Note one subtlety in the definition of limit. The requirement for $\lim_{x\to c} f(x) = L$ is that for all ϵ there exist a δ such that

$$0 < |x - c| < \delta \implies |f(x) - L| < \epsilon \quad .$$

Note the $0 < |x - c|$ bit. This tells us that the definition *does not care* what happens when x is precisely c (i.e., when $|x - c| = 0$). So changing the value of $f(x)$ at $x = c$ *has not effect* on the definition. Therefore, with the silly $f(x)$ defined above it is still the case that

$$\lim_{x\to 2} f(x) = 4 \quad .$$

do not distribute

Never mind that $f(2) = 2$. The graph of $f(x)$ wants to go through $(2, 4)$, and the limit definition formalizes this fact. The conclusion is simply that

$$\lim_{x \to 2} f(x) \neq f(2)$$

for this particular function. ⌟

It is clear that I produced the situation in Example 6.1.5 by introducing something pathological in the definition of $f(x)$: I forced the graph of $f(x)$ to contain a point, $(2, 2)$ as it happens, that 'jumps out' of the rest of the graph. In jargon, I could say that I have introduced a *discontinuity* in the graph of $f(x)$.

Definition 6.1.6. A function $f : \mathbb{R} \to \mathbb{R}$ is *continuous* at $x = c$ if

$$\lim_{x \to c} f(x) = f(c) \quad .$$

We say that f is *continuous* if it is continuous at c for all $c \in \mathbb{R}$. ⌟

For example: polynomials are continuous since, as mentioned above, if $p(x)$ is a polynomial, then $\lim_{x \to c} p(x) = p(c)$ for all c. The function in Example 6.1.5 is not continuous at $x = 2$, and therefore it is not a 'continuous function'.

6.1.3 Homework

6.1.1: Verify formally that $\lim_{x \to -1} x^2 = 1$.

▷ **6.1.2:** Prove Lemma 6.1.3.

6.2 Continuity, XX-century version.

6.2.1 Topological spaces

Historically, the definitions of limits and continuity was settled in the XIX century, and along the lines reviewed above. Such definitions are still useful, but it is fair to say that now we understand better what continuity 'means': a general point of view was developed in the early XX century, which includes the notion of continuity I discussed in §6.1.2 as an extremely particular case. I am going to attempt to give you the basic definitions of this point of view; these would be the content of the first very few lectures in a course in topology, or the first very few pages of an introductory text in topology. Keep in mind that we are only going to scratch the surface: it is certainly possible to give ten semester-long consecutive courses in topology, with no overlap in content, and not get anywhere near the state of the art in this field[2].

Still, this is a good example of the power of generalization. The definitions I am going to give are simple and abstract, they apply to a context that is much more general than that of functions $f : \mathbb{R} \to \mathbb{R}$, and yet in the end the notion of 'continuity' arising in that context will look *simpler* than the particular case I illustrated in §6.1.2.

Definition 6.2.1. A *topology* on a set T is a family $\mathcal{U}$ of subsets of T such that

- $\emptyset \in \mathcal{U}$;
- $T \in \mathcal{U}$;

[2] The same of course applies to every other substantial field of mathematics or, I imagine, of any other discipline.

do not distribute

- Finite intersections of elements of $\mathcal{U}$ are in $\mathcal{U}$;
- Arbitrary unions of elements of $\mathcal{U}$ are in $\mathcal{U}$.

Elements of $\mathcal{U}$ are called *open sets* for the topology.

A *topological space* is a set T endowed with the choice of a topology on T. ⌟

To emphasize the 'geometric' nature of topological spaces, if a set T is endowed with a topology we call its elements 'points'.

Example 6.2.2. For every set T, the family $\{\emptyset, T\}$ is a topology on T. Indeed, the first two properties are trivially satisfied; and any union or intersection of $\emptyset$ and T is again $\emptyset$ or T, so the third and fourth properties also hold.

This topology is called the *trivial* topology on T. ⌟

Example 6.2.3. The family $\mathcal{P}(T)$ of *all* subsets of T is also a topology on T. Indeed, $\emptyset$ and T are elements of $\mathcal{P}(T)$, and any intersection/union of subsets of T is still a subset of T.

This topology is called the *discrete* topology on T. ⌟

Both the trivial and the discrete topologies are not very interesting: the first one has too few open sets, and the second one has too many. Fortunately, sets have many other topologies.

Example 6.2.4. Let $T = \{*, \circ, \bullet\}$. Then the family

$$\mathcal{U} = \{\emptyset, \{*\}, \{*, \circ\}, \{*, \circ, \bullet\}\}$$

is a topology on T. (Check this!) The family

$$\mathcal{U} = \{\emptyset, \{*\}, \{\circ\}, \{*, \circ, \bullet\}\}$$

is *not* a topology on T: indeed, $\{*\}$ and $\{\circ\}$ are elements of $\mathcal{U}$, but

$$\{*, \circ\} = \{*\} \cup \{\circ\}$$

is not: thus, the fourth property listed in Definition 6.2.1 is not satisfied. ⌟

These examples are still not particularly interesting, but they show that something is going on. Here is a somewhat more interesting example.

Example 6.2.5. Let T be any set, and let $\mathcal{U}$ consist of $\emptyset$ and of all *complements of finite subsets* of T.

I claim that $\mathcal{U}$ is a topology on T. Indeed:

- $\emptyset \in \mathcal{U}$ because I put it into it by hand;
- $T \in \mathcal{U}$ since T is the complement of $\emptyset$ in T, and $\emptyset$ is a finite set.
- Let $U_1, \ldots, U_n \in \mathcal{U}$. If any of the U_i is $\emptyset$, then $U_1 \cap \cdots \cap U_n = \emptyset \in \mathcal{U}$. If none of the U_i is $\emptyset$, then $U_i = T \smallsetminus F_i$ with F_i finite. In this case,

  $$U_1 \cap \cdots \cap U_n = (T \smallsetminus F_1) \cap \cdots \cap (T \smallsetminus F_n) = T \smallsetminus (F_1 \cup \cdots \cup F_n) \quad .$$

  A *finite* union of finite sets is finite, so this set is again in $\mathcal{U}$. This verifies the third property listed in Definition 6.2.1
- The fourth property is also satisfied: the union of the complements is the complement of the intersection, and the intersection of an arbitrary family of finite sets is a finite set.

do not distribute

This topology is called the *cofinite* topology on T.

You can imagine variations on this example: the family of complements of *countable* sets also forms a topology, on any set. ⌟

Definition 6.2.6. A *closed* set in a topological space is the complement of an open set. ⌟

One can define a topology on a set by prescribing what the *closed* sets are; the topology (that is, the family of open sets) is then defined by the complements of all these sets. For example, the 'cofinite' topology defined in Example 6.2.5 can be equivalently described by saying that the closed sets are T and all the finite subsets of T.

The family of closed sets in a topological space satisfies four properties analogous to those given in Definition 6.2.1; what are they? (Exercise 6.2.2.)

6.2.2 *Homeomorphisms*

A topological space is a set T along with a particular choice of a topology on T—that is, a family of 'open' sets $\mathcal{U}$ satisfying the properties listed in Definition 6.2.1. A notation for such a topological space could be the pair $(T, \mathcal{U})$. Often, the topology is understood and one just writes T; or one uses any symbol to denote the whole information of a set and of a topology on that set.

When should two topological spaces $(T_1, \mathcal{U}_1)$ and $(T_2, \mathcal{U}_2)$ be 'isomorphic'? We can first of all require the sets T_1, T_2 to be isomorphic as sets, but this information alone does not take into account the extra 'structure' we are imposing on the sets when we consider them along with a topology.

This is a template for a very common situation in mathematics. We need a notion that keeps track of all the information we are considering. The natural approach is to consider as an isomorphism between two topological spaces $(T_1, \mathcal{U}_1)$ and $(T_2, \mathcal{U}_2)$ a bijection $f : T_1 \to T_2$ which induces a bijection between $\mathcal{U}_1$ and $\mathcal{U}_2$: that is, a function which 'preserves the structure'. As it happens, the best way to do this turns out to be to require that

$$U \in \mathcal{U}_2 \quad \Longleftrightarrow \quad f^{-1}(U) \in \mathcal{U}_1 \quad .$$

Here, $f^{-1}(U)$ denotes the set of preimages of elements of U:

$$f^{-1}(U) = \{t \in T_1 \,|\, f(t) \in U\}$$

(Definition 3.1.2). That is, we require that f induces a perfect match between the open sets in the two topological spaces. Isomorphisms of topological spaces have a special name.

Definition 6.2.7. A *homeomorphism* between two topological spaces $(T_1, \mathcal{U}_1)$ and $(T_2, \mathcal{U}_2)$ is a bijection $f : T_1 \to T_2$ such that for all subsets U of T_2,

$$U \in \mathcal{U}_2 \quad \Longleftrightarrow \quad f^{-1}(U) \in \mathcal{U}_1 \quad .$$

Two topological spaces $(T_1, \mathcal{U}_1)$, $(T_2, \mathcal{U}_2)$are *homeomorphic* if there is a homeomorphism between them. ⌟

It is easy to verify that the relation 'homeomorphic' between topological spaces is an equivalence relation. We can also say that homeomorphic topological spaces are 'isomorphic as topological spaces'.

Example 6.2.8. Two sets T_1, T_2, both endowed with the discrete topology, are homeomorphic if and only if they are isomorphic as sets.

Similarly, if T_1 and T_2 are both endowed with the trivial topology, then they are homeomorphic if and only if they are isomorphic as sets.

do not distribute

However, take any set with more than one element, for example our friend

$$T = \{*, \circ, \bullet\} \quad .$$

Consider the topological space T_{disc} obtained by endowing T with the discrete topology, and the topological space T_{triv} obtained by endowing the same set T with the trivial topology. Then T_{disc} and T_{triv} are isomorphic *as sets* (in fact they are supported on the same set T, and T is isomorphic to itself), but they are not isomorphic *as topological spaces* (that is, they are not homeomorphic). Indeed, there are 8 open sets in the discrete topology, while there are only 2 open sets in the trivial topology, so no bijection can match the two topologies. ⌟

Example 6.2.9. Consider the set $T = \{*, \circ, \bullet\}$, and the two topologies

$$\mathcal{U}_1 = \{\emptyset, \{*\}, \{*, \circ\}, \{*, \circ, \bullet\}\} \quad \text{and} \quad \mathcal{U}_2 = \{\emptyset, \{\circ\}, \{*, \circ\}, \{*, \circ, \bullet\}\}$$

on T. Then the two topological spaces $(T, \mathcal{U}_1)$, $(T, \mathcal{U}_2)$ are homeomorphic. Indeed, the bijection $T \to T$ given by

$$\begin{cases} * \to \circ \\ \circ \to * \\ \bullet \to \bullet \end{cases}$$

is a homeomorphism, as you should verify. ⌟

At this point we have a new gadget to play with. Naive set theory gave us some objects, that is, sets, and a notion of when two sets could be 'identified' from the point of view of the theory. Our new considerations gave us a new type of objects, that is, 'topological spaces', and that came with a natural notion of when two of these objects could be identified from the point of view of the study of topology. A more refined question should be the following. In set theory we have tools to shuttle information from a set to another set, that is, *functions.* What is the natural analogue of the same notion[3] for topological spaces?

6.2.3 Continuous functions

What is the correct notion of a 'function between topological spaces'? Again, this should be a function between the corresponding sets, that 'preserves the structure' determined by the topologies.

Definition 6.2.10. A *continuous function* from a topological space $(T_1, \mathcal{U}_1)$ to a topological space $(T_2, \mathcal{U}_2)$ is a set-function $f : T_1 \to T_2$ such that

$$U \in \mathcal{U}_2 \quad \Longrightarrow \quad f^{-1}(U) \in \mathcal{U}_1$$

for all $U \in \mathcal{U}_2$, that is, all open subsets U of T_2. ⌟

That is: a function f is *continuous* if and only if the *inverse image* of every open set is open.

Remark 6.2.11. In terms of 'continuity', a homeomorphism from $(T_1, \mathcal{U}_1)$ to $(T_2, \mathcal{U}_2)$ is a bijective continuous function $f : T_1 \to T_2$, such that the inverse f^{-1} is also continuous. This should sound familiar: isomorphisms should be functions that admit an inverse; in this new context, both the function and the inverse are required to be continuous. ⌟

[3] I am hint at a powerful general concept. A *category* consists of a collection of 'objects' and of 'morphisms' between objects. So there is a category of 'sets', whose objects are sets and whose morphisms are functions. There is also a category of 'topological spaces', whose objects are topological spaces. The question is: what should the 'morphisms' be, in the category of topological spaces? The reader will encounter many more categories: vector spaces, groups, rings, fields, modules, etc., all inhabit their own categories.

do not distribute

In general, a continuous bijection need not be a homeomorphism, as the following examples show.

Example 6.2.12. Let T be any set, and let T_{disc} and T_{triv} be the two topological spaces obtained by endowing T with the discrete topology, resp. the trivial topology (as in Example 6.2.8). The identity $\mathrm{id}_T : T \to T$ is a continuous function

$$T_{\text{disc}} \to T_{\text{triv}}$$

because $\emptyset = \mathrm{id}_T^{-1}(\emptyset)$ and $T = \mathrm{id}_T^{-1}(T)$ are open in the discrete topology. On the other hand, the identity is *not* a continuous function

$$T_{\text{triv}} \to T_{\text{disc}}$$

if T consists of more than one point. Indeed, if $p \in T$, then $\{p\}$ is open in T_{disc}, but $\mathrm{id}_T^{-1}(\{p\}) = \{p\}$ is not open in T_{triv}. ⌟

Example 6.2.13. Let T be the set $\{\circ, *, \bullet\}$, and consider the two topologies $\mathcal{U}_1 = \{\emptyset, \{\circ\}, \{*\}, \{\circ, *\}, T\}$, $\mathcal{U}_2 = \{\emptyset, \{\circ\}, \{\circ, *\}, T\}$. (Check that these are topologies!) Let T_1, resp. T_2 be the topological spaces obtained by taking T along with $\mathcal{U}_1$, resp. $\mathcal{U}_2$. Since $\mathcal{U}_1$ has more elements than $\mathcal{U}_2$, no function can precisely match the open sets in the two spaces, so T_1 and T_2 are not homeomorphic.

Also, $\mathrm{id}_T : T_2 \to T_1$ is not continuous: indeed, $\{*\}$ is open in T_1, but it is not open in T_2. On the other hand, $\mathrm{id}_T : T_1 \to T_2$ *is* continuous: every open set in T_2 is also open in T_1.

The function $f : T_1 \to T_2$ defined by sending $\circ \mapsto *$, $* \mapsto \circ$, $\bullet \mapsto \bullet$ is also continuous, as you can verify. ⌟

6.2.4 Homework

6.2.1: Let T be an infinite set, and consider the family $\mathcal{U}$ of subsets of T consisting of $\emptyset$ and of all *infinite* subsets of T. Is $\mathcal{U}$ necessarily a topology?

▷ **6.2.2:** What four properties are satisfied by the family of *closed* sets in a topological space?

▷ **6.2.3:** Consider the set $\mathbb{N}$ of natural numbers as a subset of $\mathbb{R}$. Is it open, or closed, or neither in the standard topology? In the cofinite topology? (Why?)

6.2.4: Consider the set

$$S = \left\{\frac{1}{n}\right\}_{n=1,2,3,\dots} = \left\{1, \frac{1}{2}, \frac{1}{3}, \frac{1}{4}, \dots\right\} \quad .$$

Is S open, closed, or neither in the standard topology? What about $S \cup \{0\}$?

6.2.5: Consider the set $T = \{*, \circ, \bullet\}$, and the two topologies

$$\mathcal{U}_1 = \{\emptyset, \{*\}, \{*, \circ\}, \{*, \circ, \bullet\}\} \quad \text{and} \quad \mathcal{U}_2 = \{\emptyset, \{\circ\}, \{*, \circ\}, \{*, \circ, \bullet\}\}$$

on T. Are the two topological spaces $(T, \mathcal{U}_1)$, $(T, \mathcal{U}_2)$ homeomorphic? If yes, find a homeomorphism between them. If not, why not?

6.2.6: Let $\cong$ denote the relation of homeomorphism, defined in a family of topological spaces: for two topological space S and T in the family, write $S \cong T$ if S and T are homeomorphic. Prove that $\cong$ is an equivalence relation.

▷ **6.2.7:** Let $\mathbb{R}_{\text{stand}}$, resp. $\mathbb{R}_{\text{cofin}}$ be the topological spaces determined by the standard, resp. cofinite topology on $\mathbb{R}$. (Recall that the closed sets in the cofinite topology are $\mathbb{R}$ and all finite sets, cf. Example 6.2.5.) Prove that the identity $\mathbb{R} \to \mathbb{R}$ is a continuous function $\mathbb{R}_{\text{stand}} \to \mathbb{R}_{\text{cofin}}$, and that it is not a homeomorphism.

do not distribute

6.2.8: Let $(T_1, \mathcal{U}_1)$, $(T_2, \mathcal{U}_2)$, and $(T_3, \mathcal{U}_3)$ be topological spaces, and let $f : T_1 \to T_2$, $g : T_2 \to T_3$ be functions. Prove that if f and g are both continuous, then $g \circ f$ is continuous.

6.3 *Continuity for functions* $\mathbb{R} \to \mathbb{R}$

6.3.1 *The standard topology on* $\mathbb{R}$

The examples of topological spaces appearing in §6.2 are all somewhat contrived, with the possible exception of the cofinite topology (Example 6.2.5). By contrast, the topology I am going to define next on $\mathbb{R}$ is not contrived; it is in a sense the 'most natural' example, the one that historically motivated the abstract notions we are reviewing.

Definition 6.3.1. A subset U of $\mathbb{R}$ is *open* with respect to the *standard* (or 'ordinary') *topology* on $\mathbb{R}$ if it is a union of open intervals. ⌟

By 'open intervals' I mean subsets $(a, b) = \{x \in \mathbb{R} \mid a < x < b\}$, including the possibilities $(-\infty, b)$, (a, ∞), $(-\infty, \infty)$ (which would anyway be unions of intervals (a, b) with finite endpoints).

We need to check something here, and that is that the family $\mathcal{U}$ of subsets of $\mathbb{R}$ that are open in the sense of Definition 6.3.1 indeed is a topology.

Proposition 6.3.2. *Let $\mathcal{U}$ be the family of unions of any set of open intervals in $\mathbb{R}$. Then*

- $\emptyset \in \mathcal{U}$;
- $\mathbb{R} \in \mathcal{U}$;
- *Finite intersections of elements of $\mathcal{U}$ are in $\mathcal{U}$;*
- *Arbitrary unions of elements of $\mathcal{U}$ are in $\mathcal{U}$.*

That is, $\mathcal{U}$ is a topology on $\mathbb{R}$

Proof. It is clear that $\emptyset \in \mathcal{U}$ and $\mathbb{R} \in \mathcal{U}$: the second is the union of all open intervals (and it is also itself the interval $(-\infty, \infty)$), and the first is the union of none of them. It is also clear that the any union of unions of open intervals is a union of open intervals; this proves the fourth point.

Thus, we just have to verify the third point. That is, we have to verify that if $U_1, \ldots, U_n$ are unions of open intervals, then

$$U_1 \cap \cdots \cap U_n$$

is a union of open intervals. By induction[4], it suffices to prove this for $n = 2$. That is, we have to verify that if

$$U = \cup_{\alpha \in I}(a_\alpha, b_\alpha) \quad \text{and} \quad V = \cup_{\beta \in J}(c_\beta, d_\beta) \quad ,$$

then $U \cap V$ is also a union of open intervals.

For this, we use our old Theorem 2.1.11, or rather the version 'for families' of part 5 in that theorem. This

[4] Do you see why induction implies this?

do not distribute

tells us that[5]

$$U \cap V = U \cap \left(\cup_\beta (c_\beta, d_\beta)\right) = \cup_\beta \left(U \cap (c_\beta, d_\beta)\right) = \cup_\beta \left(\left(\cup_\alpha (a_\alpha, b_\alpha)\right) \cap (c_\beta, d_\beta)\right)$$
$$= \cup_\beta \left(\cup_\alpha \left((a_\alpha, b_\alpha) \cap (c_\beta, d_\beta)\right)\right) \quad .$$

That is, $U \cap V$ is a union of subsets of the type

$$(a_\alpha, b_\alpha) \cap (c_\beta, d_\beta) \quad .$$

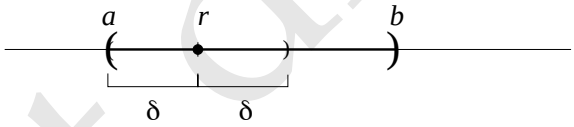
Finally, we just need to observe that these subsets are either $\emptyset$ or open intervals themselves, according to the relative position of the pairs of numbers $a_\alpha < b_\alpha$ and $c_\beta < d_\beta$. □

Definition 6.3.3. The *standard* topology on $\mathbb{R}$ is the topology in which the open sets are precisely the unions of open intervals. ⌟

When $\mathbb{R}$ is considered as a topological space, it is usually implicitly endowed with this 'standard topology' (that's why it is called 'standard'!). You will study this topology rather carefully as soon as you approach introductory material in analysis. The first thing you will prove will likely be a version of the following observation:

Proposition 6.3.4. *A subset $U \subseteq \mathbb{R}$ is open in the standard topology if and only if for all $r \in U$ there exists a $\delta > 0$ such that the open interval $(r - \delta, r + \delta)$ is contained in U.*

Proof. If U is open in the standard topology, it is a union of open intervals: $U = \cup_{\alpha \in I}(a_\alpha, b_\alpha)$. Thus, if $r \in U$, then there exists an index $\alpha \in I$ such that $r \in (a_\alpha, b_\alpha)$. Let $\delta = \min(r - a_\alpha, b_\alpha - r)$. Since $a_\alpha < r < b_\alpha$, we have $\delta > 0$; and $(r - \delta, r + \delta)$ is entirely contained in (a_α, b_α), and hence in U.



Conversely, assume that U is a subset such that for all $r \in U$ there exists an open interval $(r - \delta_r, r + \delta_r)$ contained in U. (I am denoting δ by δ_r, since this number depends on r.) Then U is the union of all such intervals, as r ranges over all its points:

$$U = \bigcup_{r \in U} (r - \delta_r, r + \delta_r) \quad .$$

This may be clear to you, and in any case you should be able to prove it formally (prove both inclusions $\subseteq$ and $\supseteq$). It follows that U is a union of open intervals, so it is open in the standard topology. □

The interval $(r - \delta, r + \delta)$ may be called the 'δ-neighborhood' of r. Proposition 6.3.4 says that if you are sitting at a point r of an open set U in the standard topology of $\mathbb{R}$, then the points 'near you' are also in U, where 'near' means 'in a neighborhood'. Topology is a way to formalize what it means to be 'near'.

[5] If you are like me, your eyes glaze over when you look at a display like this. Notation is what it is, but the content is not that bad. Work out what it says when both U and V consist of the union of just two intervals; the end-result will be that $U \cap V$ is the union of the pairwise intersection of those intervals.

do not distribute

Example 6.3.5. Open intervals (a, b) are open in the standard topology. Closed intervals $[a, b]$ are closed sets (surprise!), because the complement

$$\mathbb{R} \smallsetminus [a, b] = (-\infty, a) \cup (b, \infty)$$

is a union of open intervals, hence it is open.

Note that the interval $[a, \infty)$ is also closed: indeed, its complement is the open set $(-\infty, a)$.

By contrast, $[a, b)$ (with a, b both finite) is neither closed nor open. To see that it is not open, use Proposition 6.3.4: a is an element of $[a, b)$, and yet $[a, b)$ does not contain any neighborhood of a. The same token shows that the complement is not open (b is in the complement, but no neighborhood of b is contained in the complement). Therefore $[a, b)$ is not closed. ⌟

Example 6.3.6. Note that

$$\cap_{r>0}(-r, r) = \{0\} \quad :$$

re-do Exercise 2.1.8 if necessary. The set $\{0\}$ is not open, by Proposition 6.3.4. Thus, in the standard topology, arbitrary intersections of open sets need not be open (while finite intersections are, as we checked above). ⌟

One can define a 'standard' topology in $\mathbb{R}^n$ by mimicking the situation highlighted by Proposition 6.3.4: a set $U \subseteq \mathbb{R}^n$ is open (in the standard topology) if $\forall p \in U$, there exists a $\delta > 0$ such that the whole *open ball* with center p and radius δ is contained in U. Equivalently, open sets in the standard topology in $\mathbb{R}^n$ are unions of open balls.

6.3.2 *Continuity, reprise*

Now you may be struck by an apparent clash of terminology. Suppose $f : \mathbb{R} \to \mathbb{R}$ is a function, and view $\mathbb{R}$ as a topological space by considering it with the standard topology. According to Definition 6.2.10, we would say that f is continuous if the inverse image of a union of open intervals is a union of open intervals. On the other hand, according to the 'calculus' Definition 6.1.6, f is continuous if and only if

$$\lim_{x \to c} f(x) = f(c)$$

for all $c \in \mathbb{R}$. The calculus-based Definition 6.1.6 and the topology-based Definition 6.2.10 look *completely* different. They cannot possibly mean the same thing, can they?

They do!

Theorem 6.3.7. *Let $f : \mathbb{R} \to \mathbb{R}$ be a function, and endow $\mathbb{R}$ with the standard topology. Then f is continuous according to Definition 6.1.6 if and only if it is continuous according to Definition 6.2.10. That is,*

$$\lim_{x \to c} f(x) = f(c)$$

for all $c \in \mathbb{R}$ if and only if for every open set $U \subseteq \mathbb{R}$, $f^{-1}(U)$ is open.

Proof. What we have to show is that if $f : \mathbb{R} \to \mathbb{R}$ is a function, and we endow $\mathbb{R}$ with the standard topology, then f is continuous in the topological sense if and only if it is continuous in the sense of calculus.

Recall that f is continuous according to Definition 6.2.10 if and only if

$$U \text{ open in } \mathbb{R} \quad \Longrightarrow \quad f^{-1}(U) \text{ open in } \mathbb{R} \quad , \tag{6.3}$$

do not distribute

where here we are taking $\mathbb{R}$ with the standard topology: thus, 'open' means union of open intervals. Also recall that, by the definition of limit of functions (6.2), f is continuous according to Definition 6.1.6 if and only if

$$\forall c \in \mathbb{R}\,,\ \forall \epsilon > 0\,,\ \exists \delta > 0\ :\ |x-c| < \delta \implies |f(x)-f(c)| < \epsilon \quad . \tag{6.4}$$

Our task is to prove that (6.3) $\iff$ (6.4).

First, we prove that (6.3) implies (6.4). Let $c \in \mathbb{R}$ and $\epsilon > 0$ be arbitrary. The set

$$U = \{y \in \mathbb{R} \text{ s.t. } |y - f(c)| < \epsilon\} = (f(c)-\epsilon, f(c)+\epsilon)$$

is an open interval, so it is open. By (6.3), its inverse image $f^{-1}(U)$ is open; since $c \in f^{-1}(U)$, by Proposition 6.3.4 there exists a $\delta > 0$ such that $(c-\delta, c+\delta)$ is contained in $f^{-1}(U)$, that is:

$$x \in (c-\delta, c+\delta) \quad \implies \quad x \in f^{-1}(U) \quad .$$

Now $x \in (c-\delta, c+\delta) \iff |x-c| < \delta$; and

$$x \in f^{-1}(U) \quad \iff \quad f(x) \in U \quad \iff \quad |f(x)-f(c)| < \epsilon \quad .$$

Thus, we have verified that $\forall c \in \mathbb{R}$ and $\forall \epsilon > 0$ there exists a $\delta > 0$ such that

$$|x-c| < \delta \quad \implies \quad |f(x)-f(c)| < \epsilon \quad .$$

This is precisely what (6.4) says, so indeed (6.3) $\implies$ (6.4).

Next we will prove that (6.4) implies (6.3). Assume that f satisfies (6.4), and that U is open in $\mathbb{R}$. In order to verify (6.3), we have to prove that $f^{-1}(U)$ is open in $\mathbb{R}$. By Proposition 6.3.4, we need to show that $\forall c \in f^{-1}(U)$, there is a δ-neighborhood of c entirely contained in $f^{-1}(U)$. Let then c be an arbitrary element in $f^{-1}(U)$. Since $f(c) \in U$ and U is open, by Proposition 6.3.4 there exists an $\epsilon > 0$ such that the ϵ-neighborhood of $f(c)$ is contained in U:

$$(f(c)-\epsilon, f(c)+\epsilon) \subseteq U$$

or, equivalently:

$$\forall y \in \mathbb{R}, \quad |y - f(c)| < \epsilon \quad \implies \quad y \in U \quad .$$

Now, according to (6.4), for *every* $\epsilon > 0$ (and hence for this particular ϵ we just found) there exists a $\delta > 0$ such that

$$|x-c| < \delta \quad \implies \quad |f(x)-f(c)| < \epsilon \quad .$$

Stringing the two implications, this says that

$$\exists \delta > 0\ :\quad |x-c| < \delta \implies f(x) \in U \quad .$$

Therefore,

$$\exists \delta > 0\ :\quad x \in (c-\delta, c+\delta) \implies x \in f^{-1}(U) \quad ,$$

or simply $(c-\delta, c+\delta) \subseteq f^{-1}(U)$.

The conclusion is that if $c \in f^{-1}(U)$, then $f^{-1}(U)$ contains a δ-neighborhood of c. It follows that $f^{-1}(U)$ is open, by Proposition 6.3.4, verifying (6.3) as needed. □

do not distribute

Others may disagree, but I have a strong preference for Definition 6.2.10 over Definition 6.1.6. Definition 6.2.10 is conceptually simpler, and immeasurably more general, since it makes sense for any function from any topological space to any topological space. Definition 6.1.6 is admirable in its concreteness, but it is harder to parse once it is expanded out in its ϵ-δ form, and it applies only to functions from $\mathbb{R}$ to itself, with the standard topology. (Of course definitions in the same style can be given in a much more general setting.) If you will, Definition 6.1.6 is a gem from the XIX century; it is (I believe) due to Karl Weierstrass, and ϵ-δ considerations go further back, at least to Cauchy. It is still very useful; careful estimates in the style of ϵ-δ manipulations are still an important tool in branches of modern analysis. Definition 6.2.10 has a different flavor, and is a good example of a dominant taste in XX century mathematics. The abstract setting for topology was developed in the first half of the century by a number of mathematicians, perhaps most notably Felix Hausdorff, and reached the currently standard definition with Solomon Lefschetz. Thinking along these abstract lines may lose sight of specific contexts, where the other type of definition may be more useful. But the increased abstraction allows us to transfer tools and intuition we acquire in specific contexts to different settings, where they become invaluable. You will agree with me that Definition 6.2.10 is 'simpler' than the particular case in terms of ϵ and δ encoded by Definition 6.1.6.

The trend of unification and simplification through generalization and abstraction continues to this day.

6.3.3 *Homework*

6.3.1: Prove that a function $f : \mathbb{R} \to \mathbb{R}$ (with the standard topology) is continuous if and only if the inverse image of every open interval is a union of open intervals.

▷ **6.3.2:** Prove that the function $f : \mathbb{R} \to \mathbb{R}$ defined by $f(x) = x^2$ is continuous, by using the 'abstract' Definition 6.2.10. (In particular, $\lim_{x\to 2} x^2 = f(2) = 4$. Compare this computation of the limit to the computation in Example 6.1.1. Which one is 'easier'?)

6.3.3: Let $f : \mathbb{R} \to \mathbb{R}$ be the function considered in Example 6.1.5:

$$f(x) = \begin{cases} x^2 & \text{if } x \neq 2 \\ 2 & \text{if } x = 2 \end{cases}$$

Prove (again) that this function is *not* continuous, by using Definition 6.2.10.

6.4 *Topological invariants.*

6.4.1 *Induced topology*

Where do we go from here? How do we apply topological considerations to sets other than the toy examples included in §6.2.1?

As we have seen, $\mathbb{R}$ and all $\mathbb{R}^n$ come endowed with a 'standard' topology (cf. the end of §6.3.1). Other familiar geometric sets, such as plane figures, spheres, cylinders, other surfaces in $\mathbb{R}^3$, and so on, automatically inherit a topology from the space they sit in. More generally, a topology $\mathcal{U}$ on a set T 'induces' (that is, 'determines') a topology on every subset $S \subseteq T$. Understanding this will allow us to stretch our imagination and contemplate many more examples.

do not distribute

Definition 6.4.1. Let $(T, \mathcal{U})$ be a topological space, and let $S \subseteq T$ be a subset. The *induced topology* on S consists of the family of subsets

$$\{U \cap S\}_{U \in \mathcal{U}}$$

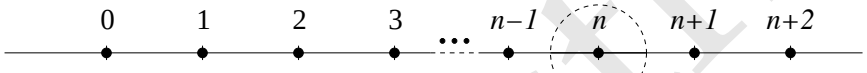
obtained by intersecting S with the open sets in $(T, \mathcal{U})$. ⌟

As always, something needs to be checked: one needs to verify that the subsets $U \cap S$, as U ranges in a topology $\mathcal{U}$, form a topology. This is a good exercise for you (Exercise 6.4.1).

For subsets of $\mathbb{R}^n$, the 'default' topology is the topology induced from the standard topology on $\mathbb{R}^n$, according to which (as we have seen) a set U is open if and only if $\forall c \in U$, U contains a small open ball centered at c. If no special mention is made of a different topology, this is the topology that is ordinarily taken for subsets of $\mathbb{R}^n$.

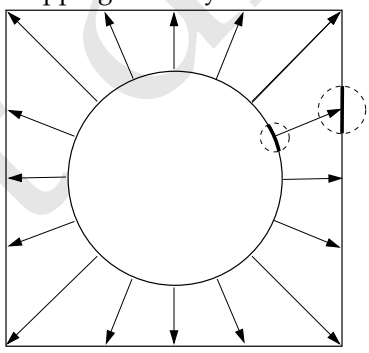
Example 6.4.2. The topology induced by the standard topology on $\mathbb{N} \subseteq \mathbb{R}$ is the discrete topology.

Indeed, for every $n \in \mathbb{N}$ the set $\{n\}$ is open in $\mathbb{N}$, because $\{n\}$ is the intersection $(n - \frac{1}{2}, n + \frac{1}{2}) \cap \mathbb{N}$ of $\mathbb{N}$ with an open interval:



Since every element of $\mathbb{N}$ is open in this topology, then so is every union of elements of $\mathbb{N}$, that is, every subset of $\mathbb{N}$. The topology in which every subset is open is the discrete topology. ⌟

Example 6.4.3. Circles are homeomorphic to (boundaries of) squares as subsets of $\mathbb{R}^2$. Indeed, we can define a bijection from a circle to a square by mapping 'radially' one to the other:

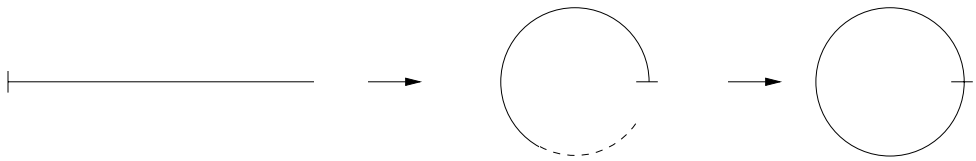


and it is clear that intersections of the square with standard open sets in $\mathbb{R}^2$ match intersections of the circle with standard open sets by means of this bijection. (The picture attempts to convey the fact that the intersection of the square with an open disk corresponds to the intersection of the circle with an open disk.)

Similarly, a sphere in $\mathbb{R}^3$ is homeomorphic to (the surface of) a cube. If you imagine the sphere as made of rubber or some other flexible material, then stretching it or adding a few bumps here and there will produce a homeomorphic surface. Such deformations 'do not change the topology', a point often made in popularizations of mathematics. ⌟

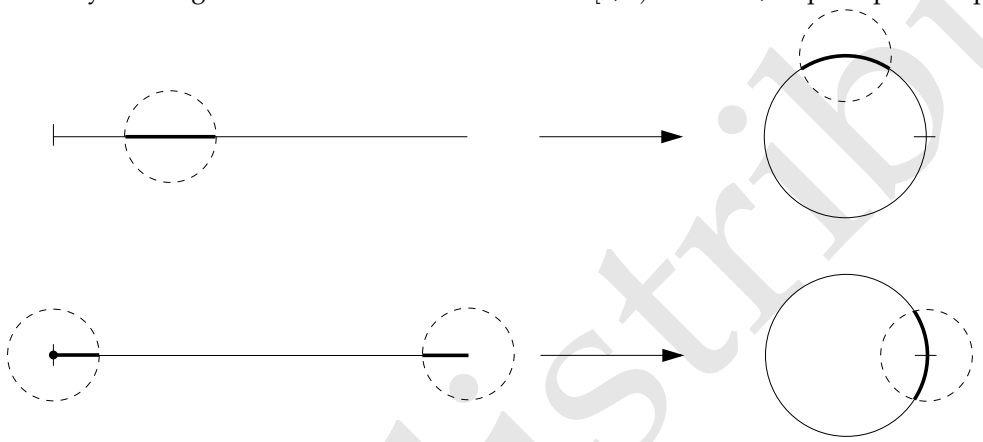
Example 6.4.4. We can map a closed-open interval $[0, 1)$ bijectively to a circle: I observed this as far back as Example 2.4.4, where I drew a picture similar to this, trying to represent the action of 'wrapping' the interval around a circle:

do not distribute



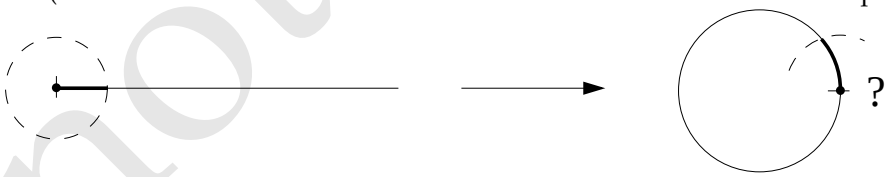
Back then we did not know enough for me to point out that the end result of this operation is a bijection, but now we do. If you want an 'analytic' way to think of this function, send the number $r \in [0,1)$ to the point on the circle determined by an angle $2\pi r$ with respect to a horizontal axis.

Now endow both $[0,1)$ and the circle with their ordinary topologies. Then it is easy to understand that this function is continuous: intersect the circle with any open disk centered at a point, and take the inverse image of the result: you will get the intersection of the interval $[0,1)$ with one, or perhaps two open intervals:



(I am drawing the intervals on the left as intersections with open balls, which makes no difference). Note that the intersection with $[0,1)$ of an open interval centered at 0 *contains* the point 0: this looks a little strange for an open set, but this is how things are according to the induced topology.

The interesting feature of this example is that the inverse of this function is not continuous. Indeed, the intersection of $[0,1)$ with an open disk centered at 0 is open in the interval, but the corresponding set in the circle is *not* open, since one of the 'endpoints' is in the set, but no neighborhood of this point is entirely contained in the set (while it would be if this were the intersection of the circle with an open set of $\mathbb{R}^2$).



Thus, this function is *not* a homeomorphism between $[0,1)$ and the circle. One can in fact verify that the interval $[0,1)$ and the circle are *not homeomorphic:* not only does this bijection fail to be a homeomorphism, but there simply is *no* homeomorphism between these two objects. As it happens, the fact that the circle has a 'loop', while the interval does not, is responsible for this fact; but you will have to study more topology to see why that is the case. I'll say one more word about this at the end of §6.4.2. ⌟

6.4.2 *Features that are invariant by homeomorphisms*

The examples in the previous section illustrate that you can bend and stretch things in topology and produce homeomorphic spaces, but you cannot introduce a loop or a 'hole' at will. You will learn how to formalize

do not distribute

such statements and prove them when you approach more advanced texts. A key idea is to study more sophisticated properties that must be preserved by homeomorphisms, or more generally introduce 'topological invariants': that is, features or quantities that one can compute and that can be proved to be the same for two topological spaces if they are homeomorphic.

For example: We say that a topological space T is *disconnected* if it contains a subset $S \neq \emptyset$, $S \neq T$ that is *both open and closed.* A space is *connected* if it is not disconnected.

Example 6.4.5. If $|T| \geq 2$, then the topological space T_{disc} obtained by taking T with the *discrete* topology is disconnected: indeed if $p \in T$, both the set $\{p\}$ and the set $T \smallsetminus \{p\}$ are open (because *every* subset of T is open in the discrete topology), so $\{p\}$ is both open and closed (and it is $\neq \emptyset$ because T has at least two points.)

In this case, the terminology goes well with the picture we may have of such a space: say T consists of exactly two points p, q; then I picture them as distinct 'pieces' of T, and this is what makes T 'disconnected'. But note that the space T_{triv} obtained by taking the same set, with the *trivial* topology, is connected: there are no open subsets of T other than $\emptyset$ and T, so T_{triv} is *not* disconnected. Pictures are helpful, but you have to make sure that they represent the situation accurately, or else they will trick your intuition. ⌟

Example 6.4.6. On the other hand, when we are dealing with the standard topology on, say, $\mathbb{R}$, then your intuition will probably be right on the mark. A set such as

$$T = [0,1] \cup [2,3]$$

does look as if it should be 'disconnected', with the topology induced from the standard topology on $\mathbb{R}$:

And indeed $[0,1]$ is open in the induced topology (because it is the intersection of T with the open interval $(-0.1, 1.1)$), and it is closed (because its complement, $[2,3]$, is the intersection of T with the open interval $(1.9, 3.1)$).

The real line $\mathbb{R}$ itself with the standard topology is connected, although it may not be too easy to verify this with the little topology you know at this point. You will become much more familiar with this notion when you approach the study of real analysis. ⌟

If T_1 and T_2 are topological spaces, and $f : T_1 \to T_2$ is a homeomorphism, then the open sets of T_1 and T_2 match via f, and so do the closed set. In particular, if $S_1 \subseteq T_1$ is both open and closed, then $f(S_1) \subseteq T_2$ must be both open and closed. It follows that *if T_1 and T_2 are homeomorphic, then they are both connected or both disconnected.* (In Exercise 6.4.4 you will prove a stronger statement.)

So here is a way to tell if two spaces are *not* homeomorphic: if one is connected and the other one is not, then necessarily they cannot be homeomorphic.

Connectedness is one of several 'features' of topological spaces, which must be the same for homeomorphic spaces. *Compactness* is another such feature. Roughly speaking, a space is compact if from every sequence of points in the space you can extract a 'subsequence' which converges (in a suitable sense, which for $\mathbb{R}$ with the standard topology agrees with the notion of 'limit' which we have visited in §6.1.2). One way to prove that $[0,1)$ and the circle (as in Example 6.4.4) are not homeomorphic to one another is to observe that $[0,1)$ is not compact (try your hand at this: Exercise 6.4.5), while the circle is. Again, you will understand this completely as soon as you begin studying e.g., real analysis.

As another example, there is a way to associate a 'group' (we touched upon this notion, at the end of §3.4.3) with a topological space, which records information about the 'loops' in the space. This is a topological

do not distribute

invariant, called the 'fundamental group'. The set $[0,1)$ has trivial fundamental group, because it contains no loops within it; the circle has fundamental group equal to $\mathbb{Z}$, because you can loop around it any integer number of times (say clockwise for positive integers, counterclockwise for negative integers). This shows again that they are not homeomorphic. You will learn about the fundamental group in graduate topology, if not sooner.

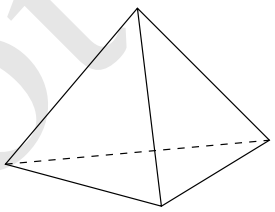
6.4.3 Dessert: the Euler characteristic

A very interesting topological invariant that is easy to contemplate (although proving anything about it rigorously requires substantial background) is the 'Euler characteristic'. This is an integer that can be computed for topological spaces satisfying some reasonable conditions (which I will not attempt to specify). For a surface, it can be heuristically defined as follows: pave your surface with triangles joined at the sides, stretching or bending them if necessary; then count the number v of vertices of your 'triangulation', the number e of edges, and the number f or faces. Then the Euler characteristic of the surface is the number

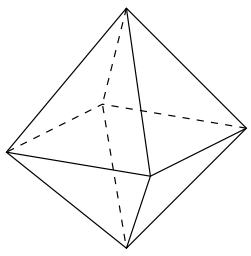
$$\chi = v - e + f \quad .$$

The first amazing observation is that this number *does not depend on how you actually triangulate your surface!* And the second observation is that if two surfaces are homeomorphic, then the corresponding number *is the same* for both. This is not too hard to believe at an intuitive level: if performing a homeomorphism amounts to stretching or bending a surface, then the effect on a triangulation will be to stretch or bend the triangles, but it will not change the number of vertices, edges, or faces. *However,* this is very far from a 'proof': homeomorphisms are much more general types of function than this intuitive idea conveys; in a simple-minded approach to a proof you would have to show that something like a 'triangle' remains a 'triangle' once you apply a homeomorphism. Good luck with that. A certain level of sophistication is needed to carry this out successfully.

But we can appreciate a few examples without worrying about these subtleties, and it is fun to do so. The tetrahedron and the octahedron are both homeomorphic to the sphere, so the number must be the same for both:



4–6+4=2

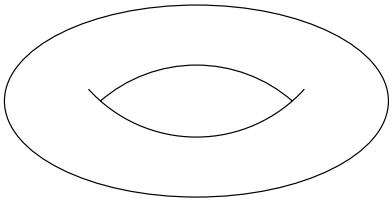


6–12+8=2

Triangulate a cube in any way you please, and you will also get 2: *the Euler characteristic of the sphere is* 2.

do not distribute

What about other surfaces? With some ingenuity, you can verify that this number is $2-2g$ for the surface of a 'pretzel' with g holes. For example, the Euler characteristic of a torus is 0.



It follows that tori are *not* homeomorphic to spheres. (This would also follow from computing the corresponding *fundamental groups*.)

Why is the Euler characteristic of a torus 0? You can figure this out for yourself! (Exercise 6.4.6.)

6.4.4 Homework

▷ **6.4.1:** Let $(T, \mathcal{U})$ be a topological space, and let $S \subseteq T$ be a subset. Verify that the collection of sets $U \cap S$, as $U \in \mathcal{U}$, forms a topology.

6.4.2: A topology $\mathcal{U}_1$ on a set T is 'finer' than a topology $\mathcal{U}_2$, and $\mathcal{U}_2$ is 'coarser' than $\mathcal{U}_1$, if $\mathcal{U}_2 \subseteq \mathcal{U}_1$: that is, if every open set with respect to $\mathcal{U}_2$ is open with respect to $\mathcal{U}_1$. Thus, the discrete topology is the finest topology on a set, and the trivial topology is the coarsest topology.

- Prove that $\mathcal{U}_1$ is finer than $\mathcal{U}_2$ if and only if the identity function $T \to T$ is continuous from the topological space $(T, \mathcal{U}_1)$ to $(T, \mathcal{U}_2)$.
- Let $(T, \mathcal{U})$ be a topological space, and let $S \subseteq T$ be a subset. Prove that the induced topology on S is the coarsest topology for which the inclusion function $i : S \to T$ is continuous.

6.4.3: Prove that $\mathbb{R}$ *with the cofinite topology* is connected.

▷ **6.4.4:** Let $f : T_1 \to T_2$ be a surjective continuous function between topological spaces. Prove that if T_1 is connected, then T_2 is connected.

▷ **6.4.5:** Find a sequence in the interval $[0, 1)$ that does not converge to a point of $[0, 1)$. (This will show that this set is not *compact*.)

▷ **6.4.6:** Prove that the Euler characteristic of a torus is 0. Hint: You can get a torus from a square by identifying opposite sides. (This observation came handy when you solved Exercise 2.4.7.) Thus you just need to triangulate a square, and keep track of vertices and edges through the identification.

do not distribute

do not distribute

Appendix — Solutions to selected problems

This appendix includes solutions to about half of the exercises listed at the end of each section; among these are the fifty or so problems directly mentioned within the text. In a sense, these form an integral part of the text, and it is fair that the reader should have access to them. However, I strongly recommend that you try to work them out before looking up my version—this will provide you with good practice and with the opportunity to develop a fresh viewpoint on the subjects covered here.

Remember that it is normal to experience a certain amount of frustration when solving exercises, particularly in a field that may be new to us. Learning to overcome such frustration on one's own is a necessary rite of passage, and the satisfaction of doing so is priceless. Glancing at a solution beforehand may occasionally save you some time, but it deprives you of one of the main benefits these notes can offer. In such cases I recommend that at the very least you try to find good examples illustrating the results of the exercises.

Chapter 1

1.1.1: Assume that $P(x)$ is some statement whose truth depends on x. Which of the following is 'grammatically correct'?

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

Solution: To fix ideas, assume that $P(x)$ is the statement 'x is a rational number'. The first sentence would read *"x is a rational number, for all"* and it sounds pleasantly zen, but I don't know what it could possibly mean. The dangling 'for all', by itself, means nothing. The third one says *"For all x in x is a rational number"*, and again I don't know what this means: 'x is a rational number' is a statement, not a set. The second one, *"For all x, x is a rational number"*, makes sense. It may be true or false, depending on the context (we don't know what x denotes here), but its meaning is clear. ⌟

1.1.2: Let S be an indexed set of real numbers:

$$S = \{s_n \,|\, n \in \mathbb{N}\} \quad .$$

Consider the statement

$$\forall x \in \mathbb{R}\,, \exists n \in \mathbb{N} : \quad s_n > x \quad .$$

- Find three examples of indexed sets S for which this statement is true, and three examples for which it is false.

do not distribute

- Find an adequate 'translation in English' of what this statement is saying.

Solution: The statement is true for $s_n = n$, $s_n = \sqrt{n}$, $s_n = n^2$, and many other analogous choices. The statement is false for $s_n = 1$, $s_n = 1/(n+1)$, $s_n = \sin n$ and many other choices. For example, to see that $s_n = \sqrt{n}$ satisfies the statement, note that $\forall x \in \mathbb{R}$ there exists some integer n for which $n > x^2$ (just take the 'next integer'): thus $\exists n \in N$ such that $(\sqrt{n})^2 > x^2$, and it follows that $\exists n \in N$ such that $\sqrt{n} > x$.

To see that setting $s_n = 1/(n+1)$ gives an indexed set for which the statement is false, keep in mind that the statement says that *for every possible* x, the statement '$\exists n \in \mathbb{N} : s_n > x$' is true. So to verify that the whole assertion is false I just need to find *one* number x for which the statement '$\exists n \in \mathbb{N} : s_n > x$' is false. Well, $x = 1$ is such a number x, because the statement

$$\exists n \in \mathbb{N} : \quad \frac{1}{n+1} > 1$$

is false. Indeed, if n is a natural number, then $n \geq 0$, so $n+1 \geq 1$, so $\frac{1}{n+1} \leq 1$, so $\frac{1}{n+1} \not> 1$. Or, even more 'concretely': the numbers $\frac{1}{n+1}$ for $n \in \mathbb{N}$ are $1, \frac{1}{2}, \frac{1}{3}, \frac{1}{4}, \ldots$: *none* of this numbers is > 1, so it is *not true* that 'there exists one of them' that is > 1.

As for a 'translation in English', the statement is telling us that *'The set S contains arbitrarily large numbers'*. ⌟

Remark: This statement is similar to the official statement defining '$\lim_{n\to\infty} s_n = \infty$', but it is not quite the same: the definition of limit would have one more quantifier. For example, I could let $s_n = n$ for n even, and $s_n = 1$ for n odd: $0, 1, 2, 1, 4, 1, 6, 1, 8, 1, \ldots$. This would satisfy the statement, but its limit would not be ∞, because s_n keeps coming back to 1.

1.2.2: Translate into a logical expression the adage *'All that glitters is not gold'*, assuming that $P(x)$ means *'x glitters'*, and $Q(x)$ means *'x is gold'*. (Watch out! this is trickier than it looks. Make sure that your translation matches the *meaning* of the sentence, not just its apparent structure.)

Solution: What *'All that glitters is not gold'* means is that *there are* things that glitter as gold would, but they turn out to be of no value. The meaning involves an existential quantifier, so the translation is

$$(\exists x) : P(x) \wedge \neg Q(x) \quad . \tag{A.1}$$

Note that taking the sentence too literally would lead to something like $(\forall x), P(x) \implies \neg Q(x)$, or a variation on the same theme. This says "*All* the things that glitter turn out not to be made of gold". A sensible statement, but not what the adage means. ⌟

The construction used in this sentence is peculiar to English. The 'not' in *'All that glitters is not gold'* does not negate 'is', as it would in an ordinary sentence. It negates the whole sentence: it means that 'It is not true that all that glitters is gold'. You can take this as another way to translate it into logic:

$$\neg(\,(\forall x) : P(x) \implies Q(x)\,) \quad , \tag{A.2}$$

another fine answer. If you want a follow-up exercise, understand that (A.1) and (A.2) are logically equivalent (as they must be, if they are both translations of the meaning of the same sentence!).

1.2.3: Consolidate the statement

$$(\exists a : p(a)) \vee (\exists b : q(b))$$

so as to use only one quantifier on only one variable. Can the statement $(\exists a : p(a)) \wedge (\exists b : q(b))$ be expressed with just one quantifier? What about $(\forall a : p(a)) \vee (\forall b : q(b))$? And $(\forall a : p(a)) \wedge (\forall b : q(b))$?

do not distribute

Solution: What the first statement claims is that either the property p is satisfied by some element, or the property q is satisfied by some element, or both possibilities occur. (Keep in mind that a and b are just labels! They may be labels for the same element, for what we know.) This may be expressed more concisely by

$$\exists a : (p(a) \vee q(a)) \quad .$$

On the other hand, $(\exists a : p(a)) \wedge (\exists b : q(b))$ means: p is satisfied by some element and q is satisfied by some element. This is not the same as saying that *the same* element may satisfy both p and q, which would be what $\exists a : (p(a) \wedge q(a))$ states. So this statement cannot be 'consolidated' in the same way.

Similarly, both quantifiers are needed for $(\forall a : p(a)) \vee (\forall b : q(b))$. This statement says that $p(a)$ is true for all a or $q(a)$ is true for all a (again, a and b are just labels, and there is no harm in replacing b by a since the scopes of the quantifiers do not overlap). A candidate replacement $\forall a : (p(a) \vee q(a))$ would mean that for every a, either $p(a)$ is true or $q(a)$ is true (or both). This is a weaker requirement: it would be true if for example p is true for half of the elements and q is true for the other half. These statements do not mean the same thing.

As to $(\forall a : p(a)) \wedge (\forall b : q(b))$, it does mean the same as $\forall a : (p(a) \wedge q(a))$: if I require $p(a)$ to be true for all a *and* $q(a)$ to be true for all a (again, there is no harm in replacing the label b by a), then I am requiring $p(a)$ and $q(a)$ to both be true for all a, and that is what $\forall a : (p(a) \wedge q(a))$ says. ⌟

1.2.8: In §1.2.5 we have seen how to get any truth table by OR-ing suitable ANDs. Explain how to get any truth table by AND-ing suitable ORs.

Solution: The strategy in that argument in §1.2.5 was to use ANDs to get columns in which all entries are F, except for one T in a desired place. Then, ORing these columns generates one with as many Ts as needed, in any place we want.

To get the same result by ANDing suitable ORs, use ORs to get columns in which all entries are T, except for one F as desired. (For example, $p \vee (\neg q) \vee (\neg r)$ is T for all choices of p, q, r except $p = F$, $q = T$, $r = T$.) ANDing the columns will produce a T when all the entries are T, and F when some of the entries are F. So we will be able to produce as many Fs as we need, in the places we need them. ⌟

For a concrete example, let's process the same truth table we worked out in §1.2.5:

p	q	r	??
T	T	T	F
T	T	F	T
T	F	T	T
T	F	F	F
F	T	T	F
F	T	F	T
F	F	T	F
F	F	F	F

I see F in the first, fourth, fifth, seventh, and eighth column. Here is how I produce five columns with exactly

do not distribute

one F in each of these places:

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

The conclusion is that my original truth table is expressed by

$$\big((\neg p) \vee (\neg q) \vee (\neg r)\big) \wedge \big((\neg p) \vee q \vee r\big) \wedge \big(p \vee (\neg q) \vee (\neg r)\big) \wedge \big(p \vee q \vee (\neg r)\big) \wedge \big(p \vee q \vee r\big) \quad .$$

1.2.9: Play with the NAND operation:

- Verify that $p \implies q$ is equivalent to p NAND $(q$ NAND $q)$.
- Show that every truth table can be obtained by exclusively using NAND operations. (Hint: By what we have seen, it suffices to write $\neg$, $\wedge$, $\vee$ by just using NANDs. Why?)
- Write $p \iff q$ by just using NANDs.

Solution: For the first point, note that q NAND q is equivalent to $\neg q$; so p NAND $(q$ NAND $q)$ is equivalent to $\neg(p \wedge (\neg q))$, which by De Morgan is equivalent to $(\neg p) \vee q$, hence to $p \implies q$ (by Lemma 1.2.1) as stated.

For the second point: we have already verified (in §1.2.5, and again in Exercise 1.2.8) that we can obtain every truth table by just using $\wedge$, $\vee$, and $\neg$. Therefore it is enough to obtain these three operations.

Recall that p NAND q is defined by $\neg(p \wedge q)$.

$\neg$) The operation $\neg$ acts on *one* statement: $\neg p$ is the negation of the statement p. This is obtained by NANDing p with itself: since $p \wedge p$ is equivalent to p, $\neg(p \wedge p)$ is equivalent $\neg p$. Therefore,

$$(\neg p) \quad \iff \quad (p \text{ NAND } p) \quad .$$

$\wedge$) The AND of p and q, $p \wedge q$, is the negation of $\neg(p \wedge q) = p$ NAND q. Encoding $\neg$ as we saw a moment ago, we get

$$p \wedge q \quad \iff \quad (p \text{ NAND } q) \text{ NAND } (p \text{ NAND } q) \quad .$$

$\vee$) By De Morgan's laws,

$$p \vee q \quad \iff \quad (\neg(\neg p)) \vee (\neg(\neg q)) \quad \iff \quad \neg((\neg p) \wedge (\neg q)) \quad .$$

This shows that

$$p \vee q \quad \iff \quad (\neg p) \text{ NAND } (\neg q) \quad .$$

Writing $\neg$ in terms of NAND as above, we get

$$p \vee q \quad \iff \quad (p \text{ NAND } p) \text{ NAND } (q \text{ NAND } q) \quad .$$

This concludes the verification.

do not distribute

For the last point, we could implement $(p \implies q) \wedge (q \implies p)$, and using the expression obtained in the first point and the considerations in the second, this could be done as follows:

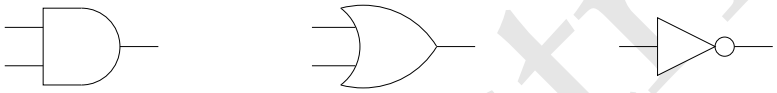
$$(p \text{ NAND } (q \text{ NAND } q)) \text{ NAND } (q \text{ NAND } (p \text{ NAND } p)) \\ \text{NAND } (p \text{ NAND } (q \text{ NAND } q)) \text{ NAND } (q \text{ NAND } (p \text{ NAND } p)) \quad .$$

This is not the most readable expression. You can find more concise alternatives, such as

$$(p \text{ NAND } q) \text{ NAND } ((p \text{ NAND } p) \text{ NAND } (q \text{ NAND } q)) \quad ;$$

this is still a little hard to parse, if you ask me. ⌟

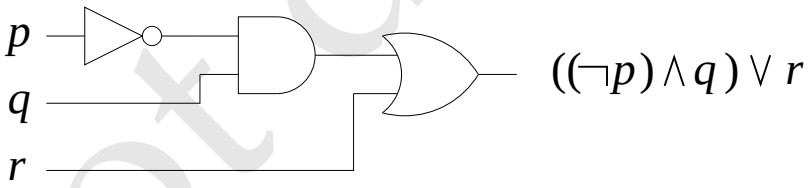
Additional, unnecessary but potentially amusing remarks: It used to be the case (and probably still is) that you could go to a store such as RadioShack and buy small chips with logic 'ports' on them, for the purpose of building electronic circuits which would encode a given logic operation. What you verified in this exercise is that *every* logic board can be constructed by just combining NAND ports together, and indeed NAND chips were the cheapest. On books and magazines, such logic circuits would be drawn with standard pictures standing for different ports:



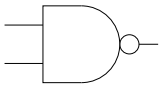
stand for AND, OR, NOT respectively; the legs on the left accept input statements p, q, and the corresponding $p \wedge q$, $p \vee q$, $\neg p$ would come out on the right. For example,

$$((\neg p) \wedge q) \vee r$$

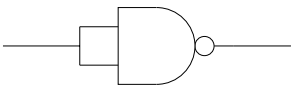
would be drawn



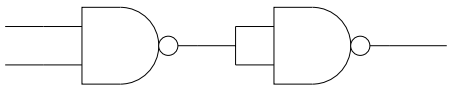
The NOT diagram is usually contracted to a small circle. Therefore, the NAND port is drawn simply



Using this graphical representation for NAND, what you verified in this exercise is that NOT is produced by

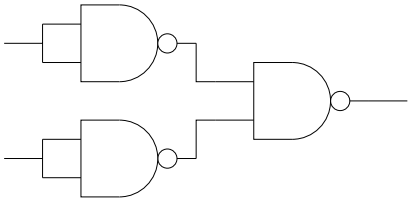


AND is



do not distribute

and OR is



1.3.2: The useful abbreviation $\forall x \in A,\ p(x)$ stands for *'the statement $p(x)$ is true for all x in the set A'*. In other words, *'for all x, if x is an element of A, then $p(x)$ is true'*. More formally, it is shorthand for

$$\forall x \quad (x \in A) \implies p(x) \quad .$$

What is the formal way to interpret the abbreviation $\exists x \in A, p(x)$?

Solution: $\exists x \in A, p(x)$ should be interpreted as

$$\exists x \quad (x \in A) \wedge p(x) \quad .$$

This should make sense without much further discussion: *there exists an x in the set A such that $p(x)$ is true* really means *there exists an x such that x is in A and $p(x)$ is true.* If this is not formal enough for you, remember that $\exists x, a(x)$ is equivalent to $\neg(\forall x, \neg a(x))$ (this was discussed in §1.3.1). By the same token, $\exists x \in A, p(x)$ is equivalent to $\neg(\forall x \in A, \neg p(x))$, or, formally,

$$\neg(\forall x, (x \in A) \implies \neg p(x)) \quad . \tag{A.3}$$

Now recall that $a \implies b$ is equivalent to $(\neg a) \vee b$; therefore, (A.3) is equivalent to

$$\neg(\forall x, (x \notin A) \vee \neg p(x))$$

therefore to

$$\exists x, \neg((x \notin A) \vee \neg p(x))$$

and finally to

$$\exists x, (x \in A) \wedge p(x)$$

by one of De Morgan's laws. ⌟

1.3.4: Are

$$\forall x > 0\ \exists N : n > N \implies |s_n| < x$$

and

$$\forall x > 0\ \exists N : n > N \implies |s_n| < 3x$$

equivalent statements? If so, prove that they are. If not, find an example of an assignment s_n which satisfies one of the statements but not the other.

Solution: These statements are equivalent: an indexed set $\{s_n\}_{n \in \mathbb{N}}$ satisfies one if and only if it satisfies the other. To see this, it may be helpful to rename the variables x and N by y and M in one of the statements (of

do not distribute

course changing the name of a quantified variable does not change the meaning of the statement[1]). So I will compare

$$\forall x > 0 \; \exists N : n > N \implies |s_n| < x \tag{A.4}$$

with

$$\forall y > 0 \; \exists M : n > M \implies |s_n| < 3y \tag{A.5}$$

and prove that these two statements are equivalent.

Proof. Assume $\{s_n\}_{n\in N}$ satisfies (A.4); I will prove it satisfies (A.5). To prove it satisfies (A.5), let y be an arbitrary given number, and let $x = 3y$. Since $\{s_n\}_{n\in\mathbb{N}}$ satisfies (A.4), there exists an N such that $n > N \implies |s_n| < x$. Since $x = 3y$, this is saying that $n > N \implies |s_n| < 3y$. Therefore, choosing $M = N$ shows that $\exists M : n > M \implies |s_n| < 3y$, verifying (A.5).

Conversely, assume $\{s_n\}_{n\in N}$ satisfies (A.5); I will prove it satisfies (A.4). To prove it satisfies (A.4), let x be an arbitrary given number, and let $y = x/3$. Since $\{s_n\}_{n\in\mathbb{N}}$ satisfies (A.5), there exists an M such that $n > M \implies |s_n| < 3y$. Since $y = x/3$, this is saying that $n > M \implies |s_n| < x$. Therefore, choosing $N = M$ shows that $\exists N : n > N \implies |s_n| < x$, verifying (A.4). □

Additional comments: I have the feeling that this is either completely clear to you, or completely mysterious. The potential for confusion I see is in the handling of the quantifiers. In the first paragraph of the proof, I want to prove that $\forall y$ something happens, so I start with an 'arbitrary given y'; but then in the same paragraph I am using that $\forall x \ldots$, and I write 'let $x = 3y$'. How come in one case I am interpreting $\forall y$ as 'y is an arbitrary number', and in the other I am interpreting $\forall x$ as 'let x be this particular number'?

This is a sensible objection, but it evaporates if you think about what we are really doing. In the first case I am supposed to *prove* that something happens *for all y:* my costumer comes in with his or her favorite y, and I have to guarantee that the fact will be true for that y. I cannot choose y, and I cannot make any assumption about it. In this sense it is 'given' and 'arbitrary'. In the second case, I *know* that something happens *for all x* (because that is my current assumption). Therefore I can choose whatever x I please, and I will know that the corresponding statement is true.

So the difference is that I am in the process of proving the statement $\forall y \ldots$ (so I can't choose y), while I already know that the statement $\forall x \ldots$ is true, so I can feed it with any x I choose.

1.3.5: Explain why the truth table for $\implies$ tells us that if p is a statement that implies a false statement, then p is itself false.
smallskip

Solution: The truth table for $p \implies q$ is

p	q	$p \implies q$
T	T	T
T	F	F
F	T	T
F	F	T

[1] I hope that this is clear. Saying *for all i from* 1 *to* 3, $p(i)$ *is true'* means: $p(1)$ is true and $p(2)$ is true and $p(3)$ is true. Saying *for all j from* 1 *to* 3, $p(j)$ *is true'* also means: $p(1)$ is true and $p(2)$ is true and $p(3)$ is true. Calling the variable i or calling it j makes no difference.

do not distribute

The part of the table for which q is F is

p	q	$p \implies q$
T	F	F
F	F	T

and we see that if p implies F, that is, if $p \implies F$ is T, then p must be F. ⌟

1.3.6: Prove 'directly' that

$$(\forall x \in \mathbb{R}\,,\, x^2 \in S) \implies S \text{ contains every nonnegative real number.}$$

(Cf. Example 1.3.2.)

Solution: We have to prove that if r is a nonnegative real number, then $r \in S$. Given r, let $s = \sqrt{r} \in \mathbb{R}$; this exists since square roots of nonnegative real numbers exist as real numbers. By hypothesis, $x^2 \in S$ for every real number x. Applying this fact to $x = s$ proves that $s^2 \in S$, and that means that $r \in S$ since $r = s^2$.

This verifies that $r \in S$, as needed. ⌟

1.3.7: Give a 'proof by contrapositive' of the following statement: *Let n be an integer; if n^2 is even, then n is even.*

Solution: The contrapositive of this statement is: *Let n be an integer; if n is odd, then n^2 is odd.* This is what we need to prove.

If n is odd, then there exists an integer k such that $n = 2k + 1$. Then

$$n^2 = (2k+1)^2 = 4k^2 + 4k + 1 = 2(2k^2 + 2k) + 1 \quad .$$

This proves that $n^2 = 2\ell + 1$, where ℓ is the integer $2k^2 + 2k$. It follows that n^2 is odd, as needed. ⌟

Chapter 2

2.1.1: Prove that if A, B are both subsets of C, then $A \cup B$ is a subset of C.

Solution: Isn't this obvious? If it seems obvious to you, that means that your mental image for what is going on is accurate. But you should be able to justify *why* it is accurate, and then you will probably produce an argument as follows. Let p, q, r denote the statements $x \in A$, $x \in B$, $x \in C$, respectively. Then $x \in A \cap B \iff (p \wedge q)$, so $A \cup B \subseteq C$ is equivalent to

$$(p \wedge q) \implies r \quad .$$

With this in mind, the statement becomes

$$\big((p \implies r) \wedge (q \implies r)\big) \implies \big((p \wedge q) \implies r\big) \quad ,$$

do not distribute

which may be verified by a truth table:

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

The logical connective $\implies$ between the last two column gives T on each row, as needed. ⌟

2.1.2: Prove that if A and B are sets, then $(A \smallsetminus B) \cup (A \cap B) = A$.

Solution: Both $A \smallsetminus B$ and $A \cap B$ are subsets of A, so $(A \smallsetminus B) \cup (A \cap B) \subseteq A$ (by Exercise 2.1.1). Therefore, it suffices to verify that $A \subseteq (A \smallsetminus B) \cup (A \cap B)$. For this, let $x \in A$. There are two possibilities: either $x \notin B$ or $x \in B$. In the first case we have $(x \in A) \wedge (x \notin B)$, therefore $x \in A \smallsetminus B$, and hence $x \in (A \smallsetminus B) \cup (A \cap B)$. In the second case we have $(x \in A) \wedge (x \in B)$, therefore $x \in A \cap B$, with the same conclusion.

Therefore $x \in A$ implies $x \in (A \smallsetminus B) \cup (A \cap B)$, as needed. ⌟

2.1.7: State and prove versions of parts (7) and (8) of Theorem 2.1.11 for arbitrary families of sets.

Solution: Part (7) states

$$S \smallsetminus (A \cap B) = (S \smallsetminus A) \cup (S \smallsetminus B)$$

for subsets A, B of a set S. In words: The complement of the intersection is the union of complements. The corresponding statement for a family $\{A_i\}_{i \in I}$ of subsets of S is

$$S \smallsetminus \bigcap_{i \in I} A_i = \bigcup_{i \in I} (S \smallsetminus A_i) \quad .$$

Similarly, the statement for families corresponding to part (8) is

$$S \smallsetminus \bigcup_{i \in I} A_i = \bigcap_{i \in I} (S \smallsetminus A_i) \quad .$$

The first statement is verified by the following equivalences:

$$\begin{aligned}
S \smallsetminus \bigcap_{i \in I} A_i &\iff (x \in S) \wedge (x \notin \bigcap_{i \in I} A_i) \\
&\iff (x \in S) \wedge \neg (x \in \bigcap_{i \in I} A_i) \\
&\iff (x \in S) \wedge \neg (\forall i \in I, x \in A_i) \\
&\iff (x \in S) \wedge (\exists i \in I, x \notin A_i) \\
&\iff \exists i \in I, x \in (S \smallsetminus A_i) \\
&\iff x \in \bigcup_{i \in I} (S \smallsetminus A_i) \quad .
\end{aligned}$$

do not distribute

The verification of the second statement is entirely analogous:

$$\begin{aligned} S \smallsetminus \bigcup_{i\in I} A_i &\iff (x \in S) \wedge (x \notin \bigcup_{i\in I} A_i) \\ &\iff (x \in S) \wedge \neg(x \in \bigcup_{i\in I} A_i) \\ &\iff (x \in S) \wedge \neg(\exists i \in I, x \in A_i) \\ &\iff (x \in S) \wedge (\forall i \in I, x \notin A_i) \\ &\iff \forall i \in I, x \in (S \smallsetminus A_i) \\ &\iff x \in \bigcap_{i\in I} (S \smallsetminus A_i) \quad . \end{aligned}$$

These assertion are the set-theoretic counterpart of the fact that negation 'switches quantifiers'. ⌟

2.1.8: What is $\bigcap_{n\in\mathbb{N}}[n,\infty)$, and why? What about $\bigcap_{a\in(0,1)}(-a,a)$?

Solution: $\bigcap_{n\in\mathbb{N}}[n,\infty) = \emptyset$. Indeed, suppose $r \in \bigcap_{n\in\mathbb{N}}[n,\infty)$; by definition of $\bigcap_{n\in\mathbb{N}}$, this means that

$$\forall n \in \mathbb{N}: \quad r \in [n,\infty) \quad ;$$

and since $[n,\infty)$ consists of all numbers $\geq n$, this means

$$\forall n \in \mathbb{N}: \quad r \geq n \quad .$$

But this would say that the fixed number r is larger than every integer, and that is a problem: there is no such number $r \in \mathbb{R}$. Therefore the intersection is $\emptyset$. ⌟

Note: If you are tempted to guess that ∞ may be in the common intersection, think again: ∞ is not a number. (And in any case the intervals $[n,\infty)$ formally exclude ∞.)

$\bigcap_{a\in(0,1)}(-a,a) = \{0\}$. Indeed, 0 is in every interval $(-a,a)$ for $a \neq 0$, so $\{0\} \subseteq \bigcap_{a\in(0,1)}(-a,a)$. To verify the other inclusion, argue by contradiction and assume $r \neq 0$ is in this intersection. By definition of $\bigcap_{a\in(0,1)}$, this means that

$$\forall a \in (0,1): \quad -a < r < a \quad ,$$

or equivalently

$$\forall a \in (0,1): \quad |r| < a \quad .$$

This would make $|r|$ a positive number that is smaller than any other positive number, which is a contradiction: there is no such r. Thus the intersection does consist of $\{0\}$ and nothing but $\{0\}$. ⌟ **2.1.9:** Prove (2.5) and (2.6): Let $\{S_i\}_{i\in I}$, $\{S_j\}_{j\in J}$ be two families of sets, and prove that

$$\begin{aligned} (\bigcap_{i\in I} S_i) \cap (\bigcap_{j\in J} S_j) &= \bigcap_{k\in I\cup J} S_k \\ (\bigcup_{i\in I} S_i) \cup (\bigcup_{j\in J} S_j) &= \bigcup_{k\in I\cup J} S_k \quad . \end{aligned}$$

Solution: By definition, $x \in \cap_{k\in I\cup J} S_k$ means

$$\forall k \in I \cup J: \quad x \in S_k \quad .$$

do not distribute

By Exercise 1.3.2, this is shorthand for

$$\forall k: \quad ((k \in I \cup j) \implies (x \in S_k)) \quad .$$

On the other hand, $k \in I \cup J$ means $(k \in I) \vee (k \in J)$. Therefore, the statement $(k \in I \cup j) \implies (x \in S_k)$ is equivalent to

$$((k \in I) \vee (k \in J)) \implies (x \in S_k) \quad :$$

it is of the form $(p \vee q) \implies r$. By Exercise 1.2.6

$$((p \vee q) \implies r) \iff ((p \implies r) \wedge (q \implies r)) \quad .$$

Therefore, $x \in \cap_{k \in I \cup J} S_k$ is equivalent to

$$\forall k: \quad (((k \in I) \implies (x \in S_k)) \wedge ((k \in J) \implies (x \in S_k)))$$

and therefore to

$$(\forall i: (i \in I) \implies (x \in S_i)) \wedge (\forall j: (j \in J) \implies (x \in S_j))$$

(you have thought through this when you solved Exercise 1.2.3; keep in mind that i, j, k are just labels!), that is to

$$(x \in \bigcap_{i \in I} S_i) \wedge (x \in \bigcap_{j \in J} S_j) \quad ,$$

and finally to

$$x \in (\bigcap_{j \in I} S_i) \cap (\bigcap_{j \in J} S_j) \quad .$$

This verifies the first formula.

The same technique of carefully parsing notation will prove the second formula. Explicitly, $x \in \cup_{k \in I \cup J} S_k$ means

$$\exists k \in I \cup J: \quad x \in S_k \quad .$$

By Exercise 1.3.2, this is shorthand for

$$\exists k: \quad ((k \in I \cup j) \wedge (x \in S_k)) \quad .$$

Again, $k \in I \cup J$ means $(k \in I) \vee (k \in J)$. Therefore, the statement $(k \in I \cup j) \wedge (x \in S_k)$ is equivalent to

$$((k \in I) \vee (k \in J)) \wedge (x \in S_k) \quad :$$

it is of the form $(p \vee q) \wedge r$. By 'distributivity' (Exercise 1.2.5)

$$((p \vee q) \wedge r) \iff ((p \wedge r) \vee (q \wedge r)) \quad .$$

Therefore, $x \in \cup_{k \in I \cup J} S_k$ is equivalent to

$$\exists k: \quad (((k \in I) \wedge (x \in S_k)) \vee ((k \in J) \wedge (x \in S_k)))$$

and therefore to

$$(\exists i: (i \in I) \wedge (x \in S_i)) \vee (\exists j: (j \in J) \wedge (x \in S_j)) \quad ,$$

do not distribute

(Exercise 1.2.3), that is to

$$(x \in \bigcup_{i\in I} S_i) \vee (x \in \bigcup_{j\in J} S_j) \quad ,$$

and finally to

$$x \in (\bigcup_{j\in I} S_i) \cup (\bigcup_{j\in J} S_j) \quad ,$$

proving the second formula. ⌟

2.1.12: Locate a discussion of *Russell's paradox*, and understand it.

One may be found at
`http://en.wikipedia.org/wiki/Russell's_paradox`

2.1.14: Let $S = \{s_1, \ldots, s_n\}$ be a finite set, and let $S' = \{s_1, \ldots, s_{n-1}\}$. Explain in what sense $\mathcal{P}(S')$ is a subset of $\mathcal{P}(S)$. Describe the complement $\mathcal{P}(S) \setminus \mathcal{P}(S')$. How large is this set, compared to $\mathcal{P}(S')$?

Solution: Elements of $\mathcal{P}(S')$ are subsets of S'. If A is a subset of S', then it may be viewed as a subset of S, since $S' \subseteq S$. In this sense every element of $\mathcal{P}(S')$ (i.e., every subset of S') is an element of $\mathcal{P}(S)$ (i.e., a subset of S). Thus $\mathcal{P}(S')$ is a subset of $\mathcal{P}(S)$.

The elements of $\mathcal{P}(S')$ consist of those subsets $A \subseteq S$ which only contain elements from $s_1, \ldots, s_{n-1}$. Therefore, the complement $\mathcal{P}(S) \setminus \mathcal{P}(S')$ consists of those subsets $A \subseteq S$ which contain the element s_n.

Every $A \subset S'$ determines one element $A \cup \{s_n\}$ of $\mathcal{P}(S) \setminus \mathcal{P}(S')$, and conversely every element $\mathcal{P}(S) \setminus \mathcal{P}(S')$ is obtained in this fashion. Therefore $\mathcal{P}(S) \setminus \mathcal{P}(S')$ contains precisely the same number of elements as $\mathcal{P}(S')$. ⌟

2.1.15: Let S be a finite set with $|S| = n$ elements. Prove that $|\mathcal{P}(S)| = 2^n$.

Solution: Choosing a subset A of S amounts to prescribing, for all n elements s of S, whether $s \in A$ or $s \notin A$. The hint was just meant to make you realize that every subset is determined by making a choice between two options for every one of the n elements of S. Altogether there are 2^n ways to do this. ⌟

2.2.4: Let $\sim_A$ be a relation on a set A, and let B be a subset of A. Define a relation $\sim_B$ on B using the information of $\sim_A$. (There is only one reasonable way to do this! We then say that $\sim_B$ is 'induced' by $\sim_A$.) If $\sim_A$ is reflexive, is $\sim_B$ necessarily reflexive? (Why/Why not?) Study the same question for the other standard properties of relations.

Solution: Given $b_1, b_2 \in B$, we note that b_1, b_2 also belong to A (because B is a subset of A). We can then set $b_1 \sim_B b_2$ to *mean* $b_1 \sim_A b_2$. This is the only way I can see to use the information carried by $\sim_A$ and B. Another viewpoint: after all, relations are subsets of cartesian products. So $\sim_B$ must be a subset of $B \times B$, and we are given $\sim_A$, a subset of $A \times A$. The only reasonable way to get a subset of $B \times B$ in this situation is to intersect it with $\sim_A$ within $A \times A$. This again defines the same $\sim_B$ as described above.

All the properties such as reflexivity, symmetry, etc. are inherited by $\sim_B$ if they are satisfied by $\sim_A$. For example, if $\sim_A$ is symmetric, then $\forall a_1, a_2 \in A$, $a_1 \sim_A a_2 \implies a_2 \sim_A a_1$. In particular, $\forall b_1, b_2 \in B$, $b_1 \sim_A b_2 \implies b_2 \sim_A b_1$. By the way we defined $\sim_B$, this says

$$\forall b_1, b_2 \in B, \quad b_1 \sim_B b_2 \implies b_2 \sim_B b_1$$

and as a consequence $\sim_B$ turns out to be symmetric as well. The verification for all other properties is analogous. ⌟

do not distribute

Note: In my experience, students seem puzzled by this question. If open-ended questions like this puzzle you, my suggestion is to think up some situation with which you are familiar and that satisfies the template from which you are supposed to extrapolate. Here you need a set A, a relation on A, and a subset B of A. For example, take $A = \mathbb{R}$, with the rlation $\leq$, and consider the subset $B = \mathbb{Z}$. The question asks for a natural definition of a relation on $\mathbb{Z}$ in this situation. I do believe that it is natural to consider the "same" relation $\leq$, but as a relation *on the subset* $\mathbb{Z}$. This means declaring that $a, b \in \mathbb{Z}$ are in relation precisely when $a \leq b$ once you view a and b as real numbers. Of course this reproduces the 'usual' relation $\leq$ on $\mathbb{Z}$ (and we then use $\leq$ as a name for this relation, whether we view it as a relation on $\mathbb{R}$ or on $\mathbb{Z}$, rather than writing $\leq_{\mathbb{R}}$, $\leq_{\mathbb{Z}}$.) This exercise calls for a general version of this type of examples.

2.2.7: Use the well-ordering principle to show that if $S \subseteq \mathbb{N}$ is a subset containing a given number $k \in \mathbb{N}$, and such that $n \in S \implies n+1 \in S$ for all $n \geq k$, then S contains all numbers $\geq k$.

Solution: We just need to adapt the proof of Proposition 2.2.11 to this apparently more general situation.

Let $T = \{n \in \mathbb{N} \mid n \geq k\} \smallsetminus S$ be the complement of S in the set of numbers $\geq k$. We have to prove that $T = \emptyset$: this will show that every integer $n \geq k$ is in S. Arguing by contradiction, assume that $T \neq \emptyset$. By the well-ordering principle, T has a least element $t \in T$.

By hypothesis, $k \in S$: so $k \notin T$, and this tells us $t > k$ since all elements of T are $\geq k$. But then we can consider the natural number $n = t - 1$. Since t is the least element in T, and $n < t$, we have $n \notin T$. On the other hand $t > k$, so $n \geq k$. By definition of T, if a number is larger than k and is not in T, then it is in S. Thus, $n \in S$. Now, for $n \geq k$, S satisfies the hypothesis

$$n \in S \implies n+1 \in S \quad .$$

Since $t = n + 1$, this would mean that $t \in S$. This is a contradiction, since $t \in T$ and t cannot belong both to S and to T.

Thus, the assumption $T \neq \emptyset$ leads to a contradiction. It follows that $T = \emptyset$, hence $S \supseteq \{n \in \mathbb{N} \mid n \geq k\}$, as needed. ⌟

2.2.13: Let $C(n,k)$ be the number of subsets with k elements of a set with n elements. Prove by induction that $C(n,k) = \binom{n}{k}$. (Hint: First prove that $C(n+1,k) = C(n,k-1) + C(n,k)$; then use this observation, Lemma 2.2.15, and induction to show that $C(n,k) = \binom{n}{k}$.)

Solution: Let's follow the hint, and first prove that $C(n+1,k) = C(n,k-1) + C(n,k)$. We don't need induction to see this; a simple counting argument suffices.

Proof. For all $n \geq 1$, I will denote by I_n the set $\{1, \ldots, n\}$. We have to count all subsets T with k elements in $I_{n+1} = \{1, \ldots, n+1\}$. If $k < 0$ or $k > n+1$, the identity $C(n+1,k) = C(n,k-1) + C(n,k)$ reads $0 = 0$ (and is therefore true); for $k = 0$, the identity says $1 = 0 + 1$ so this case is also true. Therefore, we may assume $1 \leq k \leq n+1$.

Among the $C(n+1,k)$ subsets with k elements in I_{n+1}, we can distinguish two types of subsets: those which contain $n+1$, and those which do not. The number of subsets of the first kind equals the number of subsets of I_n consisting of $k-1$ elements: these are the subsets obtained by removing $n+1$ from T. There are $C(n,k-1)$ such subsets, by definition.

As for subsets of the second kind, if T does not contain the element $n+1$ then $T \subseteq I_n$: there is one such subset for each subset of $\{1, \ldots, n\}$ with exactly k elements. The number of such subsets is $C(n,k)$, again by definition.

do not distribute

In total there are $C(n,k-1)+C(n,k)$ subsets, and this gives the statement. □

Still following the hint, let's use the fact we just proved and the analogous formula for $\binom{n}{k}$ proved in Lemma 2.2.15 to prove that

$$\binom{N}{k} = C(N,k)$$

for all integers k and $N \geq 0$. The proof will be by induction on N.

—For $N = 0$, the statement is that $\binom{0}{k} = C(0,k)$ for all k. Both sides equal 1 if $k = 0$ and 0 if $k \neq 0$, so this is true.

—Next, we have to show that if we assume $\binom{n}{k} = C(n,k)$ for a fixed n and all k, then we can deduce that $\binom{n+1}{k} = C(n+1,k)$ for the same n and all k. By Lemma 2.2.15,

$$\binom{n+1}{k} = \binom{n}{k-1} + \binom{n}{k} \quad ;$$

by the induction hypothesis,

$$\binom{n}{k-1} + \binom{n}{k} = C(n,k-1) + C(n,k) \quad ;$$

and by what we proved above,

$$C(n,k-1) + C(n,k) = C(n+1,k) \quad .$$

Putting together the last three displays gives $\binom{n+1}{k} = C(n+1,k)$, concluding the proof. ⌟

2.2.17: Put your calculus hat on for this exercise. Let f, g be real functions of one variables that can be differentiated infinitely many times, and denote by $f^{(k)}$ the k-th derivative of f. (For example, $f^{(0)} = f$, $f^{(1)} = f'$, $f^{(2)} = f''$.) Prove the following formula:

$$(fg)^{(n)} = \sum_{k=0}^{n} \binom{n}{k} f^{(k)} g^{(n-k)} \quad .$$

for all integers $n \geq 0$.

Solution: This formula is analogous to the binomial theorem, and its proof is similar.

—For $n = 0$, the statement is that $fg = fg$, which is true.

—To prove the induction step, we assume that $(fg)^{(n)} = \sum_{k=0}^{n} \binom{n}{k} f^{(k)} g^{(n-k)}$ for a fixed n, and we have to prove the analogous formula for the next derivative $(fg)^{(n+1)}$. Here is the key computation:

$$\begin{aligned} (fg)^{(n+1)} &= \left((fg)^{(n)}\right)' \\ &= \left(\sum_{k=0}^{n} \binom{n}{k} f^{(k)} g^{(n-k)}\right)' \end{aligned}$$

by the induction hypothesis,

$$= \sum_{k=0}^{n} \binom{n}{k} \left(f^{(k)} g^{(n-k)}\right)'$$

do not distribute

since differentiation is linear,

$$= \sum_{k=0}^{n} \binom{n}{k} \left(f^{(k+1)} g^{(n-k)} + f^{(k)} g^{(n-k+1)} \right)$$

by the usual product rule,

$$= \sum_{k=0}^{n} \binom{n}{k} f^{(k+1)} g^{(n-k)} + \sum_{k=0}^{n} \binom{n}{k} f^{(k)} g^{(n+1-k)}$$

splitting the summation,

$$= \sum_{k=0}^{n+1} \binom{n}{k-1} f^{(k)} g^{(n+1-k)} + \sum_{k=0}^{n+1} \binom{n}{k} f^{(k)} g^{(n+1-k)}$$

by precisely the same gymnastic as in the proof of the binomial theorem (first replace k by $\ell - 1$ in the first summation, then rename ℓ as k, while keeping track of terms that contribute 0 to the summation)

$$= \sum_{k=0}^{n+1} \left(\binom{n}{k-1} + \binom{n}{k} \right) f^{(k)} g^{(n+1-k)}$$

collecting the summations back together,

$$= \sum_{k=0}^{n+1} \binom{n+1}{k} f^{(k)} g^{(n+1-k)}$$

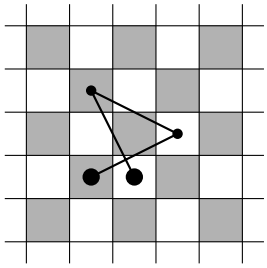
by Lemma 2.2.15.

This shows that $(fg)^{(n+1)} = \sum_{k=0}^{n+1} \binom{n+1}{k} f^{(k)} g^{(n+1-k)}$ as needed to complete the induction. ⌟

2.3.3: Say that two squares of a chessboard are 'knight-equivalent' if one can be reached from the other by a sequence of knight moves. How many equivalence classes are there?

Now say that two squares of a chessboard are 'bishop-equivalent' if one can be reached from the other by a sequence of bishop moves. How many equivalence classes are there?

Solution: Any two squares of a chessboard may be connected by a sequence of knight moves. Indeed, I can move from one square to an adjacent one (say, horizontally) as follows:



(The sequence may be modified to move to a vertically adjacent square, and must be adapted near the sides of the board.) Since I can move from one square to an adjacent one, I can move from any square to any other square. Therefore, all squares are knight-equivalent, and there is just one equivalence class.

do not distribute

As for bishop-equivalence, I can reach any white square from any white square and any black square from any black square by performing bishop moves, but I cannot connect squares of different colors. Therefore there are two equivalence classes, corresponding to the colors of the squares. ⌟

2.3.5: Let n be a fixed positive integer. Define a relation $\sim$ on $\mathbb{Z}$ by:

$$a \sim b \quad \iff \quad a \text{ and } b \text{ have the same remainder after division by } n.$$

Prove that for every given n, $\sim$ is an equivalence relation.

Solution: We have to verify that the relation $\sim$ is reflexive, symmetric, transitive.

- Reflexive: For every a, a has the same remainder as itself; therefore $a \sim a$.
- Symmetric: For all a, b, if a has the same remainder as b, then b has the same remainder as a. Therefore $a \sim b \implies b \sim a$.
- Transitive: For all a, b, c, if a has the same remainder as b, and b has the same remainder as c, then indeed a has the same remainder as c. This says that $a \sim b \wedge b \sim c$ implies $a \sim c$, as needed. ⌟

These three verifications suffice in order to prove that $\sim$ is an equivalence relation. If you are curious, contemplate the fact that two numbers a and b have the same remainder after division by n if and only if n divides the number $b - a$:

$$a \sim b \quad \iff \quad \exists k \in \mathbb{Z}\,, \quad b - a = nk \quad .$$

Reflexivity, symmetry, and transitivity correspond then to the following facts:
(R) 0 is an integer; (S) If k is an integer, then $-k$ is an integer; (T): if k and ℓ are integers, then $k + \ell$ are integers. Make sure you see why!

2.3.6: Describe explicitly the equivalence classes for the relation $\sim$ defined in Exercise 2.3.5. How many *distinct* equivalence classes are there? (This gives a description of the set $\mathbb{Z}/\sim$ in this case.)

Solution: The equivalence relation defined on $\mathbb{Z}$ in Exercise 2.3.5 is

$$a \sim b \quad \iff \quad a \text{ and } b \text{ have the same remainder after division by } n.$$

There are n possible remainders: $0, 1, \ldots, n-1$. The key observation here is that if r is one of these n numbers, then r equals the remainder of the division of itself by n. Thus, no two of these numbers are equivalent. On the other hand, if a has remainder r after division by n, then $a \sim r$, by the same token. Thus, every integer is equivalent to one of the n numbers $0, \ldots, n-1$.

The conclusion is that there are exactly n distinct equivalence classes modulo this equivalence relation. The quotient set $\mathbb{Z}/\sim$ is

$$\{[0]_\sim, \ldots, [n-1]_\sim\} \quad .$$

Here, $[r]_\sim$ consists of all integers which have remainder r after division by n. ⌟

If this is at all confusing, first look at our frequent example $n = 2$: there the equivalence classes turned out to be $[0]_\sim =$ the set of even numbers, and $[1]_\sim =$ the set of odd numbers (cf. Example 2.3.8). Understand why this exercise is a generalization of this running example. Then work it out again for some other small n. For example, for $n = 3$ the possible remainders are 0, 1, 2. The conclusion of the above discussion is that in

do not distribute

this case $\mathbb{Z}/\sim = \{[0]_\sim, [1]_\sim, [2]_\sim\}$. When $n = 3$, these classes are

$$[0]_\sim = \{\dots, -3, 0, 3, 6, 9, \dots\}$$
$$[1]_\sim = \{\dots, -2, 1, 4, 7, 10, \dots\}$$
$$[2]_\sim = \{\dots, -1, 2, 5, 8, 11, \dots\}$$

and it should be apparent that these sets do form a partition of $\mathbb{Z}$.

2.3.9: List all partitions of a set with 4 elements. How many partitions are there?

Solution: There are 15 partitions of a set with 4 elements. I will list them for the set $\{+, *, \circ, \bullet\}$:

$$\begin{aligned}
(1): &\quad \{\{+\},\{*\},\{\circ\},\{\bullet\}\}\\
(2): &\quad \{\{+,*\},\{\circ\},\{\bullet\}\}\\
(3): &\quad \{\{+,\circ\},\{*\},\{\bullet\}\}\\
(4): &\quad \{\{+,\bullet\},\{\circ\},\{*\}\}\\
(5): &\quad \{\{*,\circ\},\{+\},\{\bullet\}\}\\
(6): &\quad \{\{*,\bullet\},\{\circ\},\{+\}\}\\
(7): &\quad \{\{\circ,\bullet\},\{+\},\{*\}\}\\
(8): &\quad \{\{+,*\},\{\circ,\bullet\}\}\\
(9): &\quad \{\{+,\circ\},\{*,\bullet\}\}\\
(10): &\quad \{\{+,\bullet\},\{\circ,*\}\}\\
(11): &\quad \{\{+,*,\circ\},\{\bullet\}\}\\
(12): &\quad \{\{+,*,\bullet\},\{\circ\}\}\\
(13): &\quad \{\{+,\circ,\bullet\},\{*\}\}\\
(14): &\quad \{\{*,\circ,\bullet\},\{+\}\}\\
(15): &\quad \{\{+,*,\circ,\bullet\}\}
\end{aligned}$$

I am not doing this in any particularly intelligent way. The numbers of elements in each set of a partition must add up to the number of elements in the set, so I first wrote all possible ways to write 4 as the sum of positive numbers:

$$1+1+1+1 = 2+1+1 = 2+2 = 3+1 = 4$$

(incidentally, these are called *partitions* of the number). For each of these, I distributed the elements in all possible ways among sets of the prescribed size. For example, from $3+1$: choose the lone element to put in the set of size 1, and put the remaining 3 in the other. There are four ways to choose the lone element, so there are 4 partitions of this type. (They are (11) through (14) in my list.) ⌟

The wikipedia page on partitions:
`http://en.wikipedia.org/wiki/Partition_of_a_set`
is worth reading through quickly. It will also list the 15 partitions of a set with four elements, and arrange them in an interesting way. (But shame on you if you looked it up to find the solution to this problem!)

2.4.3: Let A be the set of *squares* in the plane. Define an equivalence relation on A by letting $S_1 \sim S_2$ if there is a rigid motion in the plane taking the square S_1 to the square S_2. Describe the quotient $A/\sim$.

do not distribute

Solution: Two squares are equivalent according to the relation defined in the statement if and only if they have the same area. Thus, there is exactly one equivalence class $[S]_\sim$ for every possible area; since the area of a square is a positive real number, we could describe $A/\sim$ as $\mathbb{R}^{>0}$, the set of positive real numbers. ⌟

2.4.5: Let A be the set of 'oriented segments' $\overrightarrow{PQ}$ in the plane. Define an equivalence relation $\sim$ on A by prescribing that $\overrightarrow{P_1Q_1} \sim \overrightarrow{P_2Q_2}$ if the segments are parallel, point in the same direction, and have the same length. (It should be clear that this is an equivalence relation.) Describe the quotient $A/\sim$.

Solution: Let O be the origin in the plane. Given any oriented segment $\overrightarrow{PQ}$, there is a unique oriented segment $\overrightarrow{OR}$ that is parallel to $\overrightarrow{PQ}$, points in the same direction, has the same length, and starts at the origin. This means that every equivalence class modulo $\sim$ has a unique representative of the form $\overrightarrow{OR}$. This representative determines and is determined by the coordinates (x, y) of the point R. Therefore, there is one equivalence class for every pair (x, y) of real numbers; we can take $\mathbb{R}^2$ as a description of the quotient $A/\sim$. ⌟

The astute reader will recognize that we are just talking about *vectors.* One way to define a vector is as 'the information of length and direction of an oriented segment'. If you have run into such a definition, and you have wondered what is meant by 'information', now you know: the information determined by $\overrightarrow{PQ}$ is its equivalence class $[\overrightarrow{PQ}]_\sim$, viewed as an element of the quotient set $A/\sim$ defined in this exercise.

2.4.7: (1) Define a relation $\sim'$ on $\mathbb{R}^2$ by prescribing

$$(x_1, y_1) \sim' (x_2, y_2) \quad \Longleftrightarrow \quad x_1 = x_2 \text{ and } y_2 - y_1 \in \mathbb{Z}.$$

(This is easily checked to be an equivalence relation.) Find a good geometric description of the quotient $\mathbb{R}^2/\sim_1$.

(2) Do the same for the equivalence relation $\sim''$ on $\mathbb{R}^2$ defined by

$$(x_1, y_1) \sim'' (x_2, y_2) \quad \Longleftrightarrow \quad x_2 - x_1 \in \mathbb{Z} \text{ and } y_2 - y_1 \in \mathbb{Z}.$$

Solution: This exercise was meant to let you stretch your imagination a little, after absorbing the example of the 'real line wrapping around a circle', Example 2.4.4 in the notes.

In Example 2.4.4, the challenge was to identify together all points $x \in \mathbb{R}$ with the same decimal part, and we settled on 'wrapping the real line around a circle': this has the effect of bringing together the points $\ldots 0.17, 1.17, 2.17, 3.17, \ldots$ (for example), which is what we were trying to do.

Just as in that example, in (1) the challenge is to identify together all points (x, y) such that y has a given decimal part, and now x is a fixed number a. So the points

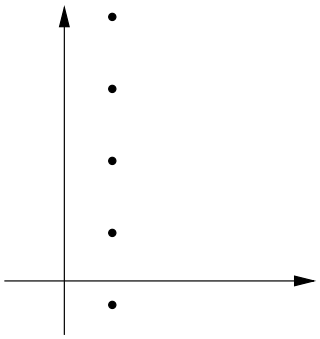
$$\ldots (0, 0.17), (0, 1.17), (0, 2.17), (0, 3.17), \ldots$$

should be brought together as an element of the quotient, since they form an equivalence class; the points

$$\ldots (0.1, 0.17), (0.1, 1.17), (0.1, 2.17), (0.1, 3.17), \ldots$$

also form an equivalence class, so they should also be brought together as another element of the quotient. Your typical equivalence class looks like this:

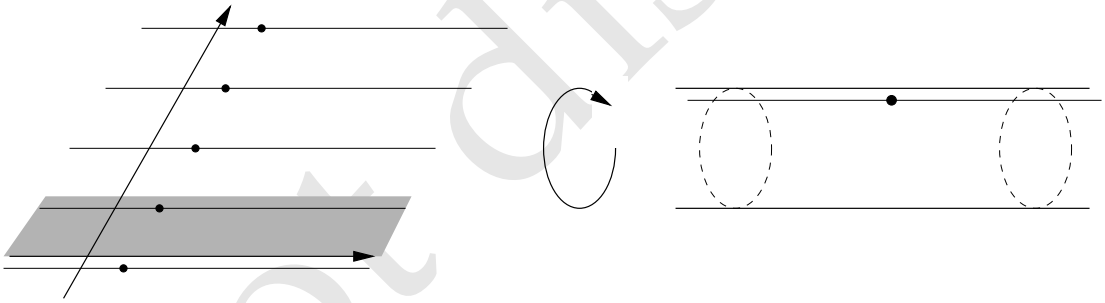
do not distribute



A moment's thought reveals that there is one equivalence class for each point in the strip

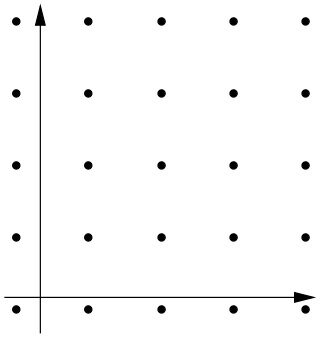
$$-\infty < x < \infty \quad , \quad 0 \leq y < 1 \quad :$$

This is already a description of the quotient, but it is not quite as good as it can be—it is the analog of the interval $[0,1)$ proposed as a first tentative description of the quotient in Example 2.4.4. The 'problem' is that the top side of the strip should really be identified with the bottom side, just as in Example 2.4.4 the two end-points of the interval should really be the same point. We solved that problem by wrapping the interval around a circle, and we can solve the current problem similarly, by wrapping the strip around a cylinder. Or (which is the same) we can simply wrap the whole plane around a cylinder:



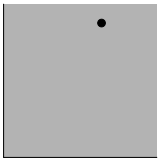
The standard description of the quotient in (1) is an infinite cylinder.

(2) is (even) more fun. A typical equivalence class looks like a grid:

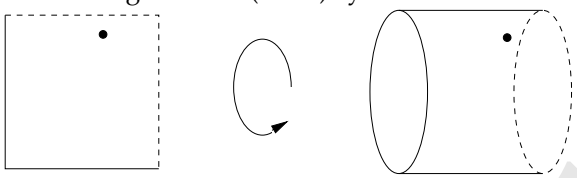


These are the points (x,y) with given fixed decimal parts for x *and* for y. Thus there is one equivalence class for every point in the square $\{(x,y) \mid 0 \leq x < 1 \wedge 0 \leq y < 1\}$:

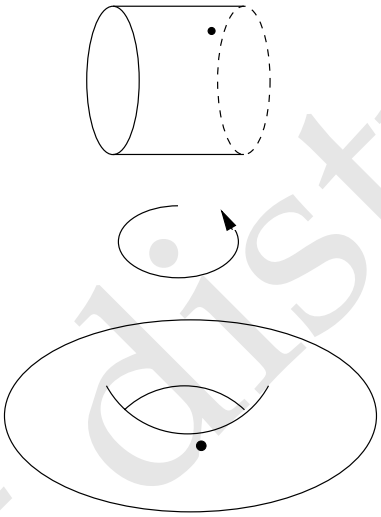
do not distribute



This is a preliminary description of the quotient, again analogous to the tentative $0 \leq x < 1$ in Example 2.4.4. To get a better description, we have to identify the bottom side to the top side of this square, and the left side to the right side. The first identification gives us a (finite) cylinder:



And then we still have to identify the left boundary of this cylinder with the right boundary. Doing this gives us the surface of a donut:



This is more properly called a *torus.* You should convince yourself (if you are not already convinced) that this double wrapping does identify together all the points in each equivalence class in $\mathbb{R}^2$ modulo $\sim''$, and that nearby 'grids' (equivalence classes) in $\mathbb{R}^2$ determine nearby points on the torus.

The standard description of the quotient in (2) is a torus. ⌟

Chapter 3

3.1.4: Let A denote a set $\{F, T\}$ with exactly two elements, and let S be any set. Prove that giving a function $S \to A$ (that is, an element of A^S) is 'the same as' giving a subset of S (that is, an element of $\mathcal{P}(S)$). In other words, explain how to define a subset of S for every function $S \to \{F, T\}$, and how to define a function $S \to \{F, T\}$ for every subset of S, in compatible ways. (This is the reason why $\mathcal{P}(S)$ is also denoted 2^S; cf. Definition 3.1.12. Here, '2' stands for a set with two elements.)

Solution: Given a function $f : S \to \{F, T\}$, let $X_f \subseteq S$ be the subset defined by

$$X_f = f^{-1}(T) = \{s \in S \mid f(s) = T\} \quad .$$

do not distribute

In this way, every function $S \to \{F, T\}$ determines a subset of S.

Conversely, given a subset $X \subseteq S$ of S, define a function $f_X : S \to \{F, T\}$ by prescribing $\forall s \in S$

$$f_X(s) = \begin{cases} T & \text{if } s \in X \\ F & \text{if } s \notin X \end{cases} .$$

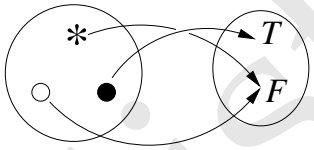
In this way, every subset of S determines a function $S \to \{F, T\}$.

These two assignments are compatible. Indeed, if you start from a function f, obtain a subset X_f from the first prescription, and then a function from X_f as in the second prescription, then this new function agrees with f. Conversely, starting from a subset $X \subseteq S$, obtaining a function f_X from the second prescription, and then a subset from f_X as in the first, then this new subset agrees with X. All these statements follow from the fact that

$$s \in X \quad \iff \quad f(s) = T$$

if $X = X_f$ or $f = f_X$, which is a consequence of the definitions given above. ⌟

All these words may obfuscate the idea, which is very simple. I have chosen to call T and F the elements of the target set because I am thinking of logic: the function f corresponds to the subset X if $f(s) = T$ when it is True that $s \in X$, and $f(s) = F$ when it is False that $s \in X$. For example, here is the set $\{*, \circ, \bullet\}$, and the function corresponding to the subset $\{\bullet\} \subseteq \{*, \circ, \bullet\}$:



If the above discussion confuses you at all, then draw the eight subsets of $\{*, \circ, \bullet\}$, and next to them the eight corresponding functions $\{*, \circ, \bullet\} \to \{F, T\}$, and all should become clear.

Also note that this is just a formalization of the hint given for Exercise 2.1.15, where I was suggesting to color an element 'Teal' if it was in the subset, and 'Fuchsia' if it wasn't. Of course that's the same as labeling an element T or F as I did in the solution to this problem. ⌟

3.1.5: Let $f : A \to B$ be a function. Prove that the set of nonempty inverse images of f forms a partition of A, and determine the corresponding equivalence relation.

Solution: I will give two proofs of the stated fact. The first proof is very direct. The second proof is instructive, since it emphasizes the connection with equivalence relations.

Proof I. Recall (Definition 2.3.10) that a partition of A is a family of distinct (i.e., different) subsets of A such that

- every element of the family is nonempty;
- every two different subsets in the family are disjoint; and
- the union of all the subsets in the family is A.

We have to verify these three properties for the family of nonempty inverse images of f. The first property is satisfied by assumption. As for the second, let b_1, b_2 be two elements in $\mathrm{im}(f)$, and consider the corresponding inverse images $f^{-1}(b_1)$, $f^{-1}(b_2)$. (I am choosing b_1, b_2 in $\mathrm{im}(f)$ to ensure that the inverse images are nonempty.) Arguing contrapositively, assume that $f^{-1}(b_1)$ and $f^{-1}(b_2)$ are *not* disjoint: then,

$$(\exists a \in A): \quad a \in f^{-1}(b_1) \cap f^{-1}(b_2) \quad ;$$

do not distribute

it follows that

$$b_1 = f(a) = b_2 \quad ,$$

and therefore that $f^{-1}(b_1) = f^{-1}(b_2)$ are not distinct, concluding the verification.

To verify the third property, note that each inverse image is a subset of A, so the union of the inverse images is a subset of A. To check the converse inclusion, let $a \in A$; then $a \in f^{-1}(f(a))$, and this shows that a belongs to a (nonempty) inverse image, and therefore to the union of the inverse images, completing the verification. □

Proof II. Let $\sim$ be the equivalence relation defined on A by prescribing $a_1 \sim a_2 \iff f(a_1) = f(a_2)$. This is an equivalence relation: $a \sim a$ for all $a \in A$ since $f(a) = f(a)$; $a \sim a'$ implies $a' \sim a$ since $f(a) = f(a')$ implies $f(a') = f(a)$; and similarly for transitivity. The equivalence class of a consists of all the elements $a' \in A$ such that $f(a') = f(a)$, and this is precisely the inverse image of the element $b = f(a) \in \mathrm{im}(f)$.

This shows that the nonempty inverse images of f are equivalence classes with respect to an equivalence relation. By Theorem 2.3.12, this implies that they form a partition of A, concluding the proof. □

3.2.3: Let $f : A \to B$ and $g : B \to C$ be two functions.

- Prove that if $g \circ f$ has a left inverse, then f has a left inverse.
- Prove that if $g \circ f$ has a right inverse, then g has a right inverse.

Solution: Assume $g \circ f$ has a left inverse h. Then

$$h \circ (g \circ f) = 1_A$$

(Definition 3.2.5). By associativity (Proposition 3.2.3), $h \circ (g \circ f) = (h \circ g) \circ f$. Thus, this shows that

$$(h \circ g) \circ f = 1_A \quad ,$$

which tells us that f has a left inverse: namely, $h \circ g$.

The argument for the second statement is entirely analogous. ⌟

3.2.4: Let $f : A \to B$ be a function, and assume that f has a left inverse g and a right inverse h. Prove that $g = h$. (Hint: Use Proposition 3.2.3.) Deduce that if f has a left and a right inverse, then it has a two-sided inverse.

Solution: Since g is a left inverse of f, then

$$g \circ f = 1_A \quad .$$

Since h is a right inverse of f, then

$$f \circ h = 1_B \quad .$$

(Cf. Definition 3.2.5.) From the first equality, and Proposition 3.2.4,

$$h = 1_A \circ h = (g \circ f) \circ h \quad .$$

By the second equality, and again Proposition 3.2.4,

$$g = g \circ 1_B = g \circ (f \circ h) \quad .$$

do not distribute

On the other hand, $g \circ (f \circ h) = (g \circ f) \circ h$ by Proposition 3.2.3. Putting all together,

$$g = g \circ (f \circ h) = (g \circ f) \circ h = h \quad ,$$

showing that $g = h$, as required.

This function $g = h$ is both a left and a right inverse for f, so f has a two-sided inverse in this case (Definition 3.2.5). ⌟

3.2.6: Let $f : A \to B$ and $g : B \to C$ be two isomorphisms of sets (cf. Definition 3.2.10). Prove that $g \circ f$ is also an isomorphism, by finding its two-sided inverse.

Solution: This follows from Exercise 3.2.5, showing that if f and g both have two-sided inverses, then so does $g \circ f$. We can also argue directly, as follows.

We can denote by f^{-1}, resp. g^{-1} the inverse of f, resp. g (Definition 3.2.10). Then I claim that $(f^{-1} \circ g^{-1})$ is a two-sided inverse for $g \circ f$. Indeed, repeatedly using associativity of composition of functions (Proposition 3.2.3):

$$\begin{aligned}(f^{-1} \circ g^{-1}) \circ (g \circ f) &= ((f^{-1} \circ g^{-1}) \circ g) \circ f \\ &= (f^{-1} \circ (g^{-1} \circ g)) \circ f \\ &= (f^{-1} \circ 1_B) \circ f \\ &= f^{-1} \circ f \\ &= 1_A\end{aligned}$$

and

$$\begin{aligned}(g \circ f) \circ (f^{-1} \circ g^{-1}) &= ((g \circ f) \circ f^{-1}) \circ g^{-1} \\ &= (g \circ (f \circ f^{-1})) \circ g^{-1} \\ &= (g \circ 1_A) \circ g^{-1} \\ &= g \circ g^{-1} \\ &= 1_C\end{aligned}$$

as needed. ⌟

3.3.1: • Suppose A and B are finite sets, and $|B| < |A|$. Prove that there are no injective functions $A \to B$. (This is called the 'pigeonhole principle': if you have n pigeons in $m < n$ holes, then necessarily at least one hole must be shared by more than one pigeon.)

• Suppose A and B are finite sets with the same number of elements, and that $f : A \to B$ is an injective function. Prove that f is also surjective.

• Now suppose that A and B are both *infinite* sets, and that $f : A \to B$ is an injective function. Is f necessarily surjective? Either give a proof or a counterexample.

Solution: • This is as obvious as it sounds. If $f : A \to B$ is a function between finite sets and $|B| < |A|$, then the images $f(a)$ as $a \in A$ cannot be all different. Then necessarily $\exists a_1, a_2 \in A$ such that $f(a_1) = f(a_2)$, that is, f is not injective.

• If A is a finite set with n elements and $f : A \to B$ is an injective function, then the image $\operatorname{im} f$ of f also consists of n elements. Indeed, there can't be more than n elements in $\operatorname{im} f$, since each element of $\operatorname{im} f$ is the

do not distribute

image of some element of A; and there can't be fewer than n, because otherwise we would get an injective function $A \to \operatorname{im} f$ even though $|\operatorname{im} f| < |A|$, contradicting the first point.

Thus, $\operatorname{im} f$ is a subset of B with n elements. If B also has n elements, as specified in this problem, then $\operatorname{im} f$ must be the whole of B. This says that every element of b is the image of some element of A, proving that f is surjective, as required. ⌟

• Injective functions between infinite sets are not necessarily surjective. For example, consider the function $f : \mathbb{Z} \to \mathbb{Z}$ defined by $f(n) = 2n$. This function is injective (because if $2n = 2m$, then $n = m$), but it is not surjective (because not all integers are even). ⌟

In time we will examine the notion of 'cardinality', and this will allow us to contemplate much stronger results. While the function $\mathbb{Z} \to \mathbb{Z}$ defined by $f(n) = 2n$ is injective but not surjective, there are injective functions $\mathbb{Z} \to \mathbb{Z}$ that are also surjective: for example, the identity is. On the other hand, *no* injective function $\mathbb{Z} \to \mathbb{R}$ can be surjective. This is not obvious; it is proved later on in these notes.

3.3.4: Let $f : A \to B$ and $g : B \to C$ be functions.

- Prove that if both f and g are injective, then the composition $g \circ f$ is also injective.
- Prove that if both f and g are surjective, then the composition $g \circ f$ is also surjective.

(Do this directly from the definitions, without appealing to Exercise 3.2.5.)

Solution: For the first point, let $a_1, a_2 \in A$, and assume that $g \circ f(a_1) = g \circ f(a_2)$. Then $g(f(a_1)) = g(f(a_2))$; since g is injective, it follows that $f(a_1) = f(a_2)$; and since f is injective we can deduce that $a_1 = a_2$. This proves that $g \circ f$ is injective.

For the second point, let $c \in C$ be any element. Since g is surjective, $\exists b \in B$ such that $g(b) = c$. Since f is surjective, $\exists a \in A$ such that $f(a) = b$. Then $g \circ f(a) = g(f(a)) = g(b) = c$. This proves that every $c \in C$ has a preimage by $g \circ f$ in A, that is, that $g \circ f$ is surjective. ⌟

3.3.8: We have seen (cf. Remark 3.2.9) that right- or left inverses are not unique in general. Discuss this fact, in light of the proof of Theorem 3.3.7: what makes the functions constructed in that proof 'not unique'?

Solution: The construction of the left inverse of an injective function relied on the arbitrary choice of an element $z \in A$, which is then given as the image by g of the elements of B that are not in $\operatorname{im} f$. Choosing a different z, or in fact choosing arbitrary elements of A to assign as images of elements not in $\operatorname{im} f$, will produce different functions g with the property that $g \circ f = 1_A$. Thus, the left inverse of an injective function is not unique if $\operatorname{im} f \neq B$, that is, if f is not surjective.

The construction of the right inverse of a surjective function also relied on a choice: we chose one preimage a_b for every element $b \in B$, and let $g(b) = a_b$. If an element b has other preimages, say $a'_b \neq a_b$, then choosing a'_b as the image of b by g will lead to a different right inverse. Thus, the right inverse of a surjective function is not unique if some element of b has more than one preimage, that is, if f is not injective. ⌟

3.4.1: Let $\mathcal{F}$ be a family of sets. Prove that the isomorphism relation $\cong$ defines an *equivalence* relation on $\mathcal{F}$. That is, show that (for all sets A, B, C in $\mathcal{F}$)

- $A \cong A$;
- If $A \cong B$, then $B \cong A$;
- If $A \cong B$ and $B \cong C$, then $A \cong C$.

do not distribute

Solution:

- $A \cong A$ because $1_A : A \to A$ is an isomorphism.
- $A \cong B \implies B \cong A$: if $f : A \to B$ is an isomorphism, then it has an inverse function $f^{-1} : B \to A$, which is also an isomorphism. (Indeed, f is its two-sided inverse.)
- If $A \cong B$ and $B \cong C$, then there are isomorphisms $f : A \to B$ and $g : B \to C$. The composition $g \circ f : A \to C$ is then an isomorphism (Exercise 3.2.6), proving that $A \cong C$. ⌟

3.4.5: List all the elements of the set of bijections of $A = \{a, b, c, d\}$ to itself, by indicating the image of each element. For example,

$$\begin{cases} a \mapsto a \\ b \mapsto c \\ c \mapsto b \\ d \mapsto d \end{cases}$$

denotes the bijection which swaps b and c.

- Find one element s that is not the identity and such that $s \circ s$ is the identity.
- Find one element t that is not the identity and such that $t \circ t \circ t$ is the identity.
- Find one element u that is not the identity, such that $u \circ u$ is not the identity, and such that $u \circ u \circ u \circ u$ is the identity.

Solution: There are 24 distinct bijections:

$$\begin{cases} a \mapsto a \\ b \mapsto b \\ c \mapsto c \\ d \mapsto d \end{cases}, \quad \begin{cases} a \mapsto a \\ b \mapsto b \\ c \mapsto d \\ d \mapsto c \end{cases}, \quad \begin{cases} a \mapsto a \\ b \mapsto c \\ c \mapsto b \\ d \mapsto d \end{cases}, \quad \begin{cases} a \mapsto a \\ b \mapsto c \\ c \mapsto d \\ d \mapsto b \end{cases}, \quad \begin{cases} a \mapsto a \\ b \mapsto d \\ c \mapsto b \\ d \mapsto c \end{cases}, \quad \begin{cases} a \mapsto a \\ b \mapsto d \\ c \mapsto c \\ d \mapsto b \end{cases}$$

$$\begin{cases} a \mapsto b \\ b \mapsto a \\ c \mapsto c \\ d \mapsto d \end{cases}, \quad \begin{cases} a \mapsto b \\ b \mapsto a \\ c \mapsto d \\ d \mapsto c \end{cases}, \quad \begin{cases} a \mapsto b \\ b \mapsto c \\ c \mapsto a \\ d \mapsto d \end{cases}, \quad \begin{cases} a \mapsto b \\ b \mapsto c \\ c \mapsto d \\ d \mapsto a \end{cases}, \quad \begin{cases} a \mapsto b \\ b \mapsto d \\ c \mapsto a \\ d \mapsto c \end{cases}, \quad \begin{cases} a \mapsto b \\ b \mapsto d \\ c \mapsto c \\ d \mapsto a \end{cases}$$

$$\begin{cases} a \mapsto c \\ b \mapsto a \\ c \mapsto b \\ d \mapsto d \end{cases}, \quad \begin{cases} a \mapsto c \\ b \mapsto a \\ c \mapsto d \\ d \mapsto b \end{cases}, \quad \begin{cases} a \mapsto c \\ b \mapsto b \\ c \mapsto a \\ d \mapsto d \end{cases}, \quad \begin{cases} a \mapsto c \\ b \mapsto b \\ c \mapsto d \\ d \mapsto a \end{cases}, \quad \begin{cases} a \mapsto c \\ b \mapsto d \\ c \mapsto a \\ d \mapsto b \end{cases}, \quad \begin{cases} a \mapsto c \\ b \mapsto d \\ c \mapsto b \\ d \mapsto a \end{cases}$$

$$\begin{cases} a \mapsto d \\ b \mapsto a \\ c \mapsto b \\ d \mapsto c \end{cases}, \quad \begin{cases} a \mapsto d \\ b \mapsto a \\ c \mapsto c \\ d \mapsto b \end{cases}, \quad \begin{cases} a \mapsto d \\ b \mapsto b \\ c \mapsto a \\ d \mapsto c \end{cases}, \quad \begin{cases} a \mapsto d \\ b \mapsto b \\ c \mapsto c \\ d \mapsto a \end{cases}, \quad \begin{cases} a \mapsto d \\ b \mapsto c \\ c \mapsto a \\ d \mapsto b \end{cases}, \quad \begin{cases} a \mapsto d \\ b \mapsto c \\ c \mapsto b \\ d \mapsto a \end{cases}$$

do not distribute

Writing them out gets old real soon unless you pause a moment and concoct some mechanism to do it systematically. In constructing this table I first sent a to a, listing the 6 permutations of b, c, d underneath; that gave me the first line; then I sent a to b, listing the 6 permutations of a, c, d, and that gave me the second line; and so on. It may have taken me five minutes to write the whole thing, and that isn't so bad, is it? (It would have taken me longer if I had written it by hand; you are hereby encouraged to start using a good typesetting system, such as LaTeX. In the end, that will save you time.)

An element s that is not the identity and such that $s \circ s$ is the identity:

$$\begin{cases} a \mapsto b \\ b \mapsto a \\ c \mapsto d \\ d \mapsto c \end{cases}$$

An element t that is not the identity and such that $t \circ t \circ t$ is the identity:

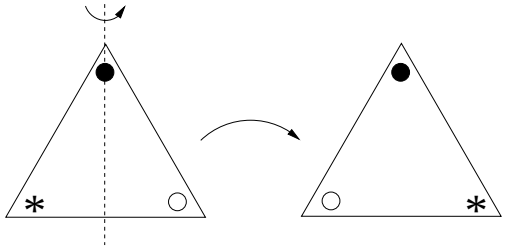
$$\begin{cases} a \mapsto b \\ b \mapsto c \\ c \mapsto a \\ d \mapsto d \end{cases}$$

An element u that is not the identity, such that $t \circ t$ is not the identity, and such that $u \circ u \circ u \circ u$ is the identity:

$$\begin{cases} a \mapsto b \\ b \mapsto c \\ c \mapsto d \\ d \mapsto a \end{cases}$$

Of course there are many other possible answers. ⌟

3.4.6: Make an equilateral triangle T out of cardboard (or imagine doing so!). Applying a (rigid) *symmetry* to T means taking your cardboard shape, moving it about in space as you wish, and then putting it back in its place, preserving its outline. Two symmetries are the same if the final positions of the triangle are the same. For example, leaving the triangle alone counts as one symmetry; flipping it about one of its axis counts as another one.



- Prove that there are exactly six distinct symmetries of an equilateral triangle.
- Find a one-to-one correspondence between the set of symmetries of an equilateral triangle and the set of bijections considered in Example 3.4.9.

do not distribute

In the second point, your one-to-one correspondence should preserve compositions, in the sense that performing one symmetry after the other should match the effect of composing the corresponding bijections.

Solution: Mark one vertex of the triangle. Any symmetry will send that one vertex to one of the three vertices, and either flip or not flip the triangle. This shows that there are $3 \cdot 2 = 6$ symmetries.

The illustrating picture was meant as a subliminal hint for the construction of the required one-to-one correspondence with the set of bijections of $\{*, \circ, \bullet\}$. Mark the corners of the triangle by $*$, $\circ$, $\bullet$; after you perform a symmetry, see where the symbols have landed, and that will give you a bijection from $\{*, \circ, \bullet\}$ to itself. For example, the symmetry in the figure corresponds to

$$\begin{cases} * \mapsto \circ \\ \circ \mapsto * \\ \bullet \mapsto \bullet \end{cases}$$

In this way, every symmetry determines a bijection of $\{*, \circ, \bullet\}$, and conversely every bijection of $\{*, \circ, \bullet\}$ determines a symmetry. Explicitly, the six bijection of $\{*, \circ, \bullet\}$ are listed in Example 3.4.9:

$$\begin{cases} * \mapsto * \\ \circ \mapsto \circ \\ \bullet \mapsto \bullet \end{cases} \quad \begin{cases} * \mapsto * \\ \circ \mapsto \bullet \\ \bullet \mapsto \circ \end{cases} \quad \begin{cases} * \mapsto \circ \\ \circ \mapsto * \\ \bullet \mapsto \bullet \end{cases} \quad \begin{cases} * \mapsto \circ \\ \circ \mapsto \bullet \\ \bullet \mapsto * \end{cases} \quad \begin{cases} * \mapsto \bullet \\ \circ \mapsto \circ \\ \bullet \mapsto * \end{cases} \quad \begin{cases} * \mapsto \bullet \\ \circ \mapsto * \\ \bullet \mapsto \circ \end{cases}$$

The corresponding symmetries are, in the same order:

1. the identity;
2. flip about the axis through $*$;
3. flip about the axis through $\bullet$;
4. 120° counterclockwise rotation;
5. flip about the axis through $\circ$;
6. 120° clockwise rotation.

This correspondence gives an isomorphism between the set of six bijections of $\{*, \circ, \bullet\}$ and the set of six symmetries of the equilateral triangle.

In fact, you can interpret the pictures drawn underneath the bijections in Example 3.4.9 as directions on how to perform the corresponding symmetry. For example, the fourth one is indeed a counterclockwise rotation by 120°. Since the bijections are just a way to represent the effect of the corresponding symmetries, it is clear that composing the bijections has the same effect as composing the corresponding symmetries. ⌟

3.5.2: Apply 'canonical decomposition' to the function $f : \mathbb{R} \times \mathbb{R} \to \mathbb{R}$ defined by $f(x, y) = x^2 + y^2$. That is: define carefully suitable sets S and T and a surjective function $\mathbb{R} \times \mathbb{R} \to S$, a bijective function $S \to T$, and an injective function $T \to \mathbb{R}$, so that the composition $\mathbb{R} \times \mathbb{R} \to S \to T \to \mathbb{R}$ is the function given above.

Solution: Define a relation $\sim$ on $\mathbb{R} \times \mathbb{R}$ by declaring that $(x_1, y_1) \sim (x_2, y_2) \iff x_1^2 + y_1^2 = x_2^2 + y_2^2$, and define S to be $(\mathbb{R} \times \mathbb{R})/\sim$. Define T to be the image of f, that is, $\mathbb{R}^{\geq 0}$. Then the decomposition of f is

$$\mathbb{R} \times \mathbb{R} \twoheadrightarrow S = (\mathbb{R} \times \mathbb{R})/\sim \longrightarrow T = \mathbb{R}^{\geq 0} \hookrightarrow \mathbb{R} \quad .$$

do not distribute

On elements: the first function maps (x, y) to its class modulo $\sim$; the function $S \to T$ maps a class $[(x, y)]_\sim$ to the nonnegative number $x^2 + y^2$; and the function $T \to \mathbb{R}$ maps this number to itself, viewed simply as a real number. ⌟

3.5.4: Let M_3 be the set of 3×3 matrices with integer entries. Consider the relation $\sim$ on M_3 defined by declaring that, $\forall A_1, A_2 \in M_3$, $A_1 \sim A_2$ if and only if $\det(A_1) = \det(A_2)$. Describe the quotient $M_3/\sim$, by using the technique explained in §3.5.2 (cf. Corollary 3.5.7).

Solution: The determinant gives a function $\det : M_3 \to \mathbb{Z}$, because the determinant of a matrix with integer entries is an integer. This function is surjective: for every $n \in \mathbb{Z}$,

$$n = \det \begin{pmatrix} n & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix} \quad .$$

By definition of $\sim$, $A_1 \sim A_2$ if and only if det has the same value on A_1 and A_2. Therefore $A/\sim \cong \mathbb{Z}$, by Corollary 3.5.7. The quotient is described by $\mathbb{Z}$. ⌟

3.5.7: Let U be any set containing $\mathbb{Z}$, and let $\mathcal{F}$ be the set of *finite* subsets of U. Consider the equivalence relation $\cong$ defined on $\mathcal{F}$ by the relation of isomorphism. (You checked that $\cong$ is an equivalence relation, in Exercise 3.4.1.) Describe the quotient $\mathcal{F}/\cong$.

Solution: $\mathcal{F}/\cong$ is isomorphic to the set $\mathbb{N} = \mathbb{Z}^{\geq 0}$ of nonnegative integers. Indeed, consider the function $f : \mathcal{F} \to \mathbb{Z}^{\geq 0}$ defined as follows: for a finite set $A \in \mathcal{F}$, let $f(A)$ be the number of elements in A. This function is surjective, since as U contains $\mathbb{Z}$, it contains at least one finite set with exactly n elements for any $n \geq 0$. (For $n = 0$, this set is $\emptyset$; for $n > 0$, we can take $\{1, \ldots, n\}$.) Since two finite sets are isomorphic if and only if they have the same number of elements (this is hopefully clear by now, and it was observed in §3.4.1),

$$A_1 \cong A_2 \quad \Longleftrightarrow \quad f(A_1) = f(A_2) \quad .$$

Corollary 3.5.7 implies that $\mathcal{F}/\cong$ is isomorphic to $\mathbb{Z}^{\geq 0}$, as claimed. ⌟

3.5.8: Consider the relation $\sim$ on $\mathbb{N} \times \mathbb{N}$ defined by setting

$$(a_1, b_1) \sim (a_2, b_2) \iff a_1 + b_2 = a_2 + b_1 \quad .$$

Describe the quotient $\mathbb{N} \times \mathbb{N}/\sim$.

Solution: Consider the function $f : \mathbb{N} \times \mathbb{N} \to \mathbb{Z}$ defined by $f(a, b) = a - b$. Then

$$f(a_1, b_1) = f(a_2, b_2) \iff a_1 - b_1 = a_2 - b_2 \iff a_1 + b_2 = a_2 + b_1 \iff (a_1, b_1) \sim (a_2, b_2) \quad .$$

To see that f is surjective, observe that if $n \in \mathbb{Z}$ is nonnegative, then $n = f(n, 0)$; if $n \in \mathbb{Z}$ is negative, say $n = -m$ with $m \in \mathbb{N}$, then $n = f(0, m)$. Thus $\operatorname{im} f = \mathbb{Z}$, and f is indeed surjective.

By Corollary 3.5.7, we can conclude that $(\mathbb{N} \times \mathbb{N})/\sim \cong \mathbb{Z}$. ⌟

This is the official way to construct the integers $\mathbb{Z}$ from $\mathbb{N}$. It is a particular case of the construction known as the 'Grothendieck group'. As mentioned in §3.4.3, $\mathbb{Z}$ is a *group* when considered with its ordinary operation of addition.

---

do not distribute

Chapter 4

4.1.2: Let $s > 1$ be an integer. Prove that if an integer n is not the s-th power of an integer, then $\sqrt[s]{n}$ is not a rational number. Deduce again that $\sqrt{2}$ is irrational, as well as e.g., the s-th roots $\sqrt[s]{p}$ of all prime integers p, for all $s > 1$. (Use Exercise 4.1.1.)

Solution: Saying that n is not the s-th power of an integer is equivalent to saying that the equation $x^s - n = 0$ has no integer roots. Since the polynomial $x^s - n$ is monic, this implies (by Exercise 4.1.1) that it has no *rational* roots, that is, that $\sqrt[s]{n}$ is not a rational number. ⌟

4.1.4: Prove that if a set A has a least upper bound, then the least upper bound is unique.

Solution: Assume that q, q' are both least upper bounds for A. Then $q \leq q'$, since q' is an upper bound and q is a least upper bound; and $q' \leq q$ since q is an upper bound and q' is a least upper bound. Since $q \leq q'$ and $q' \leq q$, we can conclude $q = q'$ since $\leq$ is an order relation and order relations are 'antisymmetric' (Definition 2.2.3). ⌟

4.2.1: Prove that the function $\mathbb{Q} \to \mathbb{D}$ defined in §4.2.1 is injective: that is, prove that if $q_1 \neq q_2$, then $\{x \in \mathbb{Q} \,|\, x < q_1\} \neq \{x \in \mathbb{Q} \,|\, x < q_2\}$.

Solution: Without loss of generality we may assume $q_1 < q_2$. Then $q_1 \notin \{x \in \mathbb{Q} \,|\, x < q_1\}$ (since $q_1 \not< q_1$), but $q_1 \in \{x \in \mathbb{Q} \,|\, x < q_2\}$ (since $q_1 < q_2$). Thus the sets differ, since they do not have the same elements. ⌟

4.2.2: Prove that the relation $\leq$ on $\mathbb{D}$ defined in Definition 4.2.5 is antisymmetric.

Solution: Assume $(A_1|B_1) \leq (A_2|B_2)$ and $(A_2|B_2) \leq (A_1|B_1)$. By definition, this means $A_1 \subseteq A_2$ and $A_2 \subseteq A_1$. Since $\subseteq$ is antisymmetric, this implies $A_1 = A_2$. It then follows that $B_2 = \mathbb{Q} \setminus A_2 = \mathbb{Q} \setminus A_1 = B_1$, proving that $(A_1|B_1) = (A_2|B_2)$. ⌟

4.2.3: Let $(A_1|B_1) < (A_2|B_2)$ be two Dedekind cuts. Prove that there exists a rational number q such that $(A_1|B_1) < q < (A_2|B_2)$.

Solution: By the definition of ordering in $\mathbb{D}$, $(A_1|B_1) < (A_2|B_2)$ implies that $A_1 \subsetneq A_2$. Let then $q \in A_2 \setminus A_1$. We may assume that q is not the minimum of B_1: indeed, if the minimum of B_1 were the only rational number in $A_2 \setminus A_1$, then it would be the maximum of A_2, while A_2 has no maximum by definition of Dedekind cut. Since $q \in B_1$ and q is not the minimum of B_1, A_1 is strictly contained in the set of rationals $< q$, so $(A_1|B_1) < q$. On the other hand we have $q \in A_2$. Since A_2 has no maximum, the set of rationals $< q$ is strictly contained in A_2; so $q < (A_2|B_2)$.

Therefore q satisfies the given requirement. ⌟

4.2.4: With notation as in the proof of Theorem 4.2.7, prove that $x < y$ for all $x \in L$ and all $y \in M$.

Solution: Since $x \in L$ (and by definition of union over families, Definition 2.1.8) there exists an index $j \in I$ such that $x \in A_j$. Since $y \in M$, $y \in B_i$ for all $i \in I$ (by definition of intersection over families), and in particular $y \in B_j$. Since $(A_j|B_j)$ is a Dedekind cut, it follows that $x < y$, as required. ⌟

4.2.5: Provide a sensible definition of the 'sum' $(A_1|B_1) + (A_2|B_2)$ of two arbitrary Dedekind cuts $(A_1|B_1)$, $(A_2|B_2)$. This should be done in such a way that if $q_1, q_2 \in \mathbb{Q}$ and $\iota(q_1) = (A_1|B_1)$, $\iota(q_2) = (A_2|B_2)$, then $(A_1|B_1) + (A_2|B_2) = \iota(q_1 + q_2)$.

do not distribute

Solution: Given two Dedekind cuts $(A_1|B_1)$, $(A_2|B_2)$, let

$$A = \{a \in \mathbb{Q} \,|\, \exists a_1 \in A_1, \exists a_2 \in A_2 : \, a = a_1 + a_2\} \quad .$$

(In words: A consists of the sums $a_1 + a_2$ of all the elements $a_1 \in A_1$ and $a_2 \in A_2$.) Let $B = \mathbb{Q} \setminus A$. Define

$$(A_1|B_1) + (A_2|B_2) := (A|B) \quad .$$

It is straightforward to verify that $(A|B)$ is a Dedekind cut. For example, if A had a maximum a, then we would have $a = a_1 + a_2$ for some $a_1 \in A_1$ and $a_2 \in A_2$, and it would easily follow that a_1 is a maximum for A_1 and a_2 is a maximum for A_2. This cannot be, since $(A_1|B_1)$ and $(A_2|B_2)$ are assumed to be Dedekind cuts, so A_1 and A_2 have no maxima.

If $(A_1|B_1) = \iota(q_1)$ and $(A_2|B_2) = \iota(q_2)$ for rational numbers q_1, q_2, then we have

$$A_1 = \{x \in \mathbb{Q} \,|\, x < q_1\} \quad , \quad A_2 = \{x \in \mathbb{Q} \,|\, x < q_2\} \quad ,$$

and I claim that the set A defined above is

$$A = \{x \in \mathbb{Q} \,|\, x < q_1 + q_2\} \quad .$$

Indeed, let $A = \{a_1 + a_2 \,|\, a_1 \in A_1, a_2 \in A_2\}$ as above, and let $S = \{x \in \mathbb{Q} \,|\, x < q_1 + q_2\}$.

To verify that $A \subseteq S$, let $a \in A$; then $a = a_1 + a_2$ with $a_1 \in A_1$ and $a_2 \in A_2$, by definition of A. By definition of A_1 and A_2, we have $a_1 < q_1$ and $a_2 < q_2$. It follows that $a = a_1 + a_2 < q_1 + q_2$, showing that $a \in S$, as needed.

To verify that $A \supseteq S$, let s be an element of S, so that $s < q_1 + q_2$. We can write s as $a_1 + a_2$ for some $a_1 < q_1$ and $a_2 < q_2$. Then $a_1 \in A_1$ and $a_2 \in A_2$, and it follows that $s = a_1 + a_2 \in A$.

Now $\iota(q_1 + q_2)$ is the Dedekind cut $(S|T)$, with S as above and $T = \mathbb{Q} \setminus S$. Thus, we have just proved that $(A|B) = \iota(q_1 + q_2)$ in this case. Therefore

$$\iota(q_1) + \iota(q_2) = (A_1|B_1) + (A_2|B_2) = (A|B) = \iota(q_1 + q_2)$$

as needed. ⌟

4.3.1: Let x, y be rational (or even real) numbers. Prove that $|x + y| \leq |x| + |y|$. This is called the 'triangle inequality'.

Solution: This is a simple case-by-case analysis, depending on whether x and y have the same sign (in which case $|x + y| = |x| + |y|$) or different signs (in which case $|x + y| < |x| + |y|$). ⌟

4.3.3: For $q \in \mathbb{Q}$ define $\iota(q)$ as in §4.3.2, that is, as the equivalence class of the constant sequence $q, q, q, \ldots$. Also define $\iota'(q)$ to the the class of the Cauchy sequence corresponding to the decimal expansion of q. Prove that $\iota(q) = \iota'(q)$.

Solution: If $q = b.a_1a_2a_3a_4\cdots$ is the decimal expansion of q, as in (4.2), then the n_th terms of the sequences $\{r_n\}_{n\in\mathbb{N}}$, $\{r'_n\}_{n\in\mathbb{N}}$ representing $\iota(q)$ and $\iota(q')$ are, respectively,

$$r_n = q \quad \text{and} \quad r'_n = b.a_1a_2a_3a_4\cdots a_n \quad .$$

For any $\epsilon > 0$, choose N so that $10^{-N} < \epsilon$. Then for all $n > N$ we have

$$|r_n - r'_n| = |q - b.a_1a_2a_3a_4\cdots a_n| = 0.0\cdots 0a_{n+1}a_{n+2}\cdots < 10^{-N} < \epsilon \quad ,$$

do not distribute

and this verifies that $\{r_n\}_{n\in\mathbb{N}} \cong \{r'_n\}_{n\in\mathbb{N}}$. Therefore

$$\iota(q) = [\{r_n\}_{n\in\mathbb{N}}]_\sim = [\{r'_n\}_{n\in\mathbb{N}}]_\sim = \iota'(q)$$

as needed. ⌟

4.3.4: Let $\alpha \in \mathbb{S}$. Prove that there exist rational numbers $q_1, q_2 \in \mathbb{Q}$ such that $q_1 < \alpha < q_2$.

Solution: Let $\{a_n\}_{n\in\mathbb{N}}$ be a representative for α. By definition of Cauchy sequence, there exists a number N such that for all $m, n > N$ we have $|a_m - a_n| < 1$. In particular, for all $n > N$ we have $|a_{n+1} - a_n| < 1$, and hence

$$a_{n+1} - 1 < a_n < a_{n+1} + 1 \quad .$$

Let $q_1 = a_{n+1} - 1$ and $q_2 = a_{n+1} + 1$. Since $q_1 < a_n$ for all $n > N$, we have $q_1 \leq \alpha$ according to Definition 4.3.10. Since $a_n < q_2$ for all $n > N$, we have $\alpha \leq q_2$. Further $\alpha \neq q_1$ and $\alpha \neq q_2$ ($\{a_n\}_{n\in\mathbb{N}}$ is clearly not equivalent to either constant sequences $\{q_1, q_1, q_1, \dots\}$ or $\{q_2, q_2, q_2, \dots\}$), therefore $q_1 < \alpha < q_2$ as required. ⌟

4.3.5: Let $\alpha, \beta \in \mathbb{S}$ and assume $\alpha < \beta$. Prove that there exists a rational number $q \in \mathbb{Q}$ such that $\alpha < q < \beta$.

Solution: Let $\{a_n\}_{n\in\mathbb{N}}$ and $\{b_n\}_{n\in\mathbb{N}}$ be representatives of α and β. Since $\alpha < \beta$, in particular $\alpha \neq \beta$, so that the representatives are not equivalent:

$$(\exists\epsilon)\,, (\forall M)\,, (\exists m > M)\ : |a_m - b_m| > \epsilon \quad . \tag{A.6}$$

In fact, since $\alpha < \beta$, we may in fact assume that $b_m > a_m + \epsilon$. Further, both sequences are Cauchy sequences; therefore there exists an N such that for all $n, m > N$,

$$|a_m - a_n| < \frac{\epsilon}{3} \quad \text{and} \quad |b_m - b_n| < \frac{\epsilon}{3} \quad .$$

Let $M = N$ in (A.6) and select the corresponding m. Finally, let q be any rational number in the interval $(a_m + \frac{\epsilon}{3}, b_m - \frac{\epsilon}{3})$. For all $n > N$ we have $a_n < a_m + \frac{\epsilon}{3}$, and therefore $a_n < q$; and $b_n > b_m - \frac{\epsilon}{3}$, and therefore $q < b_n$. This implies $\alpha < q < \beta$ as needed. ⌟

4.4.3: As you surely remember from calculus, the number e can be expressed as an infinite series:

$$e = \sum_{n=0}^{\infty} \frac{1}{n!} = 1 + \frac{1}{1!} + \frac{1}{2!} + \frac{1}{3!} + \frac{1}{4!} + \cdots \quad .$$

Is e a rational number?

- Assume e is rational: then $e = \frac{p}{q}$, where p, q are positive integers with no common factor. Show that if this is the case, then the number

$$r = q!\left(\frac{1}{(q+1)!} + \frac{1}{(q+2)!} + \frac{1}{(q+3)!} + \cdots\right) = \frac{1}{q+1} + \frac{1}{(q+1)(q+2)} + \frac{1}{(q+1)(q+2)(q+3)} + \cdots$$

  would necessarily be a positive integer.
- Show that $(q+1)r - 1 < r$, and deduce a contradiction.

The conclusion is that e is an *irrational* number.

This pretty argument is due to A. R. G. Macdivitt and Y. Yanagisawa. The first proof of this fact was obtained by Leonhard Euler in 1737, using 'continued fractions'.

do not distribute

I do not know of a similarly straightforward way to verify that π is irrational.

Solution:

- Since r is a series with positive terms, it is positive. (The series has positive terms and is dominated by $q!e$, so convergence is automatic.) We can write

$$r = q! \sum_{n=q+1}^{\infty} \frac{1}{n!} = q! \left(e - \sum_{n=0}^{q} \frac{1}{n!} \right) \quad .$$

  Since $e = \frac{p}{q}$ by assumption, the product $q!e$ is an integer. The other term equals

$$\sum_{n=0}^{q} \frac{q!}{n!} = \sum_{n=0}^{q} q(q-1)\cdots(q-n+1)$$

  and is therefore a sum of integers. It follows that r is an integer.

- We have

$$\begin{aligned}(q+1)r - 1 &= (q+1)\left(\frac{1}{q+1} + \frac{1}{(q+1)(q+2)} + \frac{1}{(q+1)(q+2)(q+3)} + \cdots \right) - 1 \\ &= \frac{1}{q+2} + \frac{1}{(q+2)(q+3)} + \frac{1}{(q+2)(q+3)(q+4)} + \cdots\end{aligned}$$

  Each term in this series is less than the corresponding term in the series for r, so $(q+1)r - 1 < r$.

- The inequality $(q+1)r < 1 + r$ is equivalent to $qr < 1$. This contradicts the fact that both q and r should be positive integers. ⌟

Chapter 5

5.1.2: Give an alternative proof of Lemma 5.1.2 by proving directly that the function $f : A \to B$ defined in the proof given in §5.1.2 is both injective and surjective (without using the inverse functions g_S).

Solution: Recall that f is defined by setting (for all $a \in A$) $f(a) = f_S(a)$, where S is the unique element of the partition $\mathcal{P}_A$ containing a.

To prove that f is injective, assume that $f(a) = f(a')$ for two elements a, a' of A. In particular, $f(a)$ and $f(a')$ must belong to one and the same element of the partition $\mathcal{P}_B$; call this element T. Since α is a bijective map $\mathcal{P}_A \to \mathcal{P}_B$, T is the image of a unique element S of the partition $\mathcal{P}_A$. By the way f is defined, a and a' must both belong to S. Since f_S is bijective, it is particular injective, and it follows that $a = a'$. Thus, we have shown that $f(a) = f(a')$ implies $a = a'$, and this shows that f is injective.

To prove that f is surjective, let b be an arbitrary element of B. Since $\mathcal{P}_B$ is a partition of B, there is a unique element T of $\mathcal{P}_B$ containing b. Since $\alpha : \mathcal{P}_A \to \mathcal{P}_B$ is a bijection, there is a unique element S of $\mathcal{P}_A$ such that $\alpha(S) = T$. Since $f_S : S \to \alpha(S)$ is bijective, it is in particular surjective; as b is an element of $T = \alpha(S)$, there exists $a \in S$ such that $f_S(a) = b$. By definition of f, $f(a) = f_S(a) = b$. Thus, we have shown that for all $b \in B$ there exists an $a \in A$ such that $f(a) = b$, and this shows that f is surjective. ⌟

5.2.1: Prove that the notation $|A| < |B|$ introduced in §5.2.1 is well-defined. (Note: This will be easy once you understand what it is asking for. Look back at §3.5.1 for a reminder on what 'well-defined' means.)

do not distribute

Solution: 'Well-defined' means that the notion is independent of the representatives. So we have to verify that if $A' \cong A$, $B' \cong B$, and A, B satisfy the condition used to define $|A| < |B|$, then A' and B' satisfy the same condition. The condition on A, B is: there is an injective function $A \hookrightarrow B$, and there is no bijective function $A \to B$.

Assume that $A' \cong A$ and $B' \cong B$. Then there are bijections $f : A' \to A$ and $g : B' \to B$. For every function $\alpha : A \to B$, consider the function $\alpha' = g^{-1} \circ \alpha \circ f$:

$$g^{-1} \circ \alpha \circ f : \quad A' \to B' \quad .$$

If α is injective, then α' is injective: indeed, compositions of injective functions are injective, and both f and g^{-1} are injective since they are bijective. Thus, if there is an injective function $\alpha : A \hookrightarrow B$, then we can deduce that there is an injective function $\alpha' : A' \hookrightarrow B'$.

On the other hand, for every function $\alpha' : A' \to B'$, consider the function $\alpha = g \circ \alpha' \circ f^{-1}$:

$$g \circ \alpha' \circ f^{-1} : \quad A' \to B' \quad .$$

If α' is bijective, then α is also bijective: indeed, compositions of bijective functions are bijective. Thus, if there is a bijective function $A' \to B'$, then there is a bijective function $A \to B$. Contrapositively, if there is *no* bijective function $A \to B$, then we can deduce that there is *no* bijective function $A' \to B'$.

Summarizing: There is an injective function $A' \hookrightarrow B'$, and there is no bijective function $A' \to B'$. Therefore $|A'| < |B'|$.

This shows that the condition used to define $|A| < |B|$ is independent of the representatives A, B chosen for the cardinality classes $|A|$, $|B|$, as needed. ⌟

5.2.2: Prove that the 'relation' $\leq$ satisfies reflexivity and transitivity.

Solution: Reflexivity would say that $|A| \leq |A|$, that is, that there is an injective function $A \to A$. This is indeed the case, since 1_A is injective.

For transitivity, assume that $|A| \leq |B|$ and $|B| \leq |C|$. Then there are injective functions $f : A \to B$, $g : B \to C$. Since f and g are injective, so is $g \circ f : A \to C$. (This is old fare.) Therefore there is an injective function $A \to C$, and this says $|A| \leq |C|$. ⌟

By comparison, 'antisymmetry' is a much subtler business.

5.2.3: Complete the alternative proof of Lemma 5.1.3 sketched in §5.2.2. That is: with notation as in the end of §5.2.2, prove that ψ does *not* equal φ_s for any $s \in S$.

Solution: Recall that $\psi : S \to 0, 1$ is defined by

$$\psi(s) = \begin{cases} 1 & \text{if } \varphi_s(s) = 0 \\ 0 & \text{if } \varphi_s(s) = 1 \end{cases}$$

Now let $s \in S$. Then either $\varphi_s(s) = 0$, or $\varphi_s(s) = 1$.

- If $\varphi_s(s) = 0$, then $\psi(s) = 1$, so $\psi(s) \neq \varphi_s(s)$. This shows that $\psi \neq \varphi_s$ in this case.
- If $\varphi_s(s) = 1$, then $\psi(s) = 0$, so $\psi(s) \neq \varphi_s(s)$. This shows that $\psi \neq \varphi_s$ in this case as well.

This shows that $\varphi_s \neq \psi$ for all $s \in S$, which was the last step to verify in order to conclude the alternative proof of Lemma 5.1.3. ⌟

do not distribute

Of course this argument mimics precisely the proof of Lemma 5.1.3. Make sure you understand that the 'diagonal argument' is just another way to formulate the idea behind the original proof of Lemma 5.1.3.

5.2.5: Explain why $\mathbb{R} \cong 2^{\mathbb{N}}$. (You don't need to be too rigorous, but you should provide a convincing sketch of an argument.)

Solution: First recall that $\mathbb{R} \cong (0,1)$ (Example 3.4.5); proving that $\mathbb{R} \cong 2^{\mathbb{N}}$ is then equivalent to proving that $(0,1)$ and $2^{\mathbb{N}}$ have the same cardinality. Now, every real number between 0 and 1 has a *binary* expansion

$$r = 0.a_0a_1a_2a_3a_4\cdots$$

where each a_i is either 0 or 1. So r determines a function $\mathbb{N} \to \{0,1\}$, defined by $n \mapsto a_n$. Conversely, every function $\mathbb{N} \to \{0,1\}$ determines a real number between 0 and 1, by providing its binary expansion.

Making this sketch into a real proof requires a bit of care: for example, the binary expansions

$$0.1000000\cdots \quad , \quad 0.01111111\cdots$$

determine the same real number ($1/2$, as it happens); and the empty set would correspond to $0.000\cdots = 0$, which is not in $(0,1)$. On the other hand, these problems only affect a countable infinity of numbers, so they do not change the result (cf. Example 5.3.9). ⌟

5.3.1: Prove that any finite product $A_1 \times \cdots \times A_r$ of countable sets is countable.

Question: is a *countable* product of countable sets, $\prod_{n\in\mathbb{N}} A_n$, necessarily countable?

Solution: —We can view $A_1 \times \cdots \times A_{r-2} \times A_{r-1} \times A_r$ as

$$(\cdots(((A_1 \times A_2) \times A_3) \times A_4) \times \cdots) \times A_r \quad .$$

By the argument at the end of Example 5.3.1, $A_1 \times A_2$ is countable if A_1 and A_2 are countable. But then $(A_1 \times A_2) \times A_3$ is also countable, since it is the product of *two* countable sets, again by the first part. And then $(((A_1 \times A_2) \times A_3) \times A_4)$ is the product of two countable sets, so it is countable. Etc.

'Etc.' does not sound very formal, and indeed the proper way to write up this solution is by using induction. I'll write a more formal version of the same argument here, as an illustration. The claim is that if A_i, $i = 1, 2, \ldots$ are all countable sets, then for all $n \geq 1$, $A_1 \times \cdots \times A_n$ is countable. The case $n = 1$ needs no proof, and the case $n = 2$ is covered by the end of Example 5.3.1. Now argue by induction: $n = 2$ works as seed; assume that $n \geq 2$, and that $A_1 \times \cdots \times A_n$ is countable, and prove that $A_1 \times \cdots \times A_{n+1}$ is countable. We have that A_{n+1} is countable by hypothesis, and

$$A_1 \times \cdots \times A_{n+1} \cong (A_1 \times \cdots \times A_n) \times A_{n+1}$$

(write out a bijection if you don't recall this!). This shows that $A_1 \times \cdots \times A_{n+1}$ is isomorphic to the product of *two* countable sets, and it follows that it is countable (Example 5.3.1 again) as needed.

This verifies the induction steps, concluding the proof.

—An infinite countable product of countable sets need not be countable. Indeed, $2^{\mathbb{N}}$ is such a product (it may be viewed as the set of infinite sequences of 0s and 1s, that is, as $\prod_{n\in\mathbb{N}} A_n$, where each A_n is the set $\{0,1\}$). We have proved that $2^{\mathbb{N}}$ is not countable, Corollary 5.2.7. ⌟

5.3.2: Prove that if S is uncountable, and $C \subseteq S$ is countable, then $|S \smallsetminus C| = |S|$.

do not distribute

Solution: Let $N \subseteq (S \setminus C)$ be an infinite countable set, and let $C' = C \cup N$. The union of two countable sets is countable (Example 5.3.1), so C' is infinite countable: $C' \cong N$. Since we have

$$S = (S \setminus C') \cup C' \quad \text{and} \quad S \setminus C = (S \setminus C') \cup N$$

it follows that $|S| = |S \setminus C|$, by Lemma 5.1.2. ⌟

5.3.3: Prove that the set $\mathbb{T}$ of transcendental number has the cardinality of the continuum. (Do not use the continuum hypothesis.)

Solution: This is an immediate application of Exercise 5.3.2: Take $S = \mathbb{R}$, and $C = \mathbb{A}$, the set of algebraic numbers; then since $\mathbb{A}$ is countable, the result of Exercise 5.3.2 shows that the set $\mathbb{T} = \mathbb{R} \setminus \mathbb{A}$ has the same cardinality as $\mathbb{R}$. ⌟

Of course this statement also follows from the continuum hypothesis: the set $\mathbb{T}$ of transcendental numbers is uncountable and contained in $\mathbb{R}$, so $|N| < |\mathbb{T}| \leq |\mathbb{R}|$, and the continuum hypothesis implies $|\mathbb{T}| = |\mathbb{R}|$. But, as shown in the above proof, the fact is true for much simpler reasons.

5.4.2: What is the cardinality of the set $\mathbb{N}^{\mathbb{N}}$ of all functions $\mathbb{N} \to \mathbb{N}$? What about the set of all relations on $\mathbb{N}$? (Hint: What is a relation? what is a function?)

Solution: The answer to the hint is: A relation on a set S is a subset of $S \times S$; a function is a particular type of relations.

Thus, a relation on $\mathbb{N}$ is a subset of $\mathbb{N} \times \mathbb{N}$. We have established that $\mathbb{N} \times \mathbb{N} \cong \mathbb{N}$ (Example 5.3.3); therefore the cardinality of the set of relations on $\mathbb{N}$ is $|2^{\mathbb{N}\times\mathbb{N}}| = |2^{\mathbb{N}}|$.

Since functions are particular types of relations, $\mathbb{N}^{\mathbb{N}}$ is a subset of the set of relations, and in particular $|\mathbb{N}^{\mathbb{N}}| \leq |2^{\mathbb{N}}|$. On the other hand, since $2 < |\mathbb{N}|$, certainly $|2^{\mathbb{N}}| \leq |\mathbb{N}^{\mathbb{N}}|$.

Therefore we have $|2^{\mathbb{N}}| \leq |\mathbb{N}^{\mathbb{N}}| \leq |2^{\mathbb{N}}|$, and we can conclude that $|\mathbb{N}^{\mathbb{N}}| = |2^{\mathbb{N}}|$ by the Cantor-Bernstein-Schröder theorem. ⌟

Chapter 6

6.1.2: Prove Lemma 6.1.3.

Solution: We have to verify that if

$$\forall \epsilon > 0\,,\ \exists \delta > 0\ :\ 0 < |x - c| < \delta \implies |f(x) - L| < \epsilon \tag{A.7}$$

and

$$\forall \epsilon > 0\,,\ \exists \delta > 0\ :\ 0 < |x - c| < \delta \implies |f(x) - L'| < \epsilon \tag{A.8}$$

then $L = L'$.

Informally: Statement (A.7) says that $f(x)$ is as close as we want to L for all x close enough to c; statement (A.8) says that $f(x)$ is as close as we want to L' for the same x. But if L and L' are not the same number, $f(x)$ cannot be arbitrarily close to both of them at the same time. Therefore necessarily $L = L'$.

This is fine, but it won't pass as a formal argument. Here is the formal version of the same idea:

Arguing by contradiction, assume that (A.7) and (A.8) are both true, and yet $L \neq L'$. Let ϵ be $|L - L'|/2$. By (A.8), $\exists \delta_1$ such that if $0 < |x - c| < \delta_1$, then $|f(x) - L| < |L - L'|/2$. By (A.8), $\exists \delta_2$ such that if $0 < |x - c| < \delta_2$,

do not distribute

then $|f(x) - L'| < |L - L'|/2$. Let δ be the minimum of δ_1 and δ_2. Then we have that if $0 < |x - c| < \delta$, then

$$|f(x) - L| < \frac{|L - L'|}{2} \quad \text{and} \quad |f(x) - L'| < \frac{|L - L'|}{2} \quad .$$

Otherwise put,

$$f(x) \in \left(L - \frac{|L - L'|}{2}, L + \frac{|L - L'|}{2}\right) \wedge f(x) \in \left(L' - \frac{|L - L'|}{2}, L' + \frac{|L - L'|}{2}\right)$$

i.e.,

$$f(x) \in \left(L - \frac{|L - L'|}{2}, L + \frac{|L - L'|}{2}\right) \cap \left(L' - \frac{|L - L'|}{2}, L' + \frac{|L - L'|}{2}\right) \quad .$$

This is a contradiction, since these two intervals are disjoint. Thus, the assumption $L \neq L'$ leads to a contradiction. Therefore, $L = L'$. ⌟

6.2.2: What four properties are satisfied by the family of *closed* sets in a topological space?

Solution: We just need to take the properties listed in Definition 6.2.1, and see what they say about the family $\mathcal{C}$ of complements of the sets in $\mathcal{U}$:

- $T \in \mathcal{C}$;
- $\emptyset \in \mathcal{C}$;
- Finite unions of elements of $\mathcal{C}$ are in $\mathcal{C}$;
- Arbitrary intersections of elements of $\mathcal{C}$ are in $\mathcal{C}$. ⌟

These look very much like the axioms satisfied by a topology, but note the switch in the requirements concerning unions and intersections. An arbitrary intersection of closed sets is closed, while an arbitrary intersection of open sets need not be open.

6.2.3: Consider the set $\mathbb{N}$ of natural numbers as a subset of $\mathbb{R}$. Is it open, or closed, or neither in the standard topology? In the cofinite topology? (Why?)

Solution: The set $\mathbb{N}$ is closed in the standard topology. Indeed, its complement is a union of open intervals:

$$\mathbb{R} \smallsetminus \mathbb{N} = (-\infty, 0) \cup (0, 1) \cup (1, 2) \cup (2, 3) \cup \cdots \quad .$$

On the other hand, $\mathbb{N}$ is neither open nor closed in the cofinite topology: it is not open because it is nonempty and its complement is not finite; and it is not closed because it is not finite and it does not equal $\mathbb{R}$. ⌟

6.2.7: Let $\mathbb{R}_{\text{stand}}$, resp. $\mathbb{R}_{\text{cofin}}$ be the topological spaces determined by the standard, resp. cofinite topology on $\mathbb{R}$. Prove that the identity $\mathbb{R} \to \mathbb{R}$ is a continuous function $\mathbb{R}_{\text{stand}} \to \mathbb{R}_{\text{cofin}}$, and that it is not a homeomorphism.

Solution: We have to verify that if a set is open in the cofinite topology, then it is open in the standard topology; and that the converse is not necessarily true.

The open sets in the cofinite topology are $\emptyset$ and the complement of finite sets. The empty set is open in the standard topology (and in any other topology). Let $F = \{r_1, \ldots, r_n\}$ be a finite subset of $\mathbb{R}$, with $r_1 < r_2 < \cdots < r_n$, and let $U = \mathbb{R} \smallsetminus F$ be its complement. Then we can write U as a union of open intervals:

$$U = (-\infty, r_1) \cup (r_1, r_2) \cup \cdots \cup (r_{n-1}, r_n) \cup (r_n, \infty)$$

do not distribute

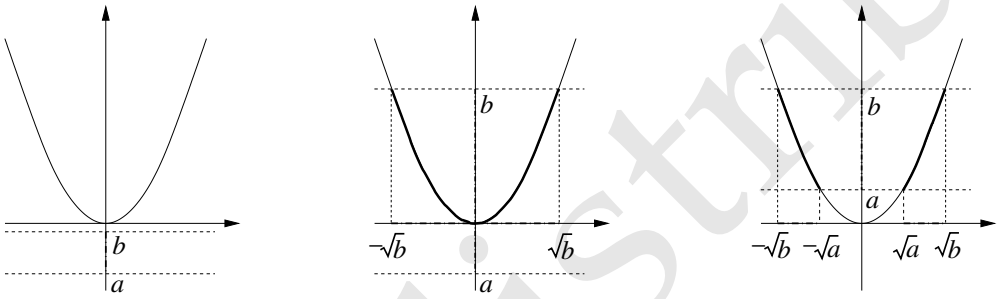
and it follows that U is open in the standard topology, as needed.

To see that open sets in the standard topology are not necessarily open in the cofinite topology, consider the open interval $(0,1)$. This is open in the standard topology, but its complement is not $\mathbb{R}$ and is not finite, so $(0,1)$ is not open in the cofinite topology. ⌟

6.3.2: Prove that the function $f : \mathbb{R} \to \mathbb{R}$ defined by $f(x) = x^2$ is continuous, by using the 'abstract' Definition 6.2.10. (In particular, $\lim_{x\to 2} x^2 = f(2) = 4$. Compare this computation of the limit to the computation in Example 6.1.1. Which one is 'easier'?)

Solution: We just need to check that $f^{-1}((a,b))$ is open, for all $a < b$; that is, that it is a union of open intervals. There are just three possibilities:

- If $a < b \leq 0$, then $f^{-1}((a,b)) = \emptyset$;
- If $a < 0$ and $b > 0$, then $f^{-1}((a,b)) = (-\sqrt{b}, \sqrt{b})$;
- And if $0 \leq a < b$, then $f^{-1}((a,b)) = (-\sqrt{b}, -\sqrt{a}) \cup (\sqrt{a}, \sqrt{b})$.



The inverse image is open in all cases, so f is indeed continuous, as claimed. ⌟

6.4.1: Let $(T, \mathcal{U})$ be a topological space, and let $S \subseteq T$ be a subset. Verify that the collection of sets $U \cap S$, as $U \in \mathcal{U}$, forms a topology.

Solution: We have to verify that the family $\mathcal{V} = \{U \cap S\}_{U\in\mathcal{U}}$ satisfies the four properties listed in Definition 6.2.1.

- $\emptyset \in \mathcal{V}$, since $\emptyset = \emptyset \cap S$ and $\emptyset \in \mathcal{U}$;
- $S \in \mathcal{V}$, since $S = T \cap S$ and $T \in \mathcal{U}$;
- Let $V_1, \ldots, V_r \in \mathcal{V}$. Then there exist elements $U_1, \ldots, U_r \in \mathcal{U}$ such that $V_i = U_i \cap S$. It follows that

  $$V_1 \cap \cdots \cap V_r = (U_1 \cap S) \cap \cdots \cap (U_r \cap S) = (U_1 \cap \cdots \cap U_r) \cap S \quad ,$$

  showing that $V_1 \cap \cdots \cap V_r \in \mathcal{V}$ (since $U_1 \cap \cdots \cap U_r \in \mathcal{U}$). Therefore, finite intersections of elements in $\mathcal{V}$ are in $\mathcal{V}$.
- Let V_α, $\alpha \in I$, be any collection of elements of $\mathcal{V}$. Then there exist elements $U_\alpha \in \mathcal{U}$ such that $V_\alpha = U_\alpha \cap S$. It follows that

  $$\cup_{\alpha\in I} V_\alpha = \cup_{\alpha\in I}(U_\alpha \cap S) = (\cup_{\alpha\in I} U_\alpha) \cap S \quad ,$$

  showing that $\cup_\alpha V_\alpha \in \mathcal{V}$ (since $\cup_\alpha U_\alpha \in \mathcal{U}$). Therefore, arbitrary unions of elements of $\mathcal{V}$ are in $\mathcal{V}$. ⌟

6.4.4: Let $f : T_1 \to T_2$ be a surjective continuous function between topological spaces. Prove that if T_1 is connected, then T_2 is connected.

do not distribute

Solution: We can prove the equivalent (contrapositive) statement: if T_2 is disconnected, then T_1 must also be disconnected. For this, let $S \subseteq T_2$ be a set that is both closed and open, and such that $S \neq \emptyset$, $S \neq T_2$. Since f is surjective, it follows that $f^{-1}(S) \neq \emptyset$, $f^{-1}(S) \neq T_1$. Also, $f^{-1}(T_2 \smallsetminus S) = T_1 \smallsetminus f^{-1}(S)$. (Make sure you are able to verify all these claims!) Since S is open and f is continuous, $f^{-1}(S)$ is open. Since S is closed, $T_2 \smallsetminus S$ is open, and hence (again since f is continuous) $T_1 \smallsetminus f^{-1}(S) = f^{-1}(T_2 \smallsetminus S)$ is open. Thus $f^{-1}(S)$ is also close.

This shows that there is a subset in T_1 as required by the definition of 'disconnected', proving that T_1 is disconnected, as needed. ⌟

6.4.5: Find a sequence in the interval $[0,1)$ that does not converge to a point of $[0,1)$. (This will show that this set is not *compact*.)

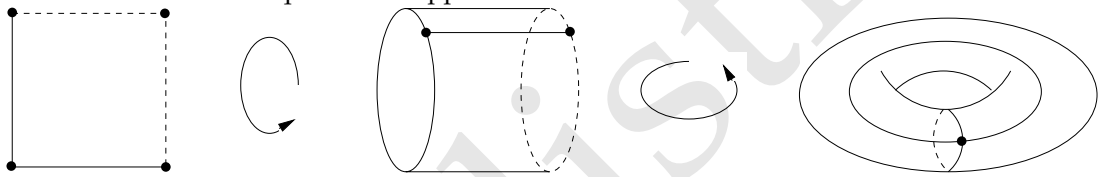
Solution: Let $s_n = 1 - \frac{1}{n}$. Then $\lim_{n\to\infty} s_n = 1$ in the sense stemming from the analogue of (6.1), that is,

$$\forall \epsilon > 0\,,\ \exists N\,,\ \forall n\ :\quad n > N \implies |1 - s_n| < \epsilon \quad .$$

(We checked in §6.1.1 that this expression holds if $1 - s_n = \frac{1}{n}$.) Since 1 is not a point of $[0,1)$, this verifies the statement. ⌟

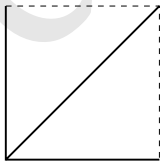
6.4.6: Prove that the Euler characteristic of a torus is 0.

Solution: View a torus as a square with opposite sides identified:

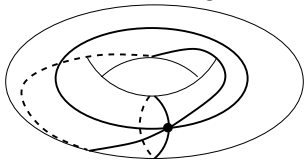


The top and bottom sides of the square become one single curve on the torus; so do the lpictureseft and right sides. The four corners are all identified with each other, determining one point on the torus.

We can triangulate the square very simply:



and this determines a triangulation of the torus, something like this:



This triangulation has 1 vertex, 3 edges, and 2 faces. Even if you don't see this from my mediocre picture, you can understand it by examining the triangulation of the square: the two triangles give two faces; the three solid sides give the three edges (the dotted sides are identified to the solid sides, so they do not contribute more edges); and the four vertices of the square are identified, giving one vertex in the triangulation. The conclusion is that

$$\chi = 1 - 3 + 2 = 0 \quad ,$$

as claimed. ⌟

do not distribute

Index

1_A, 63
2^S, 31
$A/\sim$, 52
$A \times B$, 31
B^A, 63
S^1, 55, 77
S_A, 80
$[\cdot]_\sim$, 48
$\mathbb{A}$, 127
$\mathbf{C}$, 3
$\mathbb{D}$, 96
$\mathbb{N}$, 3
$\mathbb{N}^*$, 122
$\mathbf{Q}$, 3, 55
$\mathbb{R}$, 3, 91
$\mathbb{R}^{\geq 0}$, 4
$\mathbf{S}$, 101
$\mathbb{Z}$, 3
$\bigcap_{i\in I}$, 28
$\bigcup_{i\in I}$, 28
$\mathcal{P}(S)$, 31
$\cap$, 27
$\circ$, 66
$\cong$, 74
$\cup$, 27
$\emptyset$, 3
$\exists!$, 59
$\exists$, 5
$\forall$, 5
$\hookrightarrow$, 71
id_A, 63
$\iff$, 10
$\operatorname{im} f$, 60
$\impliedby$, 12
$\implies$, 11
$\in$, 3
$\lim_{x\to c} f(x)$, 136
$|\cdot|$, 25
$\neg$, 8
$\overline{\mathbf{Q}}$, 128
$\setminus$, 27
$\subset$, 26
$\subseteq$, 25
$\subsetneq$, 26
$\supset$, 26
$\supseteq$, 26
$\twoheadrightarrow$, 71
$\vee$, 8
$\wedge$, 8
$f(a)$, 59
f^{-1}, 60, 69

algebraic
  closure, 128
  geometry, 56
  number, 127
analysis, 148
AND, 8
antisymmetric relation, 35, 130
associative
  $\wedge$ and $\vee$ are, 10
  composition of function is, 66
  operation, 80
axiom of choice, 73

Bernstein, Felix (1878-1956), 130
bijection, 70
bijective function, 70
binary, 107
binomial
  coefficients, 42
  theorem, 42

canonical decomposition, 82
Cantor, Georg Ferdinand Ludwig Philipp (1845-1918), 116
  Je le vois, mais je ne le crois pas, 130
Cantor-Bernstein-Schröder theorem, 130
cardinality, 114
  of a finite set, 25
cartesian product, 31
category, 142
Cauchy, Augustin-Louis (1789-1857)
  sequence, 100
    equivalence class of, 101
claim, 17
coarser topology, 153
codomain of a function, 60
cofinite topology, 141
Cohen, Paul Joseph (1934-2007), 129
commutative
  algebra, 56
  diagram, 66
  operation, 79
compact topological space, 151
complement, 27
complete set, 94
completion, 95
complex numbers, 3
composition of functions, 66
connected topological space, 151
connective, 8
constant
  function, 64
  sequence, 100
continuous function
  $\mathbb{R} \to \mathbb{R}$, 139
  between topological spaces, 142
continuum, 129
contradiction, 10, 21
contrapositive, 20
corollary, 17
countable set, 121

De Morgan, Augustus (1806-1871), 10
decimal expansion, 91, 100, 107

do not distribute

of a Dedekind cut, 109
rational vs. irrational, 110
Dedekind, Richard (1831-1916), 130
cut, 96
definition, 25
algebraic number, 127
antisymmetric relation, 35
cardinal number, 113
cardinality, 25, 113
cartesian product, 31
Cauchy sequence, 100
closed set, 141
complete set, 94
completion, 95
composition of functions, 66
continuous function, 139, 142
countable set, 121
Dedekind cut, 96
equivalence
class, 48
class of Cauchy sequences, 101
of Cauchy sequences, 101
relation, 46
function, 59
graph of a function, 59
homeomorphism, 141
identity function, 63
image, 60
induced topology, 148
injective function, 70
intersection, 27, 28
inverse
image, 60
of a function, 67
isomorphic sets, 74
isomorphism of sets, 69
least upper bound, 93
maximum, 36
minimum, 36
open set in the standard topology, 144
order relation, 36
ordering
on cardinalities, 118
on Cauchy sequences, 102
on Dedekind cuts, 97
partition, 49
power set, 31
preimage, 60
quotient, 52
reflexive relation, 35
relation, 35
set
of functions, 63
of parts, 31
standard topology, 145
subset, 25
surjective function, 70
topology, 139
transcendental number, 128
transitive relation, 35
uncountable set, 121
union, 27, 28
well-ordering, 37
dense, 95
denumerable set, 122
Descartes, René (1596-1650), 31
diagonal argument, 116, 119
diagram, 66
difference of sets, 27
disconnected topological space, 151
discontinuity, 139
domain of a function, 60

element, 1, 25
empty set, 3, 26
equivalence
class, 48
of Cauchy sequences, 101
relation, 36, 46
Euler, Leonhard (1707-83), 112
characteristic, 152

field of fractions, 56
finer topology, 153
Fraenkel, Abraham Halevi (1891-1965), 30
function, 59
bijective, 70
composition, 66
constant, 64
identity, 63
inclusion, 64, 71
injective, 70
natural projection, 64, 71
one-to-one, 70
onto, 70
surjective, 70
fundamental group, 152

Galois, Évariste (1811-32), 128
Gauss, Karl-Friedrich (1777-1855), 40
graph of a function, 33, 59
greatest lower bound, 36
Grothendieck, Alexander (1928-2014), 182
group, 80, 182
fundamental, 152
Galois, 128

Hausdorff, Felix (1868-1942), 148
Hilbert, David (1862-1943)
Grand Hotel, 123
Hippasus of Metapontum (530-450 BC), 93
homeomorphic topological spaces, 141
homeomorphism, 129, 141

identity function, 63
if and only if, 10
if. . . , then, 11
iff, 10
image
inverse, 60
of a function, 60
of a subset, 60
of an element, by a function, 60
inclusion function, 64, 71
indexed set, 4, 62
induced
function, 84
relation, 44
topology, 148
induction, 39
infimum, 36
injection, 70
injective function, 70
integers, 3, 113
integral domain, 56
intersection of sets, 27, 28
inverse
image, 60
left, 67
of a function, 69
right, 67
two-sided, 67
irrational, 92
e is, 112
number, decimal expansion, 110
isomorphic sets, 74
isomorphism
of sets, 69, 113
of topological spaces, 141

do not distribute

Kronecker, Leopold (1823-91), 122

lattice, 32
least upper bound, 36, 93
  property, 94
Lefschetz, Solomon (1884-1972), 148
left inverse, 67
lemma, 17
limit
  of a function, 136
  of a sequence, 136
localization, 56
logic, 8
logical connective, 8

maximum, 36, 96
minimum, 36
modulo, 48
monic polynomial, 95
morphism, 142

NAND, 15
natural
  numbers, 3, 113
  projection, 64, 71
NOT, 8
number
  p-adic, 106
  algebraic, 127
  as cardinality, 113
  cardinal, 113
  complex, 3
  integer, 113
  irrational, 92
  natural, 3, 113
  ordinal, 114
  perfect, 18
  rational, 3, 113
  real, 3, 91, 113
  theory, 56, 106, 128
  transcendental, 128

one-to-one
  correspondence, 70
  function, 70
only if, 12
onto function, 70
open set, 140
OR, 8
order relation, 36
ordered
  pair, 31
  set, 2
ordinal number, 114
ordinary topology
  on $\mathbb{R}$, 145
  on $\mathbb{R}^n$, 146

partition, 49
perfect number, 18
permutation, 80
pigeonhole principle, 73
Poincaré, Jules Henri (1854-1912), 122
power set, 31
preimage, 60
product of sets, 31
projection, 64
  natural, 64, 71
proof
  by contradiction, 21
  by contrapositive, 20
  by induction, 39
  direct, 20
proposition, 17
Pythagoras (570-495 BC), 92

quantifiers, 4
quotient, 46, 52

range, 61
rational
  numbers, 3, 113
  root, theorem, 95
real numbers, 3, 91, 113
reflexive relation, 35
relation, 35
  antisymmetric, 35
  equivalence, 36, 46
  order, 36
  reflexive, 35
  symmetric, 35
  transitive, 35
representative, 48
right inverse, 67
Russel, Bertrand (1872-1970)
  paradox, 3, 30

Schröder, Friedrich Wilhelm Karl Ernst (1841-1902), 130
set, 1, 25
  closed, 141
  countable, 121
  denumerable, 122
  empty, 3, 26
  indexed, 4, 62
  of parts, 31
  open, 140
  ordered, 2
  power, 31
  theory, 1, 25
    naive, 1
    Zermelo-Fraenkel, 30
  uncountable, 121
  universe, 30
Socrates (470-399 BC), 19
source of a function, 60
standard topology
  on $\mathbb{R}$, 145
  on $\mathbb{R}^n$, 146
subset, 25
supremum, 36
surjection, 70
surjective function, 70
syllogism, 19
symmetric relation, 35

target of a function, 60
theorem, 2, 17
  binomial, 42
  canonical decomposition, 82
  Cantor-Bernstein-Schröder, 130
  continuity in the standard topology, 146
  equivalence relations determine partitions, 50
  injective/surjective functions and existence of inverses, 72
  least upper bounds for Cauchy sequences, 105
  least upper bounds for Dedekind cuts, 98
  partitions determine equivalence relations, 51
  $\mathbb{R}$ is uncountable, 122
  rational root, 95
  $|S| < |2^S|$, 119
  there exist different cardinal infinities, 117
topological space, 129, 140
  compact, 151
  connected, 151
topology
  coarser, 153

do not distribute

cofinite, 141
discrete, 140
finer, 153
induced, 148
on a set, 139
standard, on $\mathbb{R}$, 145
standard, on $\mathbb{R}^n$, 146, 149
trivial, 140
torus, 77, 153
total ordering, 94, 98, 104
transcendental number, 128
transitive relation, 35
triangle inequality, 101, 106
triangulation, 152
truth table, 9
two-sided inverse, 67

uncountable set, 121
union of sets, 27, 28
universe set, 30
upper bound, 93
least, 93

Venn, John (1832-1923)
diagram, 27

Weierstrass, Karl Theodor Wilhelm (1815-97), 148
well-defined, 85
well-ordering, 37
principle, 37
Wittgenstein, Ludwig Josef Johann (1889-1951), 1, 122

Zermelo, Ernst Friedrich Ferdinand (1871-1953)
-Fraenkel set theory, 30

do not distribute