Introduction to Advanced Mathematics

Pete L. Clark

Contents

Introduction

1. What is this course, and what is this book?

In the contemporary American mathematical curriculum, there is a clear dichotomy between lower level courses – like calculus and differential equations – that emphasize computations and solving concrete, numerical problems – and higher level courses – like analysis and algebra – that emphasize proofs involving abstract mathematical structures. For a long time the gap between the lower and upper level courses was simply a chasm that aspriing math majors needed to jump, but about 30 years ago the need for a *transitional course* was recognized.

This is the text for such a course. Our goal is to start from scratch and lay solid foundations for future study of advanced mathematics. What is involved in this?

To study contemporary advanced mathematics one needs to be fluent in several languages. The first of these languages is **sets**: these are the building blocks of all mathematical objects, including numbers, functions and shapes. The second of these languages is **logic**: this gives the tools for correct reasoning with mathematical objects, which is crucially important since the main business of contemporary mathematics is learning, checking and making correct arguments about mathematical objects. The last language of advanced mathematics is **English**.[1] By this I mean that – in rather sharp contrast to lower level mathematics – the medium with which we express our logically correct arguments is the English language. We write in complete sentences and subject ourselves to the usual rules of English grammar. Equations and other mathematical expressions are still present but are always accompanied by words. This is true to an extent that takes some getting used to.

This text does not have a chapter on English: we all have more than one teacher! But by reading the text you will see numerous examples (good ones, I hope) of how mathematics is written in English, and solving the exercises will give you lots of practice writing mathematics in English. Rather we begin with chapters on sets (Chapters 1 and 3) and a chapter on logic (Chapter 2). The next order of business ought to be to learn *proof techniques*, which are roughly template logical arguments that occur over and over again in mathematics. However, in order to see and perform meaningful specimens of mathematical proofs, we need to introduce some further mathematical structures that we can use in our proofs. This takes place in Chapters 4 and 5, which discuss numbers and inequalities and then some basic

[1]Someone elsewhere in the world could replace "English" with their native language...at least up to a certain point. However in the year 2023, the percentage of published mathematical writing that is in English is well over 90% and rising.

number theory, especially the notion of divisibility of integers. Then in Chapter 6 we discuss basic proof techniques, especially direct proof, proof by contrapositive and proof by contradiction. In Chapter 7 we discuss at length the greatest proof technique in all of mathematics: **induction**. (In fact the justification for induction involves some background on numbers and inequalities, which is another reason we treated these in Chapters 4 and 5.) In Chapter 8 we discuss two further ubiquitous classes of mathematical structures, relations and functions. Earlier I said that sets are the building blocks of all mathematical objects, and that holds true here: indeed a relation is a certain kind of set and a function is a certain kind of relation. But as you have probably seen before, notwithstanding their set-theoretic definition, functions have a "dynamic" quality that gives their study a different character.

This forms the core of the text; other chapters are closely related but need not be part of a first course in the subject (this is discussed in more detail shortly).

2. Distinguishing Features of this Text

Compared to other texts on the same subject, this text includes more material for the "serious student" of mathematics at the advanced undergraduate level and beyond. A generation or two ago, this archetype of student was assumed to pick up sets, logic and proofs in the context of other courses. Nowadays a much higher percentage of American undergraduates take a course on the subject. I have come to feel that even students who are very serious indeed can learn a lot from this course if they are sufficiently engaged. The standard presentations of this material are, frankly, a bit boring for many students, who wake up fully during the discussion of induction in the second half of the course. Such students generally do well rather than bady, but they are far from maximizing this oportunity to increase their skills and techniques.The line between boredom and confusion is thinner than you think!

Chapter 1 of this text presents very basic material on sets, with an attempt to be as clear as possible and not too heavy-handed. However, I do like to assign Exercise 1.20 on Kuratowski's definition of the ordered pair rather early on (usually in the second problem set). This exercise is difficult for new students of set theory and really gets them to firm up thoughts and techniques about how sets work: e.g. realizing that dealing with $\{a, b\} = \{c, d\}$ involves several cases.

All subsequent chapters include some material either exploring less familiar topics or covering standard topics in extra depth. Examples:

• In Section 2.6 we discuss how all logical operators can be built out of $\wedge$, $\vee$ and $\neg$. This is actually a standard exercise in electrical engineering.

• In Section 2.9 we discuss isotone logical operators and their connection with Sperner families.

• Exercise 2.23 echoes a classic psychology experiment, the Wason Selection Task.

• Section 4.3 presents the general definition of a well-ordered subset of $\mathbb{R}$, and well-ordered subsets other than $\mathbb{N}$ and $\mathbb{Z}^+$ are explored in some exercises.

• Section 5.4 treats greatest common divisors from the "multiplicative" perspective, i.e., as a common divisor that is divisible by all other common divisors.

• Exercise 5.6 is a "zoological" generalization of the Divsion Theorem.

• In Section 6.6 we discuss the most efficient way to prove that N statements are equivalent. We show that this requires N basic implications. Then in this section and in Section 9.1 we show that the only way to do this is arrange the statements in a circle and prove that each implies the next statement in the circle.

• In Section 7.7 we give the beginnings of a general discussion on closed forms of power sums.[2]

• After a discussion of Fibonacci numbers in Section 7.12, Section 7.13 contains a more general discussion of closed forms for homogeneous linear recurrences.

• Section 7.14 treats Upward-Downward Induction and applies it to prove the Arithmetic-Geometric Mean Inequality.

• Section 7.15 gives a remarkable inductive proof of the Fundamental Theorem of Arithmetic due to Lindemann and Zermelo. We also give an inductive proof Euclid's Lemma due to Rogers (that I don't like as much, but you need not agree).

• In our dicusssion of relations in Chapter 8, we are almost as interested in partial orderings as we are in equivalence relations.[3]

• In Secion 8.4 we discuss how to build the smallest binary relation containing a given relation that is reflexive, symmetric and/or transitive.

• In our discussion of functions between finite sets in Section 8.7 we discuss Stirling numbers and Bell numbers.

At the same time, this text really is meant to be used by students who are first learning about sets, logic, how to read and write proofs, and so forth. This means that not all material is presented at the same level or in the same way. While the more optional and advanced material is often covered a bit briskly, I have tried to explain the basic material very carefully. Especially, the explanations on sets, logic and proof techniques are based on teaching the Math 3200 course several times over a period of about 15 years and represent the best I have come up with over this period of time. I hope that these parts of the text are readable: suggestions for improvement will be warmly received.

[2]This could be taken further, but we have chosen not to introduce Bernoulli numbers.

[3]If I had a magic wand to change the undergraduate curriculum, I would make partially ordered sets much more prominent.

3. A Guide to the Contents of the Text

This was written to be the course text for Math 3200, but not *just* to be the course text for Math 3200. There is in fact about twice as much material here for this (or any?) one semester undergradute course. I will describe what is covered in Math 3200 and then what is *not* covered in Math 3200 (and why it is here).

3.1. Math 3200 Coverage. In our Math 3200 course we will cover the following parts of the text:

• Chapter 1 (Sets): This chapter is covered in its entirety. Section 1.5 on indexed families of sets and Section 1.6 on partitions are less important than the others: although indexed families of sets are ubiquitous in several later undergraduate courses (especially Math 4100 Real Analysis and Math 4200 General Topology), in our course they play a subsidiary role. The material on partitions will be revisited towards the end of the course when equivalence relations are discussed but it not very important until then. The other sections of this chapter are *excruciatingly central and important* to us.

• Chapter 2 (Logic): All sections in this chapter are covered except Section 1.9 on isotone logical operators. Section 2.6 gives some structural results on logical operators. These structural results are not needed in the rest of the course, but I think these results are on the one hand quite interesting and on the other hand helpful in increasing understanding of what logical operators are all about. Once again every other section is excruciatingly central and important to us.

• Chapter 4 (Numbers, Inequalities and Rings): This short chapter is covered in its entirety. Section 4.2 on ordered field axioms has a "taking our medicine" feel: in it we provide the foundations for algebraic manipulation of inequalities. This material is unfortunately not so inherently interesting, nor is it necessary to know the ordered field axioms by name or number. Rather one must learn how to manipulate inequalities: what is permitted, what is not permitted, and when the operation flips the inequality. Section 4.3 discusses well-ordered subsets of the real numbers. To be honest, for our course and for most future undergraduate work it would be sufficient to know that for every integer N, the set $\mathbb{Z}^{geq N}$ of integers that are greater than or equal to N is well-ordered. This fact is however highly important for us: it will be used to justify our most powerful proof technique (Mathematical Induction) later on.

• Chapter 5 (Number Theory): Sections 5.1 through 5.3 are covered in our course. We also state Euclid's Lemma from Section 5.6 and prove it later.

• Chapter 6 (Fundamentals of Proof): This chapter is covered in its entirety. Sections 6.2 on direct proof, Section 6.3 and contrapositve and 6.4 on contradiction are of the highest level of importance. The other sections are less important.

• Chapter 7 (Induction): This chapter covers the most important proof technique in our course (and almost certainly, in mathematics as a whole). Most sections will be covered: the sections that *will not* be covered are Section 7 on power sums,

Section 13 on homogeneous linear recurrences and Section 14 on upward-downward induction. From Section we will present the Lindemann-Zermelo inductive proof of the Fundamental Theorem of Arithmetic, from which we can deduce Euclid's Lemma. This is not a proof that a student in the course needs to know; we cover it to show how powerful induction truly is.

• Chapter 8 (Relations): All sections will be covered *except* Section 8.3 on composition of relations and Section 8.4 on relational closures. The material of Sections 8.5 and 8.6 on functions is the most important material in the latter part of the course. We will also cover Section 9.2 on congruences during our discussion of equivalence relations, but this material will not be studied in much depth in our course (but see Math 4000 and Math 4400).

3.2. Infinite Sets. The last three chapters of this text treat Cantor's theory of infinite sets, cardinal and ordinal numbers. Our approach has gradually escalating sophistication but is always "naive" in that we do not discuss set-theoretic axioms except for the Axiom of Choice. This material had traditionally been part of the syllabus for a course on introduction to advanced mathematics. However, the first few times I taught this course I had time for only a couple of lectures on this, which moreover seemed to naturally live at a higher level of abstraction than the rest of the material.

Moreover, whereas sets, relations and functions occur in *every* future undergraduate course, even the notion of an uncountable set – let alone infinite cardinals or ordinals – is much less central to American mathematics: in our curriculum, it would come up firmly in Math 4200 (General Topology) and in passing in Math 4100 (Real Analysis). In fact, after a recent revision, the departmental syllabus for Math 3200 no longer includes this material. I have put it in this text anyway – why remove it? However, one could imagine a "foundations of *graduate* mathematics" course in which this material might find a more natural home.

3.3. Discrete Mathematics. In both Mathematics and Computer Science courses there is often a course in **Discrete Mathematics** that has highly overlapping content with an introduction to advanced mathematics. When I first starting teaching this course I found this a bit odd: this material has no greater connection to discrete mathematics than analysis, topology, algebra or geometry. However more recently I changed my mind a bit: if counting arguments involving infinite sets are part of the curriculum, then certainly counting arguments involving finite sets should be too.[4] So running contrapuntally through this text is a substantial amount of discrete mathematics – not a full course's worth, but too much less than that. This material is however kept mostly independent of the other material of the text, so that the reader can freely skip it.

Here is where this material lives in the text:

Section 2.9 on isotone logical operators is largely an excuse to introduce Sperner families of sets: i.e., families of subsets of a fixed set with no containments among distinct elements of the family.

[4] I suspect that the majority of undergraduate majors would get more out of the finite case than the infinite case.

Chapter 3 discusses basic techinques involving counting in finite sets, including the Binomial Theorem, the Inclusion-Exclusion Principle and the Pigeonhole Principle.

Section 8.7 applies Inclusion-Exclusion to give formulas for the number of surjections from an m element set to an n element set and for the number of partitions of an n-element set.

Secion 9.3 is a fairly substantial introduction to graph theory. The genesis of this, believe it or not, was Theorem 9.37, which I presented once in Math 3200 as an example of an induction proof that goes by "reducing the complexity" rather than proving a statement explicitly involving a discrete variable. Perhaps we have gotten a bit carried away in the amount of graph theory discussed (there is a course for this...) but on the one hand many of the proofs are nice applications of the material devleoped and on the other I think it is nice for an undergraduate student to hear about 20th and 21st century mathematical theorems rather than just results that are centuries old. Our discussion of Ramsey numbers and Schur numbers is part of a general knowledge of discrete mathematics that I did not receive as a student but it would be nice if contemporary students did.

Section 9.4 covers three fundamental results from the theory of finite, partially ordered sets. Once again I wish that order theory were a legitimate part of the mathematical curriculum.

CHAPTER 1

Sets

1. Introducing Sets

Sets are the first of the three languages of mathematics. They are the most basic kind of mathematical structure; all other structures are built out of them.[1]

A set is a collection of mathematical objects. This is not a careful definition; it is an informal description meant to convey the correct intuition.

1.1. Many Examples. We begin with some familiar examples.

EXAMPLE 1.1. *One can think of a set as a kind of club; some things are members; some things are not. So for instance current UGA students form a set. You are a member; I am not. Past or present presidents of the United States form a set. Barack Obama is a member. Hillary Clinton is not.*

EXAMPLE 1.2. *The positive integers*

$$\mathbb{Z}^+ = \{1, 2, 3, \ldots\}$$

are a set. The positive integer 1 *is an* ***element****, or* ***member*** *of* $\mathbb{Z}^+$*: we write this statement as*

$$1 \in \mathbb{Z}^+.$$

So is the positive integer 2*: we write*

$$2 \in \mathbb{Z}^+.$$

Similarly,

$$3 \in \mathbb{Z}^+,\ 4 \in \mathbb{Z}^+,\ \textit{and so forth.}$$

The negative integer -3 *is* not *an element of* $\mathbb{Z}^+$*. We write this as*

$$-3 \notin \mathbb{Z}^+.$$

The integer 0*, which is not positive (this is an explanation of terminology, not a mathematical fact), is not a member of* $\mathbb{Z}^+$*:*

$$0 \notin \mathbb{Z}^+.$$

Also $\frac{4}{5} \notin \mathbb{Z}$*,* $\sqrt{2} \notin \mathbb{Z}^+$ *and Barack Obama* $\notin \mathbb{Z}^+$*. Of course* lots of things *are not in* $\mathbb{Z}^+$*: we had better move on.*

[1] Like most broad, sweeping statements made at the beginning of courses, this one is not *completely* true. Mathematics is at least 2500 years old: Pythagoras died circa 495 BCE. The practice of describing all mathematical objects in terms of sets dates from approximately 1900. Many mathematicians have at least contemplated basing mathematics on other kinds of objects; something called "categories," first introduced in the 1940's by Eilenberg and Mac Lane, have long had a significant (though minority) popularity. Nevertheless every student or practitioner of mathematics must speak the language of sets, which is what we are now introducing.

EXAMPLE 1.3. *For any whole number $n \geq 1$, $\{1, 2, \ldots, n\}$ is a set, whose elements are indeed $1, 2, 3, \ldots, n$. Let us denote this set by $[n]$. So for instance*

$$5 \in [9] = \{1, 2, 3, 4, 5, 6, 7, 8, 9\}$$

and

$$9 \notin [5] = \{1, 2, 3, 4, 5\}.$$

(For whole numbers $a, b \geq 1$, we have $a \in [b]$ precisely when $a \leq b$.)

EXAMPLE 1.4. *The non-negative integers, or* ***natural numbers***

$$\mathbb{N} = \{0, 1, 2, 3, \ldots, \}$$

are a set. The only difference between $\mathbb{Z}^+$ and $\mathbb{N}$ is that $0 \in \mathbb{N}$ whereas $0 \notin \mathbb{Z}^+$. (This may seem silly, but it is actually useful to have both $\mathbb{Z}^+$ and $\mathbb{N}$ around.)

EXAMPLE 1.5. *The integers, both positive and negative*

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

form a set. This time $-3 \in \mathbb{Z}$, but still $\frac{4}{5} \notin \mathbb{Z}$, $\sqrt{2} \notin \mathbb{Z}$ and Barack Obama $\notin \mathbb{Z}$.

EXAMPLE 1.6. *A* ***rational number*** *is a quotient of two integers $\frac{a}{b}$ with $b \neq 0$. Rational numbers have many such representations, but $\frac{a}{b} = \frac{c}{d}$ exactly when $ad = bc$. The rational numbers form a set, denoted $\mathbb{Q}$. So now we have $\frac{4}{5} \in \mathbb{Q}$. Still $\sqrt{2} \notin \mathbb{Q}$ (this is an important theorem of ancient Greek mathematics that we will discuss later), and Barack Obama $\notin \mathbb{Q}$.*

EXAMPLE 1.7. *The real numbers form a set, denoted $\mathbb{R}$. A real number can be represented as an integer followed by an infinite decimal expansion. Still Barack Obama $\notin \mathbb{R}$.*

EXAMPLE 1.8. *A* ***complex number*** *is an expression of the form $a + bi$, where $a, b \in \mathbb{R}$ and $i^2 = -1$. The set of complex numbers is denoted by $\mathbb{C}$. The number i is a member; still Barack Obama $\notin \mathbb{C}$.*

Actually, apart from Example 1, Barack Obama is not going to be a member of any of the sets we will introduce. (Nor Mitt Romney, nor Olivia Rodrigo, nor...) To be honest, although we insisted that *anything* can be an element of a set, in mathematics – apart from preliminary discussions of exactly the sort you have just seen – we only consider sets that contain as members **mathematical objects**.

On the one hand, this is not surprising because mathematics is, obviously, the study of mathematical things. On the other hand, the notion of a "non-mathematical object" brings some philosophical difficulties. In particular, since sets contain objects without any notion of multiplicity, in order to form a set, given two objects x and y, we need to have a clean answer to the question "Does $x = y$?" When considering identity of non-mathematical objects we are drawn into delicate issues of spatio-temporal continuity.[2]

We did not begin by saying that "A set is a collection of mathematical objects" because we were not – and are still not now – ready to address the natural followup question "What is a mathematical object?" But here is a taste of a kind of answer:

[2]E.g.: is the you of 2023 the same person as the you of 2005? Your atoms are different. If a worm is divided into two, are the old worm and the two new worms one, two or three different objects? And so forth. These are fun questions, but they have nothing to do with mathematics.

Any set of mathematical objects is itself a mathematical object. For now this probably sounds both circular and insufficient. It turns out that neither of those things is true. You may – or may not; it's by no means necessary – understand why a bit better by the end of the course.

EXAMPLE 1.9. *The Euclidean plane forms a set, denoted $\mathbb{R}^2$. Its elements are the points in the plane, i.e., ordered pairs (x, y) with x, y real numbers: we write $x, y \in \mathbb{R}$. For a positive integer n, n-dimensional Euclidean space forms a set, whose elements are ordered tuples $(x_1, \ldots, x_n)$ of real numbers. It is denoted $\mathbb{R}^n$.*

EXAMPLE 1.10. *Here are some examples from geometry / linear algebra:*

a) *The lines ℓ in the Euclidean plane $\mathbb{R}^2$ form a set.*
b) *The planes P in Euclidean space $\mathbb{R}^3$ form a set.*

EXAMPLE 1.11. *The continuous functions $f : [0, 1] \to \mathbb{R}$ form a set.*

Some of these examples were an excuse to introduce common mathematical notation. But I hope the idea of a set is clear by now: it is a collection of (for us: mathematical!) objects. Practically speaking, this amounts to the following: if S is a set and x is any object, then exactly one of the following must hold: $x \in S$ or $x \notin S$. That's the point of a set: if you know exactly what is and is not a member of a set, then you know the set. Thus a set is like a bag of objects...but not a red bag or a cloth bag. The bag itself has no features: it is no more and no less than the objects it contains.

REMARK 1.12. *For most of the sets one initially meets in mathematics, all the elements are either numbers of some kind or points in some kind of geometric space. Examples 10 and 11 are included specifically to show that this need not be the case. In fact both of these kinds of examples – sets of subsets of some kind of space and sets of functions – are ubiquitous in higher mathematics.*

Sometimes it is helpful to think of the elements of an arbitrary set as "points," but this is just a heuristic: they need not actually be points. For that matter, "point" is not something that has an agreed upon definition throughout mathematics.

EXAMPLE 1.13. *The empty set, denoted $\varnothing$, is a set. This is a set that contains no objects whatsoever: for any object x, we have $x \notin \varnothing$.*

Not only is $\varnothing$ a legitimate set, in some sense it is the most important set!

1.2. Equality of Sets. We reiterate the following **basic principle** of sets: two sets S and T are **equal** precisely when they contain exactly the same objects: that is, for any object x, if $x \in S$ then $x \in T$, and conversely if $x \in T$ then $x \in S$.

An important consequence of this basic principle is that whereas above we said that the empty set $\varnothing$ is *a* set which contains no objects whatsoever, in fact it is *the* set which contains no such objects: any two sets which contain nothing contain exactly the same things!

1.3. Finite Lists and Finite Sets. A **finite list** of elements is something of the form $x_1, x_2, \ldots, x_n$, where n is a positive integer, and for each $1 \leq i \leq n$, we have that x_i is an object. It is convenient to also allow the **empty list** when $n = 0$. Note well that this is really a different concept from that of a finite set in that the

order is taken into account and that the objects in the list are not required to be distinct. For instance,

$$\ell_1 : 1, 1, 1 \tag{1}$$

and

$$\ell_2 : 1, 1, 1, 1, 1, 1 \tag{2}$$

are two finite lists, and they are different: ℓ_1 has length 3, while ℓ_2 has length 6.

A set is **finite** if there is a finite list $\ell : x_1, \ldots, x_n$ such that

$$S = \{x_1, \ldots, x_n\}.$$

In other words, for any object x, we have $x \in S$ precisely when $x = x_i$ for some i. We say that S is the finite set **associated** to the finite list ℓ. The empty set is associated to the empty list. A set is **infinite** if it is not finite.

The associated set of a finite list of length n has *at most* n elements, but it may have fewer: by Exercise 1.4, this happens exactly when the list has repetitions. Here is some terminology to facilitate further exploration of this point: A finite list

$$\ell : x_1, \ldots, x_n$$

is called **irredundant** if the entries are all distinct: for all $1 \leq i \neq j \leq n$ we have $x_i \neq x_j$. Thus Exercise 1.4 shows that if ℓ is a finite list of length n with associated finite set A, we have $\#A = n$ precisely when the list ℓ is irredundant.

Moreover the same finite set may be associated to two different finite lists: e.g. the finite set associated to the list ℓ_1 of (1) is $\{1\}$, and the finite set associated to the list ℓ_2 of (2) is also $\{1\}$. In fact every nonempty finite set is associated to infinitely many finite lists: Exercise 1.5.

The **cardinality** of a finite set is the least number n of elements such that the set is associated to a list of n elements: in other (perhaps simpler) terms, it is the number of elements of a defining irredundant finite list. I will denote the cardinality of a finite set by $\#S$.

(At the end of the text we explore a notion of cardinality for infinite sets. This is much more interesting!)

Example 1.14. *Finite vs. Infinite:*

a) *The set* $[n] = \{1, 2, \ldots, n\}$ *is finite, and* $\#[n] = n$.
b) *The sets,* $\mathbb{Z}^+$, $\mathbb{Z}$, $\mathbb{Q}$, $\mathbb{R}$, $\mathbb{C}$ *are all infinite.*
(In fact most interesting mathematical sets are infinite.)

We call will the process of defining a set using a finite list an **extensional** definition of a set. The other way of giving a set, called **intentional**, is by giving a defining **property** of the set. When we write

$$\mathbb{Z}^+ = \{1, 2, 3, \ldots\}$$

it looks like we're giving an extensional definition, but there is an "ellipsis" ...: what does this mean? The only honest answer to give now is that the ellipsis

stands for "and so on" and is thus a shorthand for the *intentional* concept of a positive whole number. Which is a fancy way of saying that I am assuming that you are familiar with the concept of a positive whole number and I am just referring to it, rather than giving some kind of precise, comprehensive description of it.

Thus the intentional description of a set is as the collection of objects satisfying a certain property. This description however must be taken with a grain of salt: for any set S there is a corresponding property of objects...namely the property of being in that set! Thus being an element of $\{17, 2023, \frac{7}{4}, \pi, \text{Batman}\}$ defines a property, although in the everyday sense there is certainly no evident rule that is being used to form this set. Again, think of a set as *any* collection of objects; the difficulties we have in describing or specifying a set – especially, an infinite set – are "our problem." They do not restrict the notion of a set.

1.4. Pure Sets.

EXAMPLE 1.15. *Here are some more examples of sets:*

a) $\{\varnothing\}$.
b) $\{\varnothing, \{\varnothing\}\}$.
c) $\{\varnothing, 6, \{\varnothing, \{\varnothing\}\}\}$.
d) $\{\varnothing, \{\varnothing\}, \{\varnothing, \{\varnothing\}\}, \{\varnothing, \{\varnothing\}, \{\varnothing, \{\varnothing\}\}\}\}$.

The sets above have a new feature: the elements are themselves sets! This is absolutely permissible. While we have not given a definition of an object, a set absolutely qualifies. Starting with the empty set and using our extensional method in a recursive way, we can swiftly build a large family of sets...of a sort which is actually a bit confusing and needs to be thought about carefully. Thus for instance, the sets $\varnothing$ and $\{\varnothing\}$ are certainly *not* equal: the first set has zero elements and the second set has one element, which happens to be the set which has zero elements. In other terms: we must distinguish a bag that is empty from a bag which contains, precisely, an empty bag. Part b) shows how this madness[3] can be continued. You should think carefully about the difference between the sets in parts c) and d): the set in part c) has some elements that are sets and some elements that are numbers. It also has 3 elements. Every element of the set in part d) is itself a set, and there are 4 elements.

We call a set **pure** if all its elements are sets. Although I will not try to justify this, in fact all of mathematics can be done only with pure sets. This means that everything in sight can be defined to be a set of some kind. So for instance numbers like 0 and 1 would have to be defined to be sets. I will not say anything more about this now: if this interests you, you might want to think of a reasonable definition for 0, 1, 2,... in terms of the empty set and lots of brackets.[4] If this troubles you: never mind, we move on!

2. Subsets

Let S and T be sets. We say that S is a **subset** of T if every element of S is also an element of T. Otherwise put, for all objects x, if $x \in S$ then also $x \in T$. The

[3]Not really.
[4]This can be done in more than one way.

symbol for this is

$$S \subseteq T.$$

It is useful to have vocabulary to describe $S \subseteq T$ "from T's perspective." And we do: if $S \subseteq T$, we say that T **contains** S. However this comes with a...**WARNING!!!** If $x \in S$, then we often say "S contains x." However, if $S \subseteq T$ we also say "T contains S." So if the object x happens to be a set, then saying "S contains x" is ambiguous: it could mean $x \in S$ or also $x \subseteq S$. These need not be the same thing! Thus we should not say "S contains x" when x is a set unless the context makes completely clear what is intended; if necessary we could say "S contains x as an element" to mean $x \in S$.

EXAMPLE 1.16. *Let $T = \{2, \{3\}, 4, \{4\}\}$. Then:*
- *T contains 2 as an element: this means $2 \in S$.*
- T does not *contain 2 as a subset: indeed, 2 is not even a set.*
- *T contains $\{2\}$ as a subset.*
- T does not *contain 3 as an element.*
- *T contains $\{3\}$ as an element, but not as a subset. (For any set X and any object a, X contains $\{a\}$ as a subset exactly when X contains a as an element.)*
- *T contains 4 as an element.*
- *T contains $\{4\}$* both *as an element and as a subset.*

A subset S of T is **proper** if $S \neq T$: every element of S is an element of T, but at least one element of T is not an element of S. We denote this by $S \subsetneq T$.

REMARK 1.17. *For real numbers x and y, if x is less than y we write $x < y$ rather than the more complicated $x \lneq y$. This suggests that if S is a proper subset of T we ought to write $S \subset T$. This notation is used in this way in some undergraduate texts, but **beware**: in mathematics as a whole it is much more common to use $S \subset T$ to mean merely that S is a subset of T, i.e., what we are here denoting by $S \subseteq T$. I will try to use the notation $S \subseteq T$ in this course, but I think it is likely that I will sometimes slip and write $S \subset T$: by this I mean $S \subseteq T$,* not *$S \subsetneq T$.*

EXAMPLE 1.18. *With regard to our previously defined sets of numbers, we have*

$$\mathbb{Z}^+ \subsetneq \mathbb{N} \subsetneq \mathbb{Z} \subsetneq \mathbb{Q} \subsetneq \mathbb{R} \subsetneq \mathbb{C}.$$

*The complex numbers are not "the end of the line" in any set-theoretic sense: we could take for instance the set of things which are either complex numbers or lines in the plane, and that would be bigger. There are also "number systems" that extend the complex numbers – e.g. there is something called the **quaternions** – but they are not as ubiquitous as the number systems we have given above.*

PROPOSITION 1.19. *For sets S and T, we have $S = T$ precisely when $S \subseteq T$ and $T \subseteq S$.*

PROOF. If $S = T$, then they have exactly the same elements. So every element of S is also an element of T, and every element of T is also an element of S.

Conversely, if $S \neq T$ then the two sets *do not* have exactly the same elements. That means that there is some object x such that either ($x \in S$ and $x \notin T$) or ($x \in T$ and $x \notin S$). (Both cannot hold for the same object x; but there may be one object x for which the former holds and another object y for which the latter holds.) But if there is x such that $x \in S$ and $x \notin T$ then S is not a subset of T, while if there is x such that $x \in T$ and $x \notin S$ then T is not a subset of X. □

Remark: This is our first instance of a mathematical **statement** and its **proof**. This is the sort of thing we will spend most of the course studying, after first laying down some fundamentals of sets (this Chapter) and logic (Chapter 2). So it might be more "procedurally correct" not to have proofs in this text until we nail down all the rules of logic and proof. We are not going to do this, for several reasons:
(i) The idea that a mathematical assertion should if possible be followed by an argument explaining why it is true cannot be new to you (right?!?). Even if we remove the word "mathematical," to try to understand why things are true is surely a pillar of learning and schooling. So it is not as though by giving a proof we are doing something so crazy and unfamiliar (right?!?).
(ii) Covering basic facts about sets without proofs would be a waste of time. Proofs are a route to understanding, and the goal of this course is to increase our mathematical understanding in as basic, holistic a way as possible.
(iiii) Not all proofs are equally difficult. On the contrary, proofs range from being simple enough to appear in everyday life to being so profoundly difficult that specialist mathematicians spend years trying to understand them. The proof of Proposition 1.19 is certainly not so bad, and moreover you can *use it* in a straigthforward way to give proofs of your own. If you are asked to show that two sets are equal, you should expect to show that an arbitrary element of the first set is also an element of the second set and then that an arbitrary element of the second set if also an element of the first set. Not so bad!

3. Power Sets

For a set X, the **power set** of X is the set of all subsets of X. We denote the power set of X by 2^X.

(This is a standard notation, but not *the most* standard. Another very common notation for the power set of X is $\mathcal{P}(X)$.)

EXAMPLE 1.20. *Some Small Power Sets:*

0) *The set* $\varnothing$ *has* 0 *elements. Its power set is* $2^{\varnothing} = \{\varnothing\}$*, which has* $1 = 2^0$ *elements.*
1) *The set* $[1] = \{1\}$ *has* 1 *element. Its power set is* $2^{[1]} = \{\varnothing, \{1\}\}$*, which has* $2 = 2^1$ *elements.*
2) *The set* $[2] = \{1, 2\}$ *has* 2 *elements. Its power set is* $\{\varnothing, \{1\}, \{2\}, \{1, 2\}\}$*, which has* $4 = 2^2$ *elements.*

PROPOSITION 1.21. *Let S be a finite set of cardinality n. Then the power set 2^S is finite of cardinality 2^n.*

PROOF. A finite set of cardinality n arises from an irredundant finite list

$$\ell : x_1, \ldots, x_n.$$

To form a subset, we must *choose* whether to include x_1 or not: that's two options. Then, independently, we choose whether to include x_2 or not: two more options. And so forth: all in all, we get a subset precisely by deciding, independently, whether to include or exclude each of the n elements. This gives us $2 \cdots 2$ (n times) $= 2^n$ options altogether. □

Proposition 1.21 gives some justification for our notation: for a finite set S, we have

$$\#2^S = 2^{\#S}.$$

Moving on, we observe that for sets S and T we have $T \subseteq S$ precisely when $T \in 2^S$. Thus one feature of the power set is to convert the relation $\subseteq$ to the relation $\in$.

4. Operations on Sets

We wish here to introduce some – rather familiar, I hope – operations on sets.

For sets S and T, we define their **union**

$$S \cup T$$

to be the set of all objects x which are elements of S, elements of T or both. (As we will see in the next chapter, in mathematics, the term "or" is always used inclusively.) We define their **intersection**

$$S \cap T$$

to be the set of all objects which are elements of both S and T. Two sets S and T are **disjoint** if $S \cap T = \varnothing$; i.e., they have no objects in common.

For sets S and T, we define their **set-theoretic difference**

$$S \setminus T = \{x \mid x \in S \text{ and } x \notin T\}.$$

If we are only considering subsets of a fixed set X, then for $Y \subseteq X$ we define its **complement** Y^c to be $X \setminus Y$.

EXAMPLE 1.22. *Let $X = \mathbb{Z}$, the integers. Let E be the set of even integers, i.e., integers of the form $2n$ for $n \in \mathbb{Z}$. Let O be the subset of odd integers, i.e., integers of the form $2n+1$ for $n \in \mathbb{Z}$. Then:*

a) *We have $E \cap O = \varnothing$: that is, no integer is both even and odd. Indeed, if $2m = x = 2n+1$, then $1 = 2(m-n)$, and thus $m - n = \frac{1}{2}$. But that's ridiculous: if m, n are integers, so is $m - n$, and $\frac{1}{2} \notin \mathbb{Z}$.*

b) *We have $E \cup O = \mathbb{Z}$. First note that if $x \in E$ then $x = 2m$, so $-x = -2m = 2(-m) \in E$; similarly if $x \in O$ then $x = 2n+1$, so $-x = -2n-1 = -2n-2+2-1 = 2(-n-1)+1 \in O$. Moreover $0 \in E$ and $1 \in O$, so it is enough to show that every integer $n \geq 2$ is either even or odd. The key observation is now that if for any $k \in \mathbb{Z}^+$, if $x - 2k \in E$ then $x \in E$, and if $x - 2k \in O$ then $x \in O$. Now consider $x - 2$. Since $x \geq 2$, $x - 2 \geq 0$. If $x - 2 \in \{0, 1\}$, then $x - 2$ is either even or odd, so x is either even or odd. Otherwise $x - 2 \geq 2$, so consider $x - 4$. We may continue in this way: in fact, there is a unique positive integer k such that $x - 2k \in \{0, 1\}$: if we keep subtracting 2, then eventually we will get something negative, and if we add back 2 then we must have either 0 or 1. This shows what we want.*

c) *Taking complements with respect to the fixed set X, we have $O^c = E$ and $E^c = O$. We say that E and O are complementary subsets of the integers.*

PROPOSITION 1.23 (DeMorgan's Laws for Sets). *Let A and B be subsets of a fixed set X. Then:*

a) *We have $(A \cup B)^c = A^c \cap B^c$.*

b) *We have* $(A \cap B)^c = A^c \cup B^c$.

PROOF. First a remark: I have set you up to expect to use Proposition 1.19: that is, to prove part a) we should show first that every element of $(A \cup B)^c$ lies in $A^c \cap B^c$ and second that every element of $A^c \cap B^c$ lies in $(A \cup B)^c$. This will certainly work: let's show the first one: if $x \in (A \cup B)^c$ then x *does not* lie in $A \cup B$, so x lies in A^c – if not, x lies in A, hence also in $A \cup B$ – and x lies in B^c – again, if not, x lies in B, hence also in $A \cup B$, so x lies in $A^c \cap B^c$.

The reason we will not do this is that in this case it turns out to be twice as much work as needed: in fact, we can just rephrase the condition of lying in $(A \cup B)^c$ to see that it is the same condition as lying in $A^c \cap B^c$, and similarly for part b). However, after you read this proof I encourage you to hide it and solve it by showing in each case the two containments: that's good practice.
a) Since $A \cup B$ consists of all elements of X that lie in A or in B (or both), the complement $(A \cup B)^c$ consists of all elements of X that lie in neither A nor B. That is, it consists precisely of elements that do not lie in A and do not lie in B, hence of elements that lie in the complement of A and in the complement of B.
b) Since $A \cap B$ consists of all elements of X lying in both A and B, the complement $(A \cap B)^c$ consists of all elements of X that do not lie in both A and B. An element of X does not lie in both A and B precisely when it does not lie in A or it does not lie in B (or both), i.e., we get precisely the elements of $A^c \cup B^c$. □

Although these things can be converted to "word problems" and sounded out with little trouble, many people prefer a more visual approach. For this **Venn diagrams** are useful. A Venn diagram for two subsets A and B of a fixed set X consists of a large rectangle (say) representing X and within it two smaller, overlapping circles, representing A and B. Notice that this divides the rectangle X into four regions:

- $A \cap B$ is the common intersection of the two circles.
- $A \setminus B$ is the part of A that lies outside B.
- $B \setminus A$ is the part of B that lies outside A.
- $(A \cup B)^c$ is the part of X that lies outside both A and B.

PROPOSITION 1.24. *(Distributive Laws) Let A, B, C be sets. Then:*
a) We have $(A \cup B) \cap C = (A \cap C) \cup (B \cap C)$.
b) We have $(A \cap B) \cup C = (A \cup C) \cap (B \cup C)$.
That is: intersection distributes over union and union distributes over intersection.

PROOF. a) *Now* we will use the technique of showing that two sets are equal by showing that each contains the other. Suppose $x \in (A \cup B) \cap C$. Then $x \in C$ and $x \in A \cup B$, so either $x \in A$ or $x \in B$. If $x \in A$ then $x \in A \cap C$, whereas if $x \in B$ then $x \in B \cap C$, so either way $x \in (A \cap C) \cup (B \cap C)$. Thus

$$(A \cup B) \cap C \subseteq (A \cap C) \cup (B \cap C).$$

Conversely, suppose $x \in (A \cap C) \cup (B \cap C)$. Then $x \in A \cap C$ or $x \in B \cap C$. Since both A and B are subsets of $A \cup B$, either way we have $x \in (A \cup B) \cap C$, so

$$(A \cap C) \cup (B \cap C) \subseteq (A \cup B) \cap C.$$

b) This is similar; I leave it to you as an exercise. □

REMARK 1.25. *The more familiar distributive law is that multiplication – say of complex numbers – distributes over addition: for all $x, y, z \in \mathbb{C}$ we have*

$$(x + y) \cdot z = (x \cdot z) + (y \cdot z).$$

It is interesting that in the set theoretic context each of union and intersection distributes over the other. This is a pleasant symmetry that is not present in the case of numbers: for most $x, y, z \in \mathbb{C}$ we do not have $(x \cdot y) + z = (x + z) \cdot (y + z)$. For instance try it with $x = y = z = 1$.

5. Indexed Families of Sets

We can define unions and intersections for more than two sets. Let $A_1, \ldots, A_n$ be subsets of a fixed set X. Then we define $A_1 \cap \ldots \cap A_n$ to be the set of all objects that lie in all of the A_i's, and we define $A_1 \cup \ldots \cup A_n$ to be the set of all objects that lie in at least one of the A_i's.

There is another, rather more sophisticated perspective to take on the expression $A_1, \ldots, A_n$: namely that it is a **family of sets indexed by** the set $[n] = \{1, \ldots, n\}$. Really this is a kind of **function** (although functions will not be formally defined and considered until much later in the course), by which I mean that it is an assignment of a set A_i to each $i \in \{1, \ldots, n\}$: we write

$$1 \mapsto A_1, 2 \mapsto A_2, \ldots, n \mapsto A_n.$$

More generally, a **family of sets indexed by a set I** is just a nonempty set I and an assignment of each $i \in I$ a set A_i. This is a construction that comes up widely in mathematics. For now we will just say that it makes sense to take unions and intersections over an indexed family of sets: we define the **union**

$$\bigcup_{i \in I} A_i$$

to be the set of x that lie in A_i for at least one $i \in I$ and the **intersection**

$$\bigcap_{i \in I} A_i$$

to be the elements x that lie in A_i for all $i \in I$. Thus the union is the set of elements lying in *some* set of the family and the intersection is the set of elements lying in *every* set in the family. This generalizes the kind of union and intersection we studied before when I has two elements or has finitely many elements.

EXAMPLE 1.26. *Some Examples of Indexed Families of Sets:*

a) *If $I = \mathbb{Z}^+$ then we have a* ***sequence of sets***

$$A_1, A_2, \ldots.$$

b) *Suppose $I = \mathbb{Z}$ and for all $n \in I$ we put $A_n = \{n\}$. Then*

$$\bigcup_{n \in \mathbb{Z}} \{n\} = \mathbb{Z}$$

and

$$\bigcap_{n \in \mathbb{Z}} \{n\} = \varnothing.$$

c) *More generally, let I be any set that contains more than one element, and for $i \in I$ put $A_i = \{i\}$. Then*

$$\bigcup_{i \in I} A_i = I$$

and

$$\bigcap_{i \in I} A_i = \varnothing.$$

(Why is it important here that I have more than one element?)

EXAMPLE 1.27. *Monotone Sequences of Sets:*

a) *For $n \in \mathbb{Z}^+$ we put*

$$A_n = [-n, n] \subseteq \mathbb{R}.$$

Then we have

$$A_1 = [-1, 1] \subseteq A_2 = [-2, 2] \subseteq \ldots \subseteq A_n = [-n, n] \subseteq \ldots.$$

In this case we have

$$\bigcup_{n \in \mathbb{Z}^+} A_n = \mathbb{R},$$

since every real number has absolute value less than n for some integer n. More easily, we have

$$\bigcap_{n \in \mathbb{Z}^+} A_n = [-1, 1].$$

*This sequence of sets has the interesting property that $A_n \subseteq A_{n+1}$ for all n. For any such sequence of sets, the common intersection of all the sets is just A_1. One might call this an **increasing sequence of sets**.*

b) *For $n \in \mathbb{Z}^+$ we put*

$$B_n = (\frac{-1-n}{n}, \frac{n+1}{n}) \subseteq \mathbb{R}.$$

Thus we have

$$B_1 = (-2, 2) \supseteq B_2 = (\frac{-3}{2}, \frac{3}{2}) \supseteq B_3 = (\frac{-4}{3}, \frac{4}{3}) \supseteq \ldots \supseteq B_n \supseteq \ldots$$

This time we have

$$\bigcup_{n \in \mathbb{Z}^+} B_n = B_1 = (-2, 2)$$

and the more interesting case is

$$\bigcap_{n \in \mathbb{Z}^+} B_n = [-1, 1].$$

*Thus the intersection of an infinite sequence of open intervals turns out to be a closed interval. This sequence has the interesting property that $B_n \supseteq B_{n+1}$ for all n: we call this a **decreasing sequence of sets** or a **nested sequence of sets**. For any nested sequence of sets, the union is the first set B_1: Exercise 1.17a).*

A family $\{A_i\}_{i \in I}$ of sets is **pairwise disjoint** if for all $i \neq j$ we have $A_i \cap A_j = \varnothing$.

EXAMPLE 1.28. *Let $I = \mathbb{Z}$, and for all $n \in \mathbb{Z}$ let $A_n \coloneqq \mathbb{R}$. This is a family of sets indexed by $\mathbb{Z}$ each element of which is the set of real numbers. This example illustrates that an indexed family of sets is more than just a* set *of sets; it consists of an assignment of a set to each element of an index set I. We are allowed to assign the same set to two different elements of I.*

6. Partitions

Let X be a nonempty set. A **partition** of X is, roughly, an exhaustive division of X into nonoverlapping nonempty pieces. More precisely, a partition of X is a set $\mathcal{P}$ of subsets of X satisfying all of the following properties:

(P1) $\bigcup_{S \in \mathcal{P}} S = X$.
(P2) For distinct elements $S \neq T$ in $\mathcal{P}$, we have $S \cap T = \varnothing$.
(P3) If $S \in \mathcal{P}$ then $S \neq \varnothing$.

EXAMPLE 1.29. *Let $X = [5] = \{1, 2, 3, 4, 5\}$. Then:*

a) *The set $\mathcal{P}_1 = \{\{1, 3\}, \{2\}, \{4, 5\}\}$ is a partition of X.*
b) *The set $\mathcal{P}_2 = \{\{1, 2, 3\}, \{4\}\}$ is* not *a partition of X: $5 \in X$, but 5 is not an element of any element of $\mathcal{P}_2$, so (P1) fails. However, (P2) and (P3) both hold.*
c) *The set $\mathcal{P}_3 = \{\{1, 2, 3\}, \{3, 4, 5\}\}$ is* not *a partition of X: $\{1, 2, 3\}$ and $\{3, 4, 5\}$ are not disjoint, so (P2) fails. However, (P1) and (P3) both hold.*
d) *The set $\mathcal{P}_4 = \{\{1, 2, 3, 4, 5\}, \varnothing\}$ is not a partition of X because it contains the empty set, so (P3) fails. However, (P1) and (P2) both hold.*

EXAMPLE 1.30. *More Partitions:*

a) *Let $X = [1] = \{1\}$. There is exactly one partition, $\mathcal{P} = \{X\}$.*
b) *Let $X = [2] = \{1, 2\}$. There are two partitions on X,*
$$\mathcal{P}_1 = \{\{1, 2\}\}, \ \mathcal{P}_2 = \{\{1\}, \{2\}\}.$$
c) *Let X be any set with more than one element. Then the analogues of the above partitions exist: namely there is the* ***trivial partition*** *(or* ***indiscrete partition****)*
$$\mathcal{P}_t = \{X\}$$
and the ***discrete partition***
$$\mathcal{P}_D = \{\{x\} \mid x \in X\}.$$
I hope the notation does not distract from the simplicity of what's happening here: in the trivial partition we "break X up into one piece" (or in another words, we don't break it up at all). In the discrete partition we "break X up into the largest possible number of pieces": one-element sets.
d) *If X has more than two elements then there are partitions on X other than the trivial partition and the discrete partition. For instance on $X = [3] = \{1, 2, 3\}$ there are three more:*
$$\{\{1\}, \{2, 3, \}\}, \ \{\{2\}, \{1, 3\}, \}, \ \{\{3\}, \{1, 2\}\}.$$
These three partitions share a common feature: namely for each positive integer n, they have the same number of pieces of size n. If we count

partitions on a set altogether, we find ourselves counting many similar-looking decompositions that are labelled differently, as above. It is a classic number theory problem to count partitions on $[n]$ up to the various sizes of the pieces*. In other words, in this sense* 3 *has* 3 *partitions:*

$$3 = 3 = 2 + 1 = 1 + 1 + 1.$$

Similarly, in this sense 4 *has* 5 *partitions:*

$$4 = 4 = 3 + 1 = 2 + 2 = 2 + 1 + 1 = 1 + 1 + 1 + 1.$$

For a positive integer n*, define* $P(n)$ *to be the number of partitions of* n *in this sense, so what we've seen so far is*

$$P(1) = 1,\ P(2) = 2,\ P(3) = 3,\ P(4) = 5.$$

There is an enormous amount of deep 20th century mathematics studying the asymptotic behavior of the partition function $P(n)$*: in other words, how quickly it grows as a function of* n*.*

EXAMPLE 1.31. *Let* E *be the set of even integers, and let* O *be the set of odd integers. Then* $\{E, O\}$ *is a partition of* $\mathbb{Z}$*. This serves to illustrate why partitions of sets are important: one can think of elements of the same set in a partition as sharing a common* property*, in this case the property that they are both even (if they are both in* E*) or that they are both odd (if they are both in* O*).*

Later we will see that conversely, a certain type of property of objects of a set X *– called an* ***equivalence relation*** *– determines a partition of* X *and that conversely every partition on* X *arises from an equivalence relation on* X*.*

7. Cartesian Products

Let X and Y be sets. Then the **Cartesian product** $X \times Y$ is the set of ordered pairs (x, y) where $x \in X$ and $y \in Y$.

EXAMPLE 1.32. *The main example – the trope-namer, in the terminology of the popular website* `https://tvtropes.org/` *– is* $\mathbb{R}^2 = \mathbb{R} \times \mathbb{R}$*, the Cartesian plane.*

From an operational perspective, ordered pairs are easy: for every $x \in X$ and $y \in Y$ there is an ordered pair (x, y), and for $x_1, x_2 \in X$ and $y_1, y_2 \in Y$ we have

$$(x_1, y_1) = (x_2, y_2) \iff x_1 = x_2 \text{ and } y_1 = y_2.$$

In other words, just as two points P_1 and P_2 in the plane coincide if the x-coordinate of P_1 is equal to the x-coordinate of P_2 and the y-coordinate of P_1 is equal to the y-coordinate of P_2, two ordered pairs are equal if and only if their first components are equal to each other and their second components are equal to each other.

Nevertheless, one may ask what an ordered pair (x, y) "really is." This question didn't occur to me until after I got my PhD in mathematics, so this discussion is certainly not essential, but still...one may ask. For instance, if (as is most common among mathematicians who think seriously about set theory) one pursues a "pure set theory" in which every element of a set is again a set, if x and y are sets then we don't want (x, y) to be some new kind of object: it needs to be some set defined in terms of x and y.

Kuratowski suggested the following definition:

$$(x, y) \coloneqq \{\{x\}, \{x, y\}\}.$$

Well, that is certainly a set. Our only other requirement is the just mentioned one that we want $(x_1, y_1) = (x_2, y_2)$ if and only if $x_1 = x_2$ and $y_1 = y_2$; you are asked to show this in Exercise 1.20a).

More generally, if $X_1, \ldots, X_n$ are sets then the Cartesian product $X_1 \times \ldots \times X_n$ is the set of all ordered n-tuples $(x_1, \ldots, x_n)$ with $x_1 \in X_1, \ldots, x_n \in X_n$.[5]

Please do not take Kurtatowski's definition of the Cartesian product too seriously. The main reason I introduce it is that checking that it works makes for a good exercise for a student just starting to learn about sets. But it really doesn't matter what kind of object (x, y) is; what matters is when two ordered pairs are equal, and the answer is that $(x_1, x_2) = (y_1, y_2)$ precisely when $x_1 = y_1$ and $x_2 = y_2$. Similarly, two ordered n-tuples $(x_1, \ldots, x_n)$ and $(y_1, \ldots, y_n)$ are equal precisely when $x_1 = y_1$, $x_2 = y_2$,...,$x_n = y_n$.

In the same vein of asking whether everything can be a set, we can ask what kind of set a finite list is. If ℓ is a finite list of length n with associated set S, then we may think of it as element of the n-fold Cartesian product

$$S^n \coloneqq S \times S \times \ldots \times S.$$

This *does not* make for a good exercise: the correspondence is just

$$\ell : x_1, \ldots, x_n \mapsto (x_1, \ldots, x_n).$$

There is just one minor point here: every finite list of length n is an element of some n-fold Cartesian product, but not every finite list of length n is an element of the same n-fold Cartesian product (unless we believe in a set that contains *all* objects, which may sound reasonable, but trust me for now – this has some problems). To clean this up it is helpful to conisder finite lists of length n with elements drawn from a fixed set S: these are precisely the elements of S^n. In fact, if we have sets $A_1, \ldots, A_n$, then we may view $A_1 \times A_n$ as the set of finite lists $\ell : x_1, \ldots, x_n$ of length n where for all $1 \leq i \leq n$, the ith element X_i is drawn from the set A_i. This is indeed very useful, and we will return to it soon.

Finally, if I is a nonempty set and $\{X_i\}_{i \in I}$ is an indexed family of sets, then we can consider the Cartesian product $\prod_{i \in I} X_i$. An element of this is an object $\{x_i\}_{i \in I}$: that is, for each $i \in I$, an element $x_i \in X_i$.

8. Exercises

EXERCISE 1.1. *Each of the following sets is defined intensionally – i.e., given as the set of elements satisfying some property. Give extensional definitions: i.e., list every element of the set.*

a) $\{x \in \mathbb{N} \mid -2 \leq x \leq 5\}$.
b) $\{x \in \mathbb{Z} \mid -2 \leq x \leq 5\}$.
c) $\{x \in \mathbb{R} \mid x^2 + 5x^2 = -6x\}$.

[5]One can give a Kuratowski style definition of this n-fold Cartesian product as well, but we choose not to.

EXERCISE 1.2. *Each of the following infinite sets is is defined with a ..., which is sort of an "implicit extensional" definition. Give an explicit intensional definition. E.g. given* $\{0, 2, 4, 6, 8 \ldots\}$ *you could write* $\{2x \mid x \in \mathbb{N}\}$.

Note that there is certainly more than one correct answer, but please try to find the simplest answer you can.

a) $\{0, 4, 16, 36, 64, \ldots\}$.
b) $\{1, 2, 4, 8, 16, \ldots\}$.
c) $\{-8, -3, 2, 7, 12, 17, \ldots\}$.

EXERCISE 1.3. *Let* $S \coloneqq \{1, 2\}$.

a) *Find all finite lists of length* 2 *with associated set* S.
b) *Find all finite lists of length* 3 *with associated set* S.
c) *Find all finite lists of length* 4 *with associated set* S.

EXERCISE 1.4. *Let* $\ell : x_1, \ldots, x_n$ *be a finite list, and let* $S \coloneqq \{x_1, \ldots, x_n\}$ *be the associated finite set.*

a) *Show that if* ℓ *is irredundant, then* $\#S = n$.
b) *Show that if the list has at least one repetition, then* $\#S < n$.

EXERCISE 1.5. *Let* S *be a nonempty finite set, of cardinality* n.

a) *Show that for* $k \in \mathbb{N}$ *there is a finite list of length* k *with associated finite set* S *if and only if* $k \geq n$.
b) *Deduce: there are infinitely many finite lists with associated set* S.

EXERCISE 1.6. *Let* T *be a finite set, and let* $S \subseteq T$. *Show:* $\#S \leq \#T$. *(Suggestion: start with an irredundant finite list* ℓ_T *with associated set* T*, hence of length* $\#T$. *What do you have to do to* ℓ_T *to get an analogous finite list for* S*?)*

EXERCISE 1.7. *Let* S *be a set.*

a) *Show:* $\varnothing \subseteq S$.
b) *Show that* $\varnothing \subsetneq S$ *precisely when* $S \neq \varnothing$.

EXERCISE 1.8. *Let* X *and* Y *be sets.*

a) *Suppose* $X \subseteq Y$. *Show:* $X \setminus Y = \varnothing$.
b) *Suppose* $X \setminus Y = \varnothing$. *Show:* $X \subseteq Y$.

EXERCISE 1.9. *Let* X *and* Y *be sets.*

a) *Suppose that* $X \subseteq Y$. *Show:* $2^X \subseteq 2^Y$.
b) *Suppose that* $2^X \subseteq 2^Y$. *Show:* $X \subseteq Y$.
c) *Show: if* $X = Y$*, then* $2^X = 2^Y$.
d) *Show: if* $2^X = 2^Y$*, then* $X = Y$.

EXERCISE 1.10. *Let* $X = \{2n \mid n \in \mathbb{N}\}$*, and let* Y *be the set of prime numbers.*

a) *Find* $X \setminus Y$.
b) *Find* $Y \setminus X$.

EXERCISE 1.11. *Use Venn diagrams to prove DeMorgan's Laws for Sets.*

EXERCISE 1.12. *A Venn diagram for three subsets* A, B, C *of a fixed set* X *consists of three circles inside a rectangle* X *positioned so as to divide* X *into* $8 = 2^3$ *regions in all (this comes from being in* A *vs. not being in* A*, being in* B *vs. not being in* B*, and being in* C *vs. not being in* C*). This is no problem to achieve:*

just draw three circles with the same radius and noncolinear centers sufficiently close together. Use this kind of Venn diagram to prove the distributive laws.

EXERCISE 1.13. *Show that DeMorgan's Laws extend to n sets (for any $n \geq 2$):*

$$(A_1 \cup \ldots \cup A_n)^c = A_1^c \cap \ldots \cap A_n^c$$

and

$$(A_1 \cap \ldots \cap A_n)^c = A_1^c \cup \ldots \cup A_n^c.$$

EXERCISE 1.14. *State and prove an extension of the distributive laws to n sets.*

EXERCISE 1.15. *Are there Venn diagrams for n sets with $n \geq 4$?*
(Hint: yes, but you cannot use circles.)

EXERCISE 1.16. *Let I be a nonempty set. A family of sets $\{A_i\}_{i \in I}$ is* ***mutually disjoint*** *if $\bigcap_{i \in I} A_i = \varnothing$ and is* ***pairwise disjoint*** *if for all $i \neq j$ in I, we have $A_i \cap A_j = \varnothing$. A family is* ***pairwise intersecting*** *if for all $i \neq j$ in I, we have $A_i \cap A_j \neq \varnothing$.*

a) *Give an example of a family of sets $\{A_i\}_{i \in I}$ that is neither pairwise disjoint nor pairwise intersecting.*
b) *Suppose that I contains more than one element. Show: if $\{A_i\}_{i \in I}$ is pairwise disjoint, then it is mutually disjoint.*
c) *Find a family of sets A_1, A_2, A_3 that is mutually disjoint but not pairwise disjoint.*

EXERCISE 1.17. *Let $B_1 \supseteq B_2 \supseteq \ldots \supseteq B_n \supset \ldots$ be a nested sequence of sets.*

a) *Show: $\bigcup_{n=1}^{\infty} B_n = B_1$.*
b) *Show by example that we may have $\bigcap_{n=1}^{\infty} B_n = \varnothing$ even if B_n is nonempty for all n.*

EXERCISE 1.18.

a) *Let $n \geq 2$, and let $A_1 \supseteq A_2 \supseteq \ldots \supseteq A_n$ be a finite nested sequence of sets. Show that all of the following are equivalent (i.e., each implies the others):*
   (i) *The family $\{A_i\}_{i=1}^n$ is pairwise intersecting.*
   (ii) *The family is* not *mutually disjoint: $\bigcap_{i=1}^n A_i \neq \varnothing$.*
   (iii) *The set A_n is nonempty.*
b) *Find a nested infinite sequence of subsets of $\mathbb{R}$*

$$A_1 \supseteq A_2 \supseteq \ldots \supseteq A_n \supseteq \ldots$$

that is pairwise intersecting but mutually disjoint: $\bigcap_{n=1}^{\infty} A_n = \varnothing$.

EXERCISE 1.19.

a) *Write down all partitions of the empty set. (There is 1.)*
b) *Write down all partitions of $[1] = \{1\}$. (There is 1.)*
c) *Write down all partitions of $[2] = \{1, 2\}$. (There are 2.)*
d) *Write down all partitions of $[3] = \{1, 2, 3\}$. (There are 5.)*

EXERCISE 1.20 (Defining Ordered Pairs).

a) *Above we mentioned Kuratowksi's definition of an ordered pair:*

$$(a, b) := \{\{a\}, \{a, b\}\}.$$

Show that for all objects a_1, a_2, b_1, b_2 we have

$$(a_1, b_1) = (a_2, b_2) \iff a_1 = a_2 \text{ and } b_1 = b_2$$

and thus Kuratowski's definition meets our only requirement.

b) *Do you have any comments or reservations about this definition. (For instance: is it the only possible definition, or can you think of other reasonable ones? Do you find this definition helpful?) Discuss.*

EXERCISE 1.21. *Let A, B, C and D be sets.*

a) *Suppose $A \subseteq C$ and $B \subseteq D$. Show: $A \times B \subseteq C \times D$.*

b) *Suppose that A and B are nonempty and that $A \times B \subseteq C \times D$. Show: $A \subseteq C$ and $B \subseteq D$.*

EXERCISE 1.22. *Let A, B and C be sets.*

a) *Show:*
$$(A \setminus B) \times C = (A \times C) \setminus (B \times C).$$

b) *Show:*
$$A \times (B \cap C) = (A \times B) \cap (A \times C).$$

c) *Show:*
$$A \times (B \cup C) = (A \times B) \cup (A \times C).$$

EXERCISE 1.23. *Let A and B be sets.*

a) *Show: if $A = B$ then $A \times B = B \times A$. (Yes, it is as easy as it looks.)*

b) *Show: if A and B are nonempty and $A \times B = B \times A$, then $A = B$.*

c) *Show: if either A or B is empty then $A \times B = \varnothing = B \times A$.*

CHAPTER 2

Logic

1. Statements

A **statement** is an assertion that is unambiguously true or false, and not both.

EXAMPLE 2.1. *All of the following are statements:*

a) *The smallest prime number is* 2.
b) *Every square is a rectangle.*
c) *There is no largest real number.*

In fact, they are all true.

EXAMPLE 2.2. *All of the following are statements:*

a) *The number* 57 *is prime.*
b) *Every rectangle is a square.*
c) $7^3 > 3^7$.

In fact, they are all false. (We have $57 = 3 \cdot 19$. *For instance the shape of the American flag is a rectangle that is not a square. And* $7^3 = 343 < 2187 = 3^7$.*)*

EXAMPLE 2.3. *All of the following are* not *statements:*

a) *Are we going to have Thai food tonight?*
b) *The number* 2023 *is large.*
c) *The integer* x *is prime.*

In part a), we have a question, which is clearly not a statement. In part b), we have a statement that is too fuzzy/subjective to assign a clear truth value...unless "large" is a technical term that has previously been defined.[1] *The failure of part c) to be a statement is of most relevance to us: for any* particular *integer* x*, asserting that* x *is prime is a statement. The problem is that here what appears is an* unspecified *integer* x*, and clearly the truth or falsity depends on which integer* x *is.*

Here we are not in any way taking on the task of determing what kind of syntactic constructions do or not give rise to a statement. When it comes to ordinary English, your own prior education and training far outstrips anything we could say here. In its further development logic studies **formal languages** in which one specifies rules that determine which finite strings of certain symbols constitute statements. We won't have the need for this here. For our purposes it is sufficient to imagine a supply of "primitive statements" $\{P_i \mid i \in I\}$ indexed by some set I, such that for

[1] I remember from my undergraduate career a True/False question of the form: "The largest positive integer n such that [some linear algebra assertion involving n] is true is pretty small." It turned out that the largest n that had the property in question is 1. I correctly figured this out, and answered **True**. Another – really excellent – student wrote **False**. He claimed that he also figured out that the largest such n is 1, but thought that 1 wasn't that small!

each $i \in I$ we can entertain the possibility that P_i is true and also the possibility that P_i is false and that all these possibilities can be entertained independently.

2. Logical Operations

Just as in Chapter 1 we considered operations that generate new sets from old ones, now we will do the same with statements.

The first operation we introduce is **negation**. If P is a statement, then we get a new statement $\neg P$, which we read as "the negation of P" or "not P." By definition, $\neg P$ is true exactly when P is false. So a good way of thinking of $\neg$ as something that, when applied, "toggles the truth value," much like a light switch.

To be formal about it, the following is the definition of the operation $\neg$:

P	$\neg P$
T	F
F	T

The next operation that we consider is **or**. If P and Q are statements, then we get a new statement $P \vee Q$ that is true when at least one of P and Q is true. Here is the official definition:

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

Note in particular the first column: $T \vee T = T$. This means that the logical operation $\vee$ is *inclusive*: $P \vee Q$ means P or Q (or both!). The inclusive use of "or" is completely standard throughout mathematics. In ordinary language the situation is much more complicated: sometimes the context *suggests* that an "or" is meant exclusively. E.g. if on a restaurant menu you encounter "served with a biscuit or grits" then probably you cannot order both (without paying more). On the other hand, I looked up the state of Georgia's voter identification requirements, and it contains the following passage:

"Valid employee photo ID from any branch, department, agency, or entity of the U.S. Government, Georgia, or any county, municipality, board, authority or other entity of this state."

The "or's" seem intended inclusively. For instance, suppose you show up with employee ID from the U.S. Department of Defense. The D.O.D. happens to be both a department and an agency (I just looked it up), but of course that does not invalidate this as a form of ID. Most careful writers are aware of the fuzziness inherit in the English word "or" and take pains to minimize its use when the ambiguity would lead to trouble. The use of "and/or" often appears in legal writing.[2]

[2]Strangely, "and/or" is also widely criticized in legal writing: see e.g. `https://en.wikipedia.org/wiki/And/or`. Some have criticized it for being inelegant, which seems reasonable. Others

Here is a question: *why* is $T \vee T = T$? How do we know that $T \vee T = F$ is not correct? The answer is very important: this is **our definition** of $\vee$. Definitions are whatever we say they are: there is no inherent correctness or incorrectness to them.[3] The reason that $T \vee T = T$ is no more and no less than you are listening to me now, and this is what I say that it means.[4] If you are writing a book like this, you could define $\vee$ in the exclusive sense. That might cause confusion for your students, who would be using the symbol in a way different from that of most other mathematical practitioners, but no mathematical error would be made.

There is no question that *there exists* a logical operation that applied to propositions P and Q is true if exactly one of P and Q is true. The question is what I want to call it. I will call it "exclusive or" and denote it by $\veebar$:

P	Q	$P \veebar Q$
T	T	F
T	F	T
F	T	T
F	F	F

The next logical operation we want to introduce is and: if P and Q are statements, we get a new statement $P \wedge Q$ that is true when P and Q are both true.

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

Now we introduce "if and only if": if P and Q are statements, then we say that "P if and only if Q" and write $P \iff Q$ if P and Q have the same truth value: either both are true or both are false.

P	Q	$P \iff Q$
T	T	T
T	F	F
F	T	F
F	F	T

3. Logical Equivalence, Tautologies and Contradictions

Let $P_1, \ldots, P_n$ be statements. We say that two expressions X and Y formed from these statements using logical operations are **logically equivalent** if for each of the 2^n possible combinations of truth values of $P_1, \ldots, P_n$, the two expressions have

have criticized it for being ambiguous...which I find absolutely perplexing. One begins to suspect that some lawyers simply like to argue.

[3]Please don't take this sentiment farther than intended. Some definitions will in practice lead to clarity, interesting content and so forth. Other definitions will in practice lead to confusion, trivialities or even contradiction. When you encouter a new definition you cannot **argue** with it, but you certainly can – and should! – try to understand it and figure out why it was made.

[4]" 'When I use a word,' Humpty Dumpty said in rather a scornful tone, 'it means just what I choose it to mean—neither more nor less."' – Lewis Carroll, *Through the Looking Glass* Lewis Carroll was the penname of Charles Dodgson, a mathematician. Much of his work was in logic.

the same truth value. We write this for now as $X \leftrightarrow Y$.[5]

Here are some simple first examples of logical equivalence.

EXAMPLE 2.4. *Let P and Q be statements. Then:*

a) *We have*
$$P \leftrightarrow \neg(\neg P).$$
Indeed, P is true exactly when $\neg P$ is false exactly when $\neg(\neg P)$ is true.

b) *We have*
$$(P \wedge Q) \leftrightarrow (Q \wedge P) :$$
Eeach expression is true precisely when P and Q are both true.

c) *We have*
$$(P \vee Q) \leftrightarrow (Q \vee P) :$$
Each expression is true precisely when at least one of P and Q is true.

Two logical expressions that are *equal* are certainly *logically equivalent*. The converse is not true: $\neg(\neg P)$ is a different expression from P, but they are logically equivalent. This example already shows that for many purposes logical equivalence is actually a more useful concept than equality. Consider the two statements "It is raining" and "It is not the case that it is not raining." Technically speaking they are different statements, but that just means they are expressed with different words. In all cases where we actually care about the information conveyed by the statements, they are interchangeable.

In Exercise 2.2 you are asked to establish that $\wedge$ and $\vee$ are commutative and associative, up to logical equivalence.

Here is a very useful logical equivalence.

PROPOSITION 2.5 (Logical DeMorgan's Laws). *Let P and Q be statements.*

a) *We have $\neg(P \vee Q) \leftrightarrow (\neg P) \wedge (\neg Q)$.*
b) *We have $\neg(P \wedge Q) \leftrightarrow (\neg P) \vee (\neg Q)$.*

PROOF. These and similar logical equivalences can be verified in a straightforward way: here, we actually write down all four possible combinations of the truth/falsity of P and the truth/falsity of Q and check that in each of these four cases, the first logical expression involving P and Q is true exactly when the second logical expression involving P and Q is true.
a) The following truth table establishes $\neg(P \vee Q) \leftrightarrow (\neg P) \wedge (\neg Q)$.

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

b) The following truth table establishes $\neg(P \wedge Q) \leftrightarrow (\neg P) \vee (\neg Q)$.

[5]Yes, this is similar to the previously introduced notation $X \iff Y$. The reason for this will be seen shortly.

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

□

A **tautology** is a logical expression involving statements $P_1, \ldots, P_n$ that is true for all 2^n possible truth values of the individual statements P_i. A **contradiction** is a logical expression involving statements $P_1, \ldots, P_n$ that is false for all 2^n possible truth values of the individual statements P_i.

Thus a logical expression X is a tautology precisely when its negation $\neg X$ is a contradiction.

Example 2.6.

a) *The statement $P \vee (\neg P)$ is a tautology: exactly one of P and $\neg P$ is true and thus $P \vee (\neg P)$ is always true.*

b) *The statement $P \wedge (\neg P)$ is a contradiction: exactly one of P and $\neg P$ is true and thus $P \wedge (\neg P)$ is false.*

c) *Indeed, using Proposition 2.5 we find that*

$$\neg(P \vee (\neg P)) \leftrightarrow (\neg P) \wedge (\neg\neg P) \leftrightarrow (\neg P) \wedge P \leftrightarrow P \wedge (\neg P).$$

Thus the logical expression of $P \wedge (\neg P)$ is logically equivalent to the negation of the tautology $P \vee (\neg P)$, so it is a contradiction.

Here is a simple but fundamental observation:

Proposition 2.7. *Let X and Y be logical expressions involving statements $P_1, \ldots, P_n$. Then $X \leftrightarrow Y$ holds exactly when $X \iff Y$ is a tautology.*

Proof. The assertion $X \leftrightarrow Y$ means that for each of the 2^n possible truth values of $P_1, \ldots, P_n$, the expression X is true exactly when the expression Y is true. If so, then for each of the 2^n possible truth values of $P_1, \ldots, P_n$, we have $X \iff Y$. Similarly, if $X \iff Y$ is a tautology then for each of the 2^n possible truth values of $P_1, \ldots, P_n$ the expression $X \iff Y$ is true, which means that for each of the 2^n possible truth values of $P_1, \ldots, P_n$, the expression X is true exactly when the expression Y is true. □

Proposition 2.7 shows that there is a distinction to make between $\leftrightarrow$ and $\iff$, albeit a rather fine one. Two logical expressions X and Y are either logically equivalent or they are not; thus $X \leftarrow Y$ is a **statement** that is either true or false. On the other hand $X \iff Y$ is itself a logical expression that in general may be true for some truth values of $P_1, \ldots, P_n$ and false for other truth values. However, the logical operator $\iff$ is rarely used unless X and Y are logically equivalent, so in practice this distinction does not usually need to be made.

4. Implication

We now come to the most important logical operation: implication. For statements P and Q, write $P \implies Q$ and say "P implies Q" for the operation defined as follows:

P	Q	$P \implies Q$
T	T	T
T	F	F
F	T	T
F	F	T

As mentioned above, like Humpty Dumpty I can make any definition I want. So that's what implication means *because I say so.* You can't argue!

Well....you can't argue, but you can still wonder why I made my definition or even be confused about it. For many people, the last two lines of the defining truth table are confusing: why is $P \implies Q$ true whenever P is false?

One can think of the implication $P \implies Q$ as working like a contract: if you do P, I *promise* to do Q. How can such a contract be broken? To be more concrete, suppose that P is that you have a completed Jittery Joe's punch card and Q is that you get a free drink of your choosing at Jittery Joe's. Then the whole business is encapsulated in the contract P implies Q: if you turn in your punch card, you get a free drink. What would constitute breaking this contract? The only way is that if you turn in the punchcard and they refuse to give you a free drink: $T \implies F$ is precisely what shouldn't happen. Can the contract become broken if you do not turn in your punch card? No, of course not. If you don't turn in your punch card, you might still get a free drink,[6] and if you do that certainly doesn't invalidate the contract. What if you don't turn in your punch card and don't get a free drink? Of course that's what's happening at most points in your life: that doesn't invalidate the contract either.

I find it useful to think of binary logical operations in terms of the *number* of times that they are true...in the sense of the number of T's that appear in the defining column of the table. A tautology is defined by being true four our of four times, while a contradiction is defined by being true zero out of four times. Knowing that an operator is true one, two or three times out of four doesn't determine the operator up to logical equivalence, but it is still useful information. Like $\vee$, the operation $\implies$ is true 3 out of 4 times.

Why is this useful? Well, for instance one of the mistakes that students at this level make again and again is thinking that $\neg(P \implies Q)$ comes out again as some kind of implication. But the above reasoning shows that this is not possible. Since an implication is true 3 out of 4 times, its implication is true 1 out of 4 times...hence is not an implication. In fact this reasoning puts us on the right track for a logically equivalent operation to $\neg(P \implies Q)$: what other operations do we know that are true 1 out of 4 times? The one that springs to mind is $\wedge$. This does not mean that $\neg(P \implies Q) \leftrightarrow P \wedge Q$ and in fact this isn't true: since $P \wedge Q$ is true exactly when P and Q are true, while $\neg(P \implies Q)$ is true exactly when $P \implies Q$ is false, which as just mentioned is exactly when P is true and Q is false. Aha!

PROPOSITION 2.8. *Let P and Q be statements.*

[6]As someone who has spent countless hours in various Jittery Joe's, I am here to tell you that "unearned" free drinks happen every now and again.

a) *We have* $\neg(P \implies Q) \leftrightarrow P \wedge (\neg Q)$.
b) *We have* $(P \implies Q) \leftrightarrow (\neg P) \vee Q$.

PROOF. a) This was discussed above: $\neg(P \implies Q)$ and $P \wedge (\neg Q)$ are each true when P is true and Q is false and in none of the other three cases.
b) Using Proposition 2.5 we get

$$P \implies Q$$
$$\leftrightarrow \neg(\neg(P \implies Q)$$
$$\leftrightarrow \neg(P \wedge (\neg Q))$$
$$\leftrightarrow (\neg P) \vee (\neg\neg Q) \leftrightarrow (\neg P) \vee Q.$$

□

Thus the two "three out of four" operations $\vee$ and $\implies$ are not logically equivalent...but $\implies$ is logically equivalent to an operation involving $\vee$.

Here are some further important logical equivalences involving implication.

PROPOSITION 2.9. *For statements P and Q, we have*

$$(P \iff Q) \leftrightarrow (P \implies Q) \wedge (Q \implies P).$$

PROOF. Any assertion about logical equivalence of binary logical expressions, no matter how fancy-looking, can be established just by writing out the four-rowed truth table. So here we go:

P	Q	$P \iff Q$	$(P \implies Q) \wedge (Q \implies P)$
T	T	T	T
T	F	F	F
F	T	F	F
F	F	T	T

□

One all-important consequence is that two statements are equivalent precisely when each implies the other. In practice, it is most often the case that to prove $A \iff B$ we prove $A \implies B$ and then $B \implies A$.

We now introduce three variants on the implication $P \implies Q$:

1) The **converse implication** $Q \implies P$.
2) The **inverse implication** $(\neg P) \implies (\neg Q)$.
3) The **contrapositive** $(\neg Q) \implies (\neg P)$.

The key question is which of these variations on implication are logically equivalent and especially, whether any of them is logically equivalent to $P \implies Q$. And again, no problem to answer any question of this type: just make a truth table:

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

Contemplation of the table establishes the following result.

PROPOSITION 2.10. *Let P and Q be statements.*

a) *The implication $P \implies Q$ and the contrapositive $(\neg Q) \implies (\neg P)$ are logically equivalent to each other.*
b) *The converse implication $Q \implies P$ and the inverse implication $(\neg P) \implies (\neg Q)$ are logically equivalent to each other.*
c) *The implication is not logically equivalent to either the converse implication or to the inverse implication.*

For students of mathematics, distinguishing between $P \implies Q$ and $Q \implies P$ is usually not a problem: all squares are rectangles, but not all rectangles are squares. OK! However, in everyday life confusion between these two assertions is so abundant that there is a name for it: the **converse fallacy**.

Proposition 2.10a) will later become the basis for an important proof technique, **proof by contraposition**.

5. The Logic of Contradiction

PROPOSITION 2.11. *For any statements P and Q, the following is a tautology:*

$$((\neg P) \implies (Q \wedge \neg Q)) \implies P.$$

PROOF. Of course we could establish this with a truth table. But instead, let us talk it out: like any implication, it is true unless its hypothesis is true and its conclusion is false, so we must rule that out. Namely, we must rule out that both

$$(\neg P) \implies (Q \wedge \neg Q)$$

and

$$\neg P$$

hold. But if so, then $Q \wedge (\neg Q)$ holds, but whether Q is true or false, it follows that $Q \wedge (\neg Q)$ is false. □

In Proposition 2.11 we could replace $Q \wedge (\neg Q)$ with *any* logical contradiction: that is, with any statement that evalues to false whether P itself is true or false. However the idea of Proposition 2.11 is that it models the most common form of a proof by contradiction: you want to establish P, so for the sake of argument suppose that P is false. If from this you can deduce two contradictory statements Q and $\neg Q$, then your argument has reached a false conclusion so you must have a false premise. But your only premise was that P was false, so it must be that P is true.

6. Logical Operators Revisited

Let us think a bit more about logical expressions in the statements $P_1, \ldots, P_n$. What is such a thing?

If we construe syntactically different expressions as different, then even using the expressions we have already defined, there are clearly infinitely many different logical expressions involving even one statement P, e.g.

$$P, (P \vee P), \neg\neg P, (P \wedge P \wedge P)$$

and so forth. So it makes much more sense to consider logical expressions in $P_1, \ldots, P_n$ only up to logical equivalence. How many such expressions are there, and can we describe them all?

To define a logical expression up to logical equivalence really means filling in a column of a truth table with 2^n rows – each row corresponding to one of the possible combinations of the truth values of $P_1, \ldots, P_n$. Since we have 2^n entries to freely fill in with either T or F, it follows that there are precisely 2^{2^n} inequivalent logical expressions involving the statements $P_1, \ldots, P_n$.

We will use the term **logical operator** to mean a logical expression, taken up to logical equivalence, so for instance $P \wedge Q$ and $Q \wedge P$ are different but equivalent logical expressions that determine the same logical operator.[7] This is essentially the same as what we previously called "logical operations," since these were defined in terms of their truth tables.

EXAMPLE 2.12. *Let's write down the logical operators in the statement P. There are $2^{2^1} = 4$ of them. We have seen them all already: they are P, $\neg P$, T – i.e., the tautology: always true – and F – i.e., the contradiction: always false.*

We haven't defined a "tautology symbol" or a "contradiction symbol" yet – do we need to? No, as mentioned before we have

$$T \leftrightarrow P \vee (\neg P), \ F \leftrightarrow P \wedge (\neg P).$$

So all of the inequivalent logical expressions in P are obtained from P, $\vee$, $\wedge$ and $\neg$.

EXAMPLE 2.13. *Let's write down all the logical operators in the statements P and Q. There are $2^{2^2} = 16$ of them. Here they are.*

a) $T \leftrightarrow P \vee (\neg P)$.
b) $P \vee Q$.
c) $(P \implies Q) \leftrightarrow (\neg P) \vee Q$.
d) $(Q \implies P) \leftrightarrow (\neg Q) \vee P$.
e) $P \implies (\neg Q) \leftrightarrow (\neg P) \vee (\neg Q)$.
f) P.
g) $\neg P$.
h) Q.
i) $\neg Q$.
j) $P \iff Q \leftrightarrow (P \wedge Q) \vee (\neg P \wedge \neg Q)$.
k) $P \iff (\neg Q) \leftrightarrow (P \wedge \neg Q) \vee (\neg P \wedge Q)$.
l) $P \wedge Q$.
m) $(\neg P) \wedge Q$.
n) $P \wedge (\neg Q)$.
o) $(\neg P) \wedge (\neg Q)$.
p) $F \leftrightarrow P \wedge (\neg P)$.

How do we know all these operators are different – i.e., no two of the given logical expressions are logically equivalent? All we have to do is write down the truth tables – you are asked to do this in Exercise 2.4. When the most familiar version of this expression is not built up out of P, Q, $\neg$, $\vee$ and $\wedge$ we have given an equivalent

[7]The term "logical operator" is reasonable but not completely standard. The more commonly used term **Boolean function** means exactly the same thing.

expression that is built up in this way. This shows that there are no "essentially new" logical operators in P and Q: the ones we have already introduced can be combined to yield every possible logical operator.

EXAMPLE 2.14. *There are $2^{2^3} = 256$ logical operators in the statements P_1, P_2 and P_3. We could list them all in a truth table with $2^3 = 8$ rows and $2^{2^3} = 256$ columns. Then we can, one by one, try to write these operators in terms of P_1, P_2, P_3, $\neg$, $\vee$ and $\wedge$. Do you want to do this? Me neither – let us try to come up with a more general, conceptual approach.*

Let $P_1, \ldots, P_n$ be statements. For any list $\ell : a_1, \ldots, a_n$ of length n with entries in $\{T, F\}$, we can build a logical operator out of $P_1, \ldots, P_n$, $\neg$ and $\wedge$ that is true precisely when for all $1 \leq i \leq n$ the truth value of P_i is a_i: for $1 \leq i \leq n$ let us put

$$P_i^{a_i} := \begin{cases} P_i & \text{if } a_i = T \\ \neg P_i & \text{if } a_i = F \end{cases}.$$

Then the operator we have in mind is

$$(P_1^{a_1}) \wedge (P_2^{a_2}) \wedge \ldots \wedge (P_n^{a_n}).$$

For instance, if the list is $\ell : T, T, F, F, T$ then the expression is

$$X_\ell : P_1 \wedge P_2 \wedge (\neg P_3) \wedge (\neg P_4) \wedge P_5.$$

This operator evaluates to true if and only if the truth value of P_i is a_i for all i, so it does what we want. This corresponds to a logical operator whose corresponding column in the truth table has a single T in the row corresponding to the list ℓ and has all $2^n - 1$ other entries F.

As above, the contradiction F can be achieved as $P_1 \wedge (\neg P_1)$. Every other logical operator has some number $1 \leq m \leq 2^n$ entries equal to T, and these entries correspond to certain lists $\ell_1, \ldots, \ell_m$ of length n with entries in $\{T, F\}$. Therefore the logical operator that gives the correct column in the truth table is

$$X_{\ell_1} \vee X_{\ell_2} \vee \ldots \vee X_{\ell_m}.$$

We have shown the following result:

PROPOSITION 2.15. *Every logical operator in $P_1, \ldots, P_n$ can be built up out of $P_1, \ldots, P_n$, $\neg$, $\vee$ and $\wedge$.*

Could we go further? That is, do we need all of $\neg$, $\vee$ and $\wedge$? Well, just by taking $n = 1$ we see that we need $\neg$, since $P \wedge P \leftrightarrow P \vee P \leftrightarrow P$. On the other hand, given that we use $\neg$, we do not need both of $\vee$ and $\wedge$: Proposition 2.5 tells us how to express either one of $\vee$, $\wedge$ in terms of the other and $\neg$. Thus we get:

PROPOSITION 2.16.

a) *Every logical operator in $P_1, \ldots, P_n$ can be built up out of $P_1, \ldots, P_n$, $\neg$ and $\vee$.*
b) *Every logical expression in $P_1, \ldots, P_n$ can (up to logical equivalence) be built up out of $P_1, \ldots, P_n$, $\neg$ and $\wedge$.*

Proposition 2.16 looks like the end of this road: using $P_1, \ldots, P_n$ and $\neg$ it is clear that we can build up exactly $2n$ logical operators:

$$P_1,\ \neg P_1,\ P_2, \neg P_2, \ldots, P_n, \neg P_n.$$

When $n = 1$ this gives us all 2^n inequivalent expressions, but for $n \geq 2$ it does not. However we *can* go further: I claim there is a binary logical operator $X(P_1, P_2)$ such that using P_1, P_2 and $X(P_1, P_2)$ one can build $\neg$ and $\wedge$ and thus every logical operator. In fact, I claim that precisely 2 of the 16 binary logical operators in Example 2.13 have this property. You are asked to show this in Exercise 2.5.

Is Exercise 2.5 significant? Not to us, no...though it seems interesting. But there are connections between logical expressions and electrical engineering: one can think of an n-ary logical circuit (or "gate") as something that takes n different wires as input and has 1 output wire. Depending upon which of the input wires carry current, the circuit decides whether the output wire carries current. Letting T correspond to "carries current" and F correspond to "does not carry current," the possible n-ary logical circuits correspond to the 2^{2^n} logical expressions in $P_1, \ldots, P_n$. Now 2^{2^n} grows very rapidly: e.g. there are $65,536$ different 4-ary logical circuits. It would be ridiculous for someone to need $65,536$ different kinds of circuits on hand. Proposition 2.15 shows that we can get away with one unary circuit (a "not gate" that converts current into no current and vice versa) and two binary circuits corresponding to $\wedge$ and $\vee$. Proposition 2.16 shows that we only need one of the $\wedge$ and the $\vee$. Finally, Exercise 2.5 shows that one in fact needs just *one binary circuit* out of which all other logical circuits can be built. In my experience this fact is indeed better known to electrical engineers and computer scientists than mathematicians.

7. Open Sentences and Quantifiers

Consider the following sentence:

$P(x, y)$: $y = x^2$.

The sentence $P(x, y)$ is *not* a statement, because its truth value depends on what x and y are, and they are unspecified. Suppose that we consider x and y ranging over the real numbers. Then, of course, there are some pairs $(x, y) \in \mathbb{R}^2$ for which P is true and other pairs $(x, y) \in \mathbb{R}^2$ for which $P(x, y)$ is false. Indeed the **graph** of the function $f(x) = x^2$ is precisely

$$\{(x, y) \in \mathbb{R}^2 \mid P(x, y) \text{ is true}\},$$

so it is a parabola in the Cartesian plane.

We say that $P(x, y)$ is an **open sentence**. In general, an open sentence P is like a statement, but it contains **variables** $x_1, \ldots, x_n$ which we understand to range over certain nonempty sets $S_1, \ldots, S_n$. We sometimes speak of $\prod_{i=1}^n S_i$ as the **domain** of the open sentence P. Thus for each element $(x_1, \ldots, x_n) \in \prod_{i=1}^n S_i$ we can "plug it into P" and thereby get a statement $P(x_1, \ldots, x_n)$. The truth value of $P(x_1, \ldots, x_n)$ may of course depend upon the choice of $(x_1, \ldots, x_n)$, and then as above we can consider the **truth locus** of P, namely

$$\mathbb{T}(P) \coloneqq \{(x_1, \ldots, x_n) \in \prod_{i=1}^n S_i \mid P(x_1, \ldots, x_n) \text{ is true}\}.$$

PROPOSITION 2.17. *Let $P = P(x)$ be an open sentence with variable x ranging over the nonempty set S. Then we have*

$$\mathbb{T}(\neg P) = S \setminus \mathbb{T}(P).$$

You are asked to prove Proposition 2.17 in Exercise 2.10.

Thus an open sentence is a bit more interesting than a statement: whereas a statement is either true or false, or an open sentence we get to ask for which values it is true and for which values it is false. However, it is very useful to be able to "collapse" an open sentence into a statement, which we can do by asserting something about its truth locus $\mathbb{T}(P)$. There are two standard ways to do this.

Universal quantifier: We introduce a new symbol $\forall$ that we read as "for all." Suppose that $P = P(x)$ is an open sentence involving a variable x that ranges over the nonempty set S. Then

$$\forall x \in S,\ P(x) \tag{3}$$

is a statement: in words, it is "For all elements x in S, we have that $P(x)$ is true." It is true if and only if $P(x)$ is true for all $x \in S$, or equivalently if the truth locus $\mathbb{T}(P)$ is all of S. We call $\forall$ the **universal quantifier**. This terminology is because (I think) we are viewing S as our universal set, so if the truth locus is S then the sentence is "universally true" – that is, it holds for all elements of the universal set.

EXAMPLE 2.18.

a) *We start with the open sentence $P(x) : x^2 + 1 > 0$, where x ranges over the real numbers. Applying the universal quantifier we get the statement*

$$\forall x \in \mathbb{R},\ x^2 + 1 > 0.$$

*The quantified statement is **true**, since for any real number x we have $x^2 \geq 0$ and thus $x^2 + 1 \geq 1 > 0$.*

b) *We start with the open sentence $Q(x) : x^2 - 1 > 0$, where x ranges over the real numbers. Applying the universal quantifier we get the statement*

$$\forall x \in \mathbb{R},\ x^2 - 1 > 0.$$

*The quantified statement is **false**: in order to be true it would have to hold for all $x \in \mathbb{R}$, but it is false e.g. for $x = 0$. More precisely, we have*

$$\mathbb{T}(Q) = \{x \in \mathbb{R} \mid x^2 - 1 > 0\} = (-\infty, -1) \cup (1, \infty).$$

So the truth locus of Q is the union of two intervals on the real line. That locus is not all of $\mathbb{R}$, so the universally quantified sentence is false.

c) *We start with the open sentence $R(x) : x^2 + 1 \leq 0$, where x ranges over the real numbers. Applying the universal quantifier we get the statement*

$$\forall x \in \mathbb{R},\ x^2 + 1 \leq 0.$$

*The quantified statement is **false**. Indeed, $R(x) = \neg P(x)$, so since the sentence $P(x)$ holds for all $x \in \mathbb{R}$, the sentence $Q(x)$ holds for no $x \in \mathbb{R}$: its truth locus $\mathbb{T}(R)$ is $\varnothing$.*

Existential quantifier: We introduce a new symbol $\exists$ that we read as "there exists." Suppose that $P = P(x)$ is an open sentence involving a varaible x that ranges over the nonempty set S. Then

$$\exists x \in S,\ P(x) \tag{4}$$

is a statement: in words, it is "There exists x in S such that $P(x)$ is true." (Notice that in translating (3) to words we added "we have": this was just for reasons of English usage, as it it better not to start a phrase with a symbol. The words "we have" add nothing to the meaning. Similarly, in translating (4) to words we added "such that": this is necessary in order to get a grammatically correct English sentence, but it does not make any mathematical change.) It is true if and only if $P(x)$ is true for some (i.e., at least one) $x \in S$, or equivalently if the truth locus $\mathbb{T}(P)$ is nonempty. We call $\exists$ the "existential quantifier" for reasons that seem more clear: elements of $\mathbb{T}(P)$ exist.

EXAMPLE 2.19. *We revisit Example 2.18 with $\exists$ in place of $\forall$.*

a) *Consider the statement*

$$\exists x \in \mathbb{R},\ x^2 + 1 > 0.$$

This is true, and to verify that it is enough to find one real number x such that $x^2 + 1 > 0$. Since $0^2 + 1 = 1 > 0$, that suffices. Previously we established that $\forall x \in \mathbb{R},\ x^2 + 1$ was true. Observe that this is a stronger statement: if something is true for all $x \in \mathbb{R}$ it is most certainly true for some *$x \in \mathbb{R}$. This observation is completely general: for any open sentence $P(x)$ with x ranging over the nonempty[8] set S, we have*

$$(\forall x \in S,\ P(x)) \implies (\exists x \in S,\ P(x)).$$

b) *Consider the statement*

$$\exists x \in \mathbb{R},\ x^2 - 1 > 0.$$

Since $2^2 - 1 = 3 > 0$, indeed there is an $x \in \mathbb{R}$ such that $x^2 - 1 > 0$. Thus the statement is true. This was not a surprise, since earlier we observed that $\mathbb{T}(Q) = (-\infty, -1) \cup (1, \infty)$. In particular $\mathbb{T}(Q) \neq \varnothing$, which is all we need to see that the existentially quanitified statement is true.

c) *Consider the statement*

$$\exists x \in \mathbb{R}, x^2 + 1 \leq 0.$$

This is false: since for all $x \in \mathbb{R}$ we have $x^2 + 1 > 0$, for no $x \in \mathbb{R}$ do we have $x^2 + 1 \leq 0$. Here we started with an open sentence $P(x)$ that was universally true and passed to its negation $\neg P(x)$, which is therefore universally false: the truth set switches from the universal set to the empty set. Symbolically:

$$\forall x \in S,\ P(x) \iff \neg(\exists x \in S,\ \neg P(x)). \tag{5}$$

We record an equivalent version of this observation as Proposition 2.20a), coming up next.

PROPOSITION 2.20. *Let $P(x)$ be an open sentence with a variable x ranging over the nonempty set S.*

[8]We are actually using here that the set S is nonempty.

a) *We have*

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

b) *We have*

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

PROOF. a) By Proposition 2.17 we have

$$\mathbb{T}(\neg P) = S \setminus \mathbb{T}(P).$$

So $\neg(\forall x \in S,\ P(x))$ holds if and only if $\mathbb{T}(P) \subsetneq S$ if and only if $\mathbb{T}(\neg P) = S \setminus \mathbb{T}(P) \neq \varnothing$ (a subset of a universal set is proper if and only if its complement is nonempty) if and only if $\exists x \in S,\ \neg P(x)$.

Alternately, we may apply $\neg$ to both sides of (5).

b) Since $\exists x \in S,\ P(x)$ holds if and only if $\mathbb{T}(P) \neq \varnothing$, we have $\neg(\exists x \in S,\ P(x))$ if and only if $\mathbb{T}(P) = \varnothing$ if and only if $\mathbb{T}(\neg P) = S$ if and only if $(\forall x \in S,\ \neg P(x))$. □

One might expect there to be other quantifiers besides $\exists$ and $\forall$, but I only know of one that is in common use: when we write

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

and say "There is a unique $x \in S$ such that $P(x)$ holds" to mean that the set $\mathbb{T}(P)$ of $x \in S$ such that $P(x)$ holds has exactly one element. (In general, the word "unique" is used in mathematics to mean that there is at most one.) It is somewhat tempting to invent new quantifiers, for instance $\exists^\infty$ could mean "There are infinitely many." However it turns out that one can do quite well enough with the old standbys of $\forall$ and $\exists$ (and occasionally $\exists!$). I will let you reflect on why this might be the case.

We introduced open sentences as having several variables, but our discussion of quantifiers has thus far been with sentences involving only one variable. In general, if you have an open sentence $P(x_1, \ldots, x_n)$ in which x_i ranges over S_i for $1 \leq i \leq n$, one can apply either $\forall$, $\exists$ or neither one to each variable separately. In order to get a statement we should apply either quantifier to *all* of the variables; if we only quantify some of the variables then we are left with an open sentence whose domain is the Cartesian product of the unquantified variables. In logic one often speaks of a variable as being **bound** by a quantifier and as **unbound** or **free** if no quantifier is applied. This terminology will be (only) occasionally helpful to us.

EXAMPLE 2.21. *One can think of $\exists$ as inducing a **projection** on the truth set. We illustrate this with two examples.*

a) *Consider the open sentence*

$$P(x, y) : x = y^2$$

with domain $(x, y) \in \mathbb{R} \times \mathbb{R}$. Its truth locus is the horizontally opening parabola. Now consider

$$Q(x) : \exists y\ (x = y^2),$$

with domain $x \in \mathbb{R}$. For $x \in \mathbb{R}$, $Q(x)$ is true if and only if x is the square of some other real number y, which holds if and only if $x \geq 0$. Thus the truth locus of Q is $[0, \infty)$. But now notice that if we project the parabola $\{(x, y) \in \mathbb{R}^2 \mid x = y^2\}$ onto the x-axis we get $\mathbb{T}(Q) = [0, \infty)$.

b) *Consider the sentence*

$$P(x,y): \frac{x^2}{4}+\frac{y^2}{9}=1$$

with domain $(x,y)\in\mathbb{R}\times\mathbb{R}$. *Its truth locus is an ellipse centered at the origin. The truth locus of*

$$Q(x): \exists y\ (\frac{x^2}{4}+\frac{y^2}{9}=1)$$

is the set of $x\in\mathbb{R}$ *such that* $9(1-\frac{x^2}{4})$ *has a real square root, which happens if and only if* $1-\frac{x^2}{4}\geq 0$, *which happens if and only if* $|x|\leq 2$, *so* $\mathbb{T}(Q)=[-2,2]$. *This is the projection of the ellipse onto the* x*-axis. Similarly for*

$$R(y): \exists x\ (\frac{x^2}{4}+\frac{y^2}{9}=1)$$

the truth locus of $R(y)$ *is the set of* $y\in\mathbb{R}$ *such that* $4(1-\frac{y^2}{9})$ *has a real square root, which happens if and only if* $|y|\leq 3$, *so* $\mathbb{T}(R)=[-3,3]$. *This is the projection of the ellipse onto the* y*-axis.*

EXAMPLE 2.22. *We consider various quantifications of the open sentence*

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

with domain $(x,y)\in\mathbb{R}\times\mathbb{R}$.

a) *If we don't quanitify at all, then* $\mathbb{T}(P)$ *is the set of points in the plane lying over the 45 degree line* $y=x$.

b) *Consider*

$$Q(x): \forall y\in\mathbb{R},\ y>x.$$

Here we have written $Q(x)$ *because the variable* y *is bound by the quantifier* y *and therefore "used up": we can't plug in values for* y. *However, we can still plug in values for* x *and thus we get an open sentence with domain* $x\in\mathbb{R}$. *If we plug in* $x\in\mathbb{R}$, *the statement* $Q(x)$ *asserts that* x *is less than every real number. This is false no matter what* x *is, so* $\mathbb{T}(Q)=\varnothing$.

c) *Consider*

$$R(x): \exists y\in\mathbb{R},\ y>x.$$

Again we get an open sentence with domain $x\in\mathbb{R}$. *If we plug in* $x\in\mathbb{R}$, *the statement* $Q(x)$ *asserts that some real number is greater than* x. *This is true no matter what* x *is: we can take* $y=x+1$. *So* $\mathbb{T}(R)=\mathbb{R}$.

d) *Consider the statement*

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

This statement asserts that $R(x)$ *is true for all* $x\in\mathbb{R}$, *which as we saw above is true, so the statement is true.*

e) *The statement*

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

asserts that every real number is greater than every other real number. This is false.

f) *The statement*

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

asserts that there are real numbers x and y such that $y > x$. This is true: $1 > 0$.

g) *Consider the statement*

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

This statement asserts that there is a real number x that is smaller than every real number. This is false.

Now for a very important observation: the statements

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

and

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

are identical except for the order of the quantifiers: the first one has $\forall$ then $\exists$, whereas the second has $\exists$ then $\forall$. The first statement is true and the second statement is false. Therefore:

Changing the order of the quantifiers can change the meaning and truth value of a statement.

More precisely, swapping $\exists$ and $\forall$ can change the meaning of a statement: in fact it usually does in a dramatic way. Moving two existential quantifiers past each other does not change the meaning of truth value of a statement (or the meaning or truth locus of an open sentence): $\exists x_1 \in S_1,\ \exists x_2 \in S_2$ means "There is x_1 in S_1 and x_2 in S_2 such that..." which means the same thing as "there is $x_2 \in S_2$ and $x_1 \in S_2$ such that..."

Often it is a reasonable perspective to count repeated instances of the same quantifier as one single quantifier: this is because e.g.

$$\forall x_1 \in S_1, \forall x_2 \in S_2\ P(x_1, x_2)$$

means the same thing as

$$\forall (x_1, x_2) \in S_1 \times S_2\ P(x_1, x_2).$$

On the other hand, the more "alternating quantifiers" a statement has, the more complex it is both logically and mathematically. A statement with a single quantifier is usually immediately understood. Statements with two quantifiers often require a little thought. Statements with at least three quantifiers often require special training to properly understand.

EXAMPLE 2.23. *Let $f : \mathbb{R} \to \mathbb{R}$ be a function. The definitions of "$\lim_{x \to c} f(x)$" and "f is continuous at c" involve three quantifiers. It is probably because of their logical complexity that they are not introduced in freshman calculus. For instance, here is the definition of continuity at* 0*: the function f is continuous at $x = 0$ if:*

$$\forall \epsilon > 0\ \exists \delta > 0\ \forall x \in \mathbb{R},\ |x| < \delta \implies |f(x) - f(c)| < \epsilon.$$

A note on uses of quantifiers: throughout much of mathematics, explicit use of the symbols $\forall$ and $\exists$ is not much encountered in formal mathematical writing: it is often preferable to use the words "for all" and "there exists" (or even "there is") instead. These symbols are more commonly encountered on the blackboard, in handwritten notes, and other less formal mathematical presentations. On the other hand there is a branch of mathematics, **mathematical logic** in which logic and its interactions with other parts of mathematics are studied (many of these interactions are more interesting and less "foundational" or "philosophical" than one might expect), and in this part of mathematics the symbols $\exists$ and $\forall$ abound.

Perhaps because undergraduate "foundations" courses such as this one are a fairly recent invention, mathematicians also differ among each other in how explicit they are in their quantification. All mathematicians make use of quantifiers, but some mathematicians prefer to leave certain quantifications implicit. I am not one of those mathematicians, and I think explicit use of quantifiers is a key part of accurate mathematical communication. But you should be aware of and keep an eye out for "implied quantification." When this happens, it is usually the case that the implied quantifier is $\forall$, not $\exists$. This creates problems for some students, who when in doubt may supply the missing quantifier as $\exists$ because after all that is a weaker assertion and therefore easier to prove.

EXAMPLE 2.24. *Horace is asked to prove:*

"A non-negative real number has a real square root."

His answer:

"The real number 9 *is non-negative, and since* $3^2 = 9$*, we have that* 3 *is a real square root of* 9*."*

Do you see the issue?

Horace has interpreted the statement as "There exists a non-negative real number that has a real square root." Outside of the context of mathematical discourse, one can defend his decision: after all the sentence begins with "A". Perhaps Horace, as a student, is used to thinking of mathematical statements as problems being posed rather than logical statements that are either true or false. If one tacked the word "Find" onto the beginning of the sentence, we would not have a statement to prove but a command to follow: a command that Horace correctly followed.

But any mathematician would hear the unspoken universal quantifier and interpret the statement as "For all real numbers $a \geq 0$*, there is* $b \in \mathbb{R}$ *such that* $b^2 = a$*." This is a much deeper assertion. It can be solved using the Intermediate Value Theorem from calculus (a result which, by the way, is not usually proved until a more advanced course) as follows: let* $a \geq 0$*. If* $a = 0$ *then* 0 *is a real square root of* a*, so we may assume* $a > 0$*. Now consider the function* $f : \mathbb{R} \to \mathbb{R}$ *defined by*

$$f(x) = x^2 - a.$$

We have $f(0) = -a < 0$*. We also have*

$$f(a+1) = (a+1)^2 - a = a^2 + a + 1 > a^2 - 2a + 1 = (a-1)^2 \geq 0.$$

Since f *is a continuous function,* $f(0) < 0$ *and* $f(a+1) > 0$*, the Intermediate Value Theorem implies that there is* $b \in \mathbb{R}$ *with* $0 < b < a + 1$ *such that*

$$0 = f(b) = b^2 - a.$$

This means that $b^2 = a$, so b is a square root of a.

In mathematics we usually want to prove "general statements" and not "examples." So in the vast majority of mathematical results, we have a universally quantified statement. In fact the most common form of a mathematical theorem is:

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

where P and Q are both open sentences with domain x. (The sentences P and Q may themselves involve other quantifiers and often do.)

8. Negating Statements

In this section we have the relatively modest goal of giving some tips and practice on negating statements.

Let us begin here: why would we want to negate statements? How does this come up in practice?

It is often said that the main goal in pure mathematics is to prove theorems. This is approximately true. But it omits something important: since we (not exclusively, but especially and most excitingly) want to prove *new* theorems, and a proof of a theorem is how we know the theorem is true, a more accurate statement of mathematical practice is that we are interested in either proving or *disproving* statements. How does one go about disproving a statement P?

I can think of two good ways to do this.

First Method: Prove $\neg P$.

This works of course, since P is false if and only if $\neg P$ is true. Construed this way, the task of disproving has no formal difference from the task of proving: we're just trying to prove a different statement.

Second Method: Assume that P is true and deduce a logical contradiction.

This works too, of course, since if it cannot be the case that P is true, then it must be that P is false.

Of these, the first method involves negating P. The second ostensibly does not.

EXAMPLE 2.25. a) *The Pythagorean school famously disproved that $\sqrt{2}$ is a rational number. The proof, which is perhaps the most famous of all mathematical proofs, will be given in full later, but it really does begin by assuming that $\sqrt{2}$ is a rational number, writing $\sqrt{2} = \frac{a}{b}$ for integers a and b with $b \neq 0$, and deducing a contradiction. There is no negation here (at least, not yet).*

b) *Goldbach's Conjecture states that every even integer $n \geq 4$ is the sum of two prime numbers. It is widely believed to be true and known to hold for all $4 \leq n \leq 4 \cdot 10^{18}$, but it is not currently proved. How could we disprove Goldbach's Conjecture? It does not seem plausible to disprove Goldbach's Conjecture* by contradiction*: we assume that every even integer $n \geq 4$ is*

the sum of two primes and then try to reach a contradiction – how in the world are we going to do that?!?

In this case the only plausible way of disproving Goldbach's conjecture is to prove its negative. Goldbach's conjecture is that for all even $n \geq 4$, there are prime numbers p_1 and p_2 such that $n = p_1 + p_2$. So its negation is: there exists an even $n \geq 4$ such that for all prime numbers p_1 and p_2, we have $n \neq p_1 + p_2$. If this (the negation) is true and I give you the n, it is easy to check: for all prime numbers $p < n$ you compute that $n - p$ is not prime.

Negating statements also comes up when we are trying to *prove* P, not just disprove it. Most mathematical theorems have the form:

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

There are three basic ways of doing this:

I. Direct Proof: Let $x \in S$. (What follows either must work for all $x \in S$ at the same time or be divided into cases. We will discuss this in much more detail later.) Assume that $P(x)$ is true, and deduce that $Q(x)$ is true.

II. Proof by Contrapositive: Let $x \in S$. (Same parenthetical comment as above). Assume that $\neg Q(x)$ is true and deduce that $\neg P(x)$ is true.

III. Proof by Contradiction: Let $x \in S$. (Same parenthetical....) Assume that $P(x)$ is true and that $\neg Q(x)$ is true and deduce a contradiction.

Two out of the three methods involve negating statements. Taken together, that is pretty good motivation!

Next issue: isn't the process of negation pretty trivial? Here, I will negate an arbitrary statement P:

$$\neg P.$$

In words, for any statement P its negation is "It is not the case that P holds."

However, although this negation just by tacking "It is not the case that..." on the front of a statement is logically valid, it is practically not very useful. The point is that mathematical reasoning is all about deduction. We can build on something that *is* the case; it is much harder to build directly on what is *not* the case. Here is a first example:

PROPOSITION 2.26. *For all $x \in \mathbb{Z}$, if x^2 is even, then x is even.*

PROOF. We will use the contrapositive: for $x \in \mathbb{Z}$, if x is not even, then x^2 is not even. Okay, but what does it mean not to be even? In pre-university mathematics one often defines an even number to be a whole number that is twice another whole number and an odd number to be a whole number that is not even...but this is not directly helpful. How do we build on the fact that x *cannot* be written as $2n$ for some $n \in \mathbb{Z}$? The key to moving on is to find an equivalent "positive" statement. As mentioned in Example 1.22 (and as we will prove later on), it turns out that if

an integer x is not even then it is of the form $2n+1$ for some $n \in \mathbb{Z}$. This finally is telling us something about x, and we can use it: if $x = 2n+1$ then

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

is also odd, completing the proof. □

That an integer that is not even is of the form $2n+1$ is an instance of "good negation." The best (most useful) kind of negation doesn't have the word "not" in it at all: instead of asserting the absence of something it is asserts the presence of some other, complementary property. Exactly how this works depends on the situation, but at a formal level our goal is to "work the negation from the outside in": you will know that you've negated a logical expression in a good way when the negations just get applied to basic propositions, not to more complicated expressions.

EXAMPLE 2.27. *Some Good Negations:*

a) *A good negation of* $P \implies Q$ *is* $P \wedge (\neg Q)$.
(This is as far as we can go in so much generality, but in practice you are left with the task of finding a good negation of Q*.)*

This example also comes with a warning: ***the negation of an implication is NOT an implication of any kind****. In my experience, many students know this in pleasant circumstances but upon being sufficiently pressured (e.g. in an exam setting) will write something like*

$$\text{"}\neg(P \implies Q) \leftrightarrow (\neg P) \implies (\neg Q)\text{."}$$

This is TERRIBLY WRONG*. A "quantitative" approach should help here: as I stressed above, any implication is a "3/4 operator": one that is true for 3 out of 4 of the possible combinations of the basic truth values. Its negation therefore must be a "1/4 operator," which is not an implication of any kind (i.e., with* P *and/or* Q *swapped and/or negated).*

b) *A good negation of* $P_1 \wedge P_2 \wedge P_3 \wedge P_4$ *is*

$$(\neg P_1) \vee (\neg P_2) \vee (\neg P_3) \vee (\neg P_4).$$

A good negation of $Q_1 \vee Q_2 \vee Q_3$ *is*

$$(\neg Q_1) \wedge (\neg Q_2) \wedge (\neg Q_3).$$

These are logical DeMorgan's Laws.

c) *A good negation of* $\forall x \in S,\ P(x)$ *is*

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

A good negation of $\exists x \in S,\ P(x)$ *is*

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

Again, this shows that we want to work our negations "from the outside to the inside."

We now move on to some practical examples.

EXAMPLE 2.28. *We negate the statement "If I'm lying, I'm dying.'*
This statement has the form $P \implies Q$ *where* P *is "I'm lying" and* Q *is "I'm dying." Following the template of Example 2.27a), a good negation would be:*

I'm lying and I'm not dying.

It might be even better to express "I'm not dying" in a positive way, but I don't know a simple English construction that expresses this meaning exactly.

EXAMPLE 2.29. *We negate the statement "For every real number x there is a larger real number y." More symbolically this is*

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

Let's negate it from the outside in:

$$\neg(\forall x \in \mathbb{R}\ \exists y \in \mathbb{R},\ y > x)$$

$$\leftrightarrow \exists x \in \mathbb{R}(\neg(\exists y \in \mathbb{R},\ y > x))$$

$$\leftrightarrow \exists x \in \mathbb{R}(\forall y \in \mathbb{R}, \neg(y > x))$$

$$\leftrightarrow \exists x \in \mathbb{R}\ \forall y \in \mathbb{R},\ y \leq x.$$

Thus the negated statement asserts that there is a largest real number, which means that the original statement could have been more pithily expressed as "There is no largest real number."

EXAMPLE 2.30. *We negate the following sentence*[9]*: "You can fool some of the people all of the time and all of the people some of the time, but you cannot fool all of the people all of the time."*

First we have to realize that the "but" is playing the logical role of an "and" so the statement has the form $A \wedge B \wedge C$ with

$$A:\ \textit{You can fool some of the people all of the time.}$$

$$B:\ \textit{You can fool all of the people some of the time.}$$

$$C:\ \textit{You cannot fool all of the people all of the time.}$$

So the negation is $\neg(A \wedge B \wedge C) \leftrightarrow (\neg A) \vee (\neg B) \vee (\neg C)$. Now we have to work on A, B and C individually, which are all themselves quantified statements.

Statement A takes the form: "For all times, there exists a person you can fool," or more schematically "$\forall t\ \exists P$ such that you can fool P." So its negation is: there exists a time at which you can't fool anyone.

Statement B takes the form "There exists a time such that you can fool all the people," or more schematically "$\exists t\ \forall P$ you can fool P." So its negation is: At all times, there is at least one person you can't fool.

Happily, C has the form of a "cheaply negated" statement: "It is not the case that you can fool all of the people all of the time." So its negation is "At all times, you can fool all of the people." Final answer:

At least one of the following holds: (i) There is at least one time where you can't fool anyone, or (ii) At all times there is someone you can't fool, or (iii) At all times, you can fool everyone.

[9]This sentence is traditionally attributed to Abraham Lincoln, but it does not seem clear whether he actually said it.

9. Isotone Logical Operators

Let X be a logical operator involving the basic statements $P_1, \ldots, P_n$. We can think of X as taking as input any length n list ℓ with entries in the set $\{T, F\}$ – let us call this a **Boolean list** of length n – and returning a value T or F. (The ith entry of the list ℓ tells us whether to evaluate P_i as true or false.) For such lists ℓ, we write $\ell_1 \prec \ell_2$ if for all $1 \leq i \leq n$, if the ith entry of ℓ_1 is T, then the ith entry of ℓ_2 is also T. Another way to say this is that $\ell_1 \prec \ell_2$ if $\ell_2 = \ell_1$ or ℓ_2 can be obtained from ℓ_1 by changing some of the F's in ℓ_1 to T's. We then say that the n-ary logical operator X is **isotone** if for all such lists ℓ_1 and ℓ_2 with $\ell_1 \prec \ell_2$, if $X(\ell_1)$ is true, then also $X(\ell_2)$ is true.[10]

We can also think of an n-ary logical operator as a "gate" in which n wires go in and 1 wire goes out. Each of these n wires may or may not have current flowing through it ("is on"), and the operator is the rule that determines whether in each of thse 2^n cases the outward wire has a current flowing through it ("is on"). In this interpretation a logical operator is isotone if whenever some input configuration of currents leads to an "on" output, if we then change some of the off input currents to on, the output current remains on.

EXAMPLE 2.31. *Let $P_1, \ldots, P_n$ be basic statements.*

a) *The tautology operator T – i.e., the operator that evaluates every length n Boolean list to true – is an isotone operator. Indeed, changing the entries of the list never changes the output. For the same reasons, the contradiction operator F – i.e., the operator that evaluates every length n Boolean list to false – is an isotone operator.*

b) *If we view P_1 as a logical operator, it is an isotone operator: if it evaluates to true, then P_1 is true. If we then change some of the truth values of $P_2, \ldots, P_n$ from false to true, then the operator P_1 still evaluates to true. Similarly, for all $2 \leq i \leq n$, the logical operator P_i is isotone.*

c) *The binary logical operator $\vee$ is isotone: $P_1 \vee P_2$ evaluates to true on each of the lists (T, T), (T, F) and (F, T) and evaluates to false on (F, F). We cannot get from any of the first three lists to the fourth list by changing F's to T's.*

d) *The binary logical operator $\wedge$ is isotone: $P_1 \wedge P_2$ evaluates to true on the list (T, T) and evaluates to false on (T, F), (F, T) and (F, F). The list (T, T) does not have any F's, so we cannot get from it to any of the other lists by changing F's to T's.*

e) *The unary logical operator $\neg$ is* not *isotone: it evaluates to true on (F), but if we change the F to a T we get (T), on which $\neg$ is false.*

f) *The binary logical operator $\implies$ is* not *isotone: it evaluates to true on (F, F), but it we change the first F to T we get (T, F) on which $\implies$ evaluates to false.*

g) *The binary logical operator $\iff$ is* not *isotone, for exactly the same reasons as part f).*

EXAMPLE 2.32. *Above we found 6 isotone binary logical operators (let us call the two basic statements P and Q): T, $P \vee Q$, P, Q, $P \wedge Q$ and F. It is not hard*

[10] Our terminology seems reasonable but is not standard: instead of "isotone logical operators" it is more common to speak of **monotone Boolean functions**.

to see that the other $2^{2^2} - 6 = 10$ *binary logical operators are not isotone: you are asked to do this in Exercise 2.16.*

What if we wanted to extend Example 2.32 to ternary logical operators (operators in P_1, P_2, P_3)? There are $2^{2^3} = 256$ such operators, so it would be nice to have something faster than checking each one for being isotone. It turns out that we can get some traction from yet another connection between logic and sets.

Let $X(P_1, \ldots, P_n)$ be an n-ary logical operator. Then X is determined by the set of length n Boolean strings ℓ such that X evaluates to true on ℓ. For each length n Boolean string ℓ we may in turn associate a subset $S(\ell)$ of $[n]$, namely

$$S(\ell) \coloneqq \{1 \leq i \leq n \mid \text{ the } i\text{th entry of } \ell \text{ is } T\}.$$

All in all, we associate to the operator X a family $\mathcal{F}(X)$ of subsets of $[n]$, namely

$$\mathcal{F}(X) \coloneqq \{S(\ell) \mid \ \ell \text{ is a length } n \text{ Boolean string on which } X \text{ evaluates to } T\}.$$

EXAMPLE 2.33. *Consider the ternary logical operator* $X \coloneqq P_1 \wedge (P_2 \vee P_3)$*. The set of binary strings on which this operator evaluates to true is:*

$$(T,T,F),\ (T,F,T),\ (T,T,T).$$

Thus we associate the the family of subsets

$$\mathcal{F}(X) = \{\{1,2\},\ \{1,3\},\ \{1,2,3\}\}.$$

The operator X *is isotone: if we change the* F *in* (T,T,F) *to* T*, we get* (T,T,T)*, on which* X *also evaluates to* T*, and similarly if we change the* F *in* (T,F,T) *to* T*, we get* (T,T,T)*, on which* X *also evaluates to* T*. The question we are about to address is how to discern the isotonicity of* X *from the family* $\mathcal{F}(X)$*.*

The correspondence $X \mapsto \mathcal{F}(X)$ is a perfect set theoretic encoding of X: if we know $\mathcal{F}(X)$ we can recover X. Moreover each family $\mathcal{F}$ of subsets of $[n]$ comes from a logical operator. In the language of Chapter 8, we are giving a *bijection* from the set of n-ary logical operators to $2^{2^{[n]}}$, the set of sets of subsets of $[n]$.

EXAMPLE 2.34. *If* $n = 3$ *and*

$$\mathcal{F} = \{\varnothing, \{1\}, \{1,3\}, \{2\}, \{1,2,3\}\}, \tag{6}$$

then the corresponding operator X *evaluates to true on the strings* (F,F,F)*,* (T,F,F)*,* (T,F,T)*,* (F,T,F) *and* (T,T,T) *and evaluates to false on* (F,F,T)*,* (T,T,F) *and* (F,T,T)*. This operator is* not *isotone: an isotone operator that evaluates to true on* (F,F,F) *must evaluate to true on every string, i.e., must be the tautology.*

At this point it is natural to ask for a property of a family of subsets $\mathcal{F}(X)$ of $[n]$ that holds if and only if the corresponding logical operator X is isotone. The set-theoretic analogue of changing F's to T's is adding elements to a member of $\mathcal{F}(X)$. So we find that: the logical operator X is isotone if and only if the family $\mathcal{F}(X)$ is closed under passage to supersets: that is, for all subsets $A \subseteq B \subseteq [n]$, if $A \in \mathcal{F}(X)$ then also $B \in \mathcal{F}(X)$. For any fixed set S, let us call a family $\mathcal{F}$ of subsets of S **isotone** if it is closed under passage to supersets.

We are not done, but let us stop and record the progress made:

PROPOSITION 2.35. *Let $n \in \mathbb{Z}^+$. The map $X \mapsto \mathcal{F}(X)$ is a bijection from the set of all n-ary logical operators to the set $2^{2^{[n]}}$ of all families of subsets of $[n]$. This bijection restricts to give a bijection from the set of all isotone loigcal operators to the set of all isotone families.*

It turns out that the isotone families of subsets of $[n]$ are in turn in bijection with a set of "smaller" families of subsets; using these smaller families is more efficient. To see this, let us look back at the isotone family of subsets of [3] considered above:

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

Because $\{1,2\}$ is an element of $\mathcal{F}$ and the family is isotone, we know it must contain the set $\{1,2,3\}$ because $\{1,2,3\} \supsetneq \{1,2\}$. So we could get away without writing down the set $\{1,2,3\}$: we know it must be there. However since $\{\{1,2\},\{1,2,3\}\}$ is also an isotone family, the same cannot be said about $\{1,3\}$: we do need to record it along with $\{1,2\}$. So in this case it seems that the isotone family $\mathcal{F}$ can be recovered from the subfamily

$$\underline{\mathcal{F}} = \{\{1,2\},\ \{1,3\}\};$$

namely, $\mathcal{F}$ is the set of all subsets Y of [3] such that $Y \supseteq X$ for some $X \in \underline{\mathcal{F}}$.

As another example, if we know that an isotone family $\mathcal{F}$ of subsets of $[n]$ has $\varnothing$ as an element, then it must contain every subset of $[n]$ and thus must be $2^{[n]}$. So to specify this family we only need to record that it has $\varnothing$ as an element.

In Exercise 2.17 you are asked to show that for any set S, if $\mathcal{F}$ is a family of subsets of S, then

$$\mathcal{F}^{\uparrow} := \{T \in 2^S \mid T \supseteq U \text{ for some } U \in \mathcal{F}\}$$

is an isotone family of subsets of S. However, different families $\mathcal{F}$ may give rise to the same isotone family $\mathcal{F}^{\uparrow}$: e.g. if $\mathcal{F}$ is any family of subsets of S with $\varnothing \in \mathcal{F}$, then $\mathcal{F}^{\uparrow} = 2^S$. Coming back to the case of $S = [n]$ we will see that there is always a "smallest" choice of $\mathcal{F}$ that yields a given isotone family $\mathcal{F}^{\uparrow}$.

Let $\mathcal{F}$ be any set of sets. We say that an element $A \in \mathcal{F}$ is **minimal** if there is no $B \in \mathcal{F}$ such that B is a proper subset of A. In the family of (6) the unique minimal element is $\varnothing$. To give another example, the minimal elements of

$$\mathcal{G} := \{\{1\},\{1,3\},\{2,4\}\}$$

are $\{1\}$ and $\{2,4\}$. This example shows that if all the sets in the family are finite, then any set of smallest cardinality is necessarily minimal (because a set that it properly contained would have to have smaller cardinality), but there may be other minimal sets of larger cardinality.

For any family $\mathcal{F}$ of sets, we let $m(\mathcal{F})$ be the set of minimal elements of the family: that is, the set of elements of $\mathcal{F}$ that do not properly contain any element of $\mathcal{F}$. We say that $\mathcal{F}$ is a **Sperner family** (or an **antichain**) if $\mathcal{F} = m(\mathcal{F})$: in other words, $\mathcal{F}$ is a Sperner family if for no two distinct sets X and Y in $\mathcal{F}$ do we have $X \subseteq Y$.

LEMMA 2.36. *Let $\mathcal{F}$ be a family of sets.*

a) *We have $m(\mathcal{F}^{\uparrow}) = m(\mathcal{F})$.*

b) *If $X \in \mathcal{F}$ is a finite set, then there is $Y \in m(\mathcal{F})$ such that $Y \subseteq X$.*

PROOF. a) Let $X \in m(\mathcal{F})$, so X is a minimal element of $\mathcal{F}$. Since $\mathcal{F} \subseteq \mathcal{F}^\uparrow$, certainly X is an element of $\mathcal{F}^\uparrow$, and we need to see that X is a minimal element. Suppose not: then there is $Y \in \mathcal{F}^\uparrow$ such that $Y \subsetneq X$. By definition of $\mathcal{F}^\uparrow$ there is then $Z \in \mathcal{F}$ such that $Z \subseteq Y$. If follows that Z is an element of $\mathcal{F}$ that is a proper subset of X, contradicting our assumption that X is a minimal element of $\mathcal{F}$.

Now suppose that $X \in m(\mathcal{F}^\uparrow)$. Again there is $Y \in \mathcal{F}$ such that $Y \subseteq X$, but since X is a minimal element of $\mathcal{F}^\uparrow$ we must have $Y = X$, so $X \in \mathcal{F}$. Similarly, if X were not a minimal element of $\mathcal{F}$ there would be $Z \in \mathcal{F}$ with $Z \subsetneq X$, but since also $Z \in \mathcal{F}^\uparrow$ this again contradicts the minimality of X. So $X \in m(\mathcal{F})$.

b) Suppose $\#X = n \in \mathbb{N}$. If $n = 0$ then $X = \varnothing$, which is a minimal element of any family of sets it belongs to. Otherwise, if X is itself minimal then it is a minimal element contained in itself, so we may assume that X is not minimal, so there is $Y \in \mathcal{F}$ that is properly contained in X and thus $\#Y \leq n-1$. We may continue this argument in the same manner: if Y is minimal we are done; otherwise it contains a proper subset Z that is an element of $\mathcal{F}$ and has at most $n-2$ elements, and so forth. Because the size of the finite set decreases by at least one each time, we can pass from a set in the family to a proper subset that also lies in the family at most n times. So the process must end eventually, and when that happens we get a minimal element of $\mathcal{F}$ that is contained in X. □

Exercise 2.19 establishes that Lemma 2.36b) need *not* hold when X is infinite.

THEOREM 2.37. *Let $n \in \mathbb{Z}^+$. Then:*

a) *For any isotone family $\mathcal{F} \subseteq 2^{[n]}$, we have*

$$m(\mathcal{F})^\uparrow = \mathcal{F}.$$

b) *For any Sperner family $\mathcal{F} \subseteq 2^{[n]}$, we have*

$$m(\mathcal{F}^\uparrow) = \mathcal{F}.$$

PROOF. a) Suppose $\mathcal{F}$ is an isotone family of subsets of $[n]$. Let $X \in \mathcal{F}$. Then $X \subseteq [n]$ is finite, so by Lemma 2.36b) there is $Y \subseteq X$ such that Y is a minimal element of $\mathcal{F}$. It follows that $X \in m(\mathcal{F})^\uparrow$.

Conversely, let $X \in m(\mathcal{F})^\uparrow$. Then there is a minimal element $Y \in \mathcal{F}$ such that $Y \subseteq X$; since $\mathcal{F}$ is isotone, it follows that $X \in \mathcal{F}$.

b) To say that $\mathcal{F}$ is a Sperner family is to say that $\mathcal{F} = m(\mathcal{F})$; Lemma 2.36a) tells us that $m(\mathcal{F}) = m(\mathcal{F}^\uparrow)$, so $\mathcal{F} = m(\mathcal{F}^\uparrow)$. □

Theorem 2.37 says that for families of subsets of $[n]$ (or really, for families of subsets of any finite set), we may trade in any isotone family for a Sperner family and we may trade in any Sperner family for an isotone family and these trades are "perfect" in the sense that if we trade and then trade back in either order, we get back to the family we started with. Or in the language of Chapter 8, we have a bijection from the set of isotone families of subsets of $[n]$ to the set of Sperner families of subsets of $[n]$. Since above we found a bijection from the set of isotone n-nary logical operators to the set of isotone families of subsets of $[n]$, composing these bijections, we get:

THEOREM 2.38. *Let $n \in \mathbb{Z}^+$. Then $X \mapsto m(\mathcal{F}(X))$ is a bijection from the set of n-ary logical operators to the set of Sperner families of subsets of $[n]$.*

EXAMPLE 2.39. *We will find all Sperner families of subsets of* $[3]$.
• *"Every" family of sets with zero elements is a Sperner family. The scare quotes are because there is only one family of sets with zero elements:* $\varnothing$.
• *Every family of sets with one element is a Sperner family. These are all of the form* $\{S\}$ *for a subset* $S \subseteq [3]$*, so there are eight of them.*
Notice that no family of subsets of $[n]$ *with more than one set can contain either* $\varnothing$ *or* $[n]$*, since for any nonempty subset* $A \subseteq [n]$ *we have* $\varnothing \subsetneq A$ *and for any proper subset* $A \subsetneq [n]$ *we have...*$A \subsetneq [n]$*. So every element of a family of at least* 2 *sets must be a set with either* 1 *or* 2 *elements.*
• *Suppose that* $\mathcal{F}$ *is a family of subsets of* $[3]$ *such that every set in* $\mathcal{F}$ *has* 1 *element. Then, since no one element set can properly contain another, all such families are Sperner families. There are as many such families as there are subsets of* $[3]$*, so there are* 8 *of them. But we have already counted* $\varnothing$*,* $\{\{1\}\}$*,* $\{\{2\}\}$ *and* $\{\{3\}\}$*, so we get four new families this way:* $\{\{1\},\{2\}\}$*,* $\{\{1\},\{3\}\}$*,* $\{\{2\},\{3\}\}$ *and* $\{\{1\},\{2\},\{3\}\}$.
• *Suppose that* $\mathcal{F}$ *is a family of subsets of* $[3]$ *such that every set in* $\mathcal{F}$ *has* 2 *elements. Since no two element subset can properly contain another, all such families are Sperner families. All such families arise as the complement of a unique Sperner family in which each set has one element, so we have* 8 *of these too, but as above, four of them have either* 0 *or* 1 *element so have been counted already. This gives* 4 *new families:* $\{\{1,2\},\{1,3\}\}$ *,*$\{\{1,2\},\{2,3\}\}$*,* $\{\{1,3\},\{2,3\}\}$ *and* $\{\{1,2\},\{1,3\},\{2,3\}\}$.
• *Finally we must find the Sperner families that contain a* 1 *element set and also a* 2 *element set. For this, the key point is that a two element subset of* $[3]$ *must exclusde exaclty one of* 1, 2, 3 *and the one element set must be the set containing that excluded element. So we get three such Sperner families:*

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

In total we found $1+8+4+4+3=20$ *Sperner families of subsets of* $[3]$*, whereas there are* $2^{2^3}=256$ *families of subsets of* $[3]$ *altogether. By Theorem 2.37 this means that* 20 *of the* 256 *ternary logical operators are isotone.*

For a positive integer n, we define the **nth Dedekind number** D_n to be the number of isotone n-ary Boolean functions, which by Theorem 2.37 is also equal to the number of Sperner families of subsets of $[n]$.

In truth, though Theorem 2.37 is *helpful* to compute the Dedekind numbers, their computation is nevertheless a daunting task. The following result records all Dedekind numbers that are exactly known. Note how recent the last part is!

THEOREM 2.40.

a) *(Church 1940)* $D_4 = 168$ *and* $D_5 = 7581$.
b) *(Ward 1946)* $D_6 = 7828354$.
c) *(Church 1965, Berman and Köhler 1976)* $D_7 = 2414682040998$.
d) *(Wiedemann 1991)* $D_8 = 561304372286875579077 8$.
e) *(Jäkel, Van Hirtum-De Causmaecker-Goemare-Kenter-Riebler-Lass-Plessl 2023)* $D_9 = 286386577668298411128469151667598498812366$.

Theorem 2.37 has other uses. For instance, it leads to an enlightening characterization of isotone logical operators. For this, we first observe that if we build up an expression out of isotone logical operations, the operator we get is again isotone.

To express this idea in full generality would require some heavy notation. In fact the following cases will be all we really need:

PROPOSITION 2.41. *Let $X(P_1, \ldots, P_n)$ and $Y(P_1, \ldots, P_n)$ be isotone logical operators. Then $X \wedge Y$ and $X \vee Y$ are also isotone.*

PROOF. We will treat the case of $X \wedge Y$ and leave $X \vee Y$ as Exercise 2.20. Let ℓ_1 and ℓ_2 be length n Boolean strings with $\ell_1 \prec \ell_2$ and suppose that $X \wedge Y$ evaluates to true on ℓ_1: let's abbreviate this as $(X \wedge Y)(\ell_1) = T$. Then $X(\ell_1) = T$ and $Y(\ell_1) = T$. Since X and Y are both isotone and $\ell_1 \prec \ell_2$, this means that $X(\ell_2) = T$ and $Y(\ell_2) = T$, so $(X \wedge Y)(\ell_2) = T$. So $X \wedge Y$ is isotone. □

The converse of Proposition 2.41 is almost true: if $X(P_1, \ldots, P_n)$ is a isotone logical operator *other than* the tautology (always true) or the contradiction (always false), then X can be expressed in terms of $P_1, \ldots, P_n$, $\wedge$ and $\vee$. You are asked to prove this in Exercise 2.21. Above we saw that we can build *all* logical operators in $P_1, \ldots, P_n$ using $\wedge$, $\vee$ and $\neg$ and that we did need $\neg$ to get them all. So this answers the more refined question of exactly which operators we *can* build without using $\neg$.

10. Exercises

EXERCISE 2.1. *Determine which of the following are statements. For whose which are statements, determine whether they are true or false. Please briefly explain your answers.*[11]

a) *For integers x and y, if $13x = 13y$, then $x = y$.*
b) $e^x > 0$.
c) *The integer x is a multiple of 16.*
d) *If an integer x is a multiple of 16, then x is a multiple of 32.*

EXERCISE 2.2. *Let P, Q and R be statements.*

a) *Show:* $(P \vee Q) \leftrightarrow (Q \vee P)$.
b) *Show:* $(P \wedge Q) \leftrightarrow (Q \wedge P)$.
c) *Show:* $(P \wedge Q) \wedge R \leftrightarrow P \wedge (Q \wedge R)$.
d) *Show:* $(P \vee Q) \vee R \leftrightarrow P \vee (Q \vee R)$.

In light of parts c) and d) we usually write $P \wedge Q \wedge R$ in place of either $(P \wedge Q) \wedge R$ or $P \wedge (Q \wedge R)$ and $P \vee Q \vee R$ in place of either $(P \vee Q) \vee R$ or $P \vee (Q \vee R)$.

EXERCISE 2.3. *(Logical Distributive Laws) Let P, Q and R be statements.*

a) *Show:* $P \vee (Q \wedge R) \leftrightarrow (P \vee Q) \wedge (P \vee R)$.
b) *Show:* $P \wedge (Q \vee R) \leftrightarrow (P \wedge Q) \vee (P \wedge R)$.

EXERCISE 2.4. *Write down a truth table with 4 rows and 16 columns that verifies that the 16 logical expressions written down in Example 2.13 are all logically inequivalent.*

EXERCISE 2.5. *Let P_1 and P_2 be statements.*

a) *Show that there is a logical expression $X(P_1, P_2)$ such that using P_1, P_2 and $X(P_1, P_2$ one can build both $\neg$ and $\wedge$. Hence all 2^{2^n} n-ary logical expressions can be built up out of $P_1, \ldots, P_n$ and the expression $X(P_1, P_2)$. (Hint: in fact we will have $X(P_1, P_1) = \neg P_1$.)*

[11] Warning: this exercise is somewhat subjective.

b) *Show that precisely* 2 *out of the* 16 *binary logical expressions have this property.*

EXERCISE 2.6. *Prove the following tautology:*[12] *For statements* P *and* Q*,*

$$(P \wedge (P \implies Q)) \implies Q.$$

EXERCISE 2.7. *Prove the following tautology: for any statements* P*,* Q *and* R *we have*

$$((P \implies Q) \wedge (Q \implies R)) \implies (P \implies R).$$

EXERCISE 2.8. *Prove the following tautology: for any statements* P*,* Q *and* R *we have*

$$((P \vee Q) \implies R) \iff ((P \implies R) \wedge (Q \implies R)).$$

EXERCISE 2.9. *Let* I *be a nonempty set, and for* $i \in I$*, let* P_i *be a statement. Show that the following are equivalent:*

(i) $(\forall i \in I,\ P_i$ *is true*$) \vee (\forall i \in I,\ P_i$ *is false*$)$.
(ii) *For all* $i, j \in I$*, we have* $P_i \implies P_j$.

EXERCISE 2.10. *Prove Proposition 2.17.*

EXERCISE 2.11. *How is Proposition 2.20 related to DeMorgan's Laws? Discuss.*

EXERCISE 2.12. *Let* A *and* B *be subsets of a "universal" set* X*. We write* A^c *and* B^c *for* $X \setminus A$ *and* $X \setminus B$*.*

a) *Show:* $A \subseteq B \iff B^c \subseteq A^c$.
b) *Explain why part a) is a set-theoretic analogue of the contrapositive.*

EXERCISE 2.13. *Let* $P(x, y)$ *be an open sentence with domain* $(x, y) \in \mathbb{R}^2$*.*

a) *Let*

$$\mathbb{T} := \{(x, y) \in \mathbb{R}^2 \mid P(x, y) \text{ is true}\}$$

be the truth locus of $P(x, y)$*, and let*

$$\mathbb{T}_x := \{x \in \mathbb{R} \mid \exists y \in \mathbb{R}\ P(x, y)\}$$

be the truth locus of $\exists y \in \mathbb{R}\ P(x, y)$*. In Example 2.21 there is a discussion of why* $\mathbb{T}_x$ *is the* ***projection*** *of* $\mathbb{T}$ *to the* x*-axis. Explain this in your own words, and draw at least one picture.*

b) *Let*

$$S := \{x \in \mathbb{R} \mid \forall y \in \mathbb{R}\ P(x, y)\}.$$

Describe S *in terms of* $\mathbb{T}$*.*

EXERCISE 2.14. *Let* S_1 *and* S_2 *be nonempty sets, and let* $P(x, y)$ *be an open sentence with domain* $(x, y) \in S_1 \times S_2$*. Let*

$$\mathbb{T} := \{(x, y) \in S_1 \times S_2 \mid\ P(x, y) \text{ is true}\}$$

be the truth locus of P*.*

a) *Consider the statement*

$$Q : \exists x \in S_1\ \forall y \in S_2\ P(x, y).$$

Explain what it means for Q *to be true in terms of* $\mathbb{T}$*.*
(Hint: this is closely related to Exercise 2.13b).)

[12] It is called **modus ponens**. This name is not however much used by those in mathematics but rather in other academic subjects in which logic arises, like philosophy.

b) *Consider the statement*

$$R : \forall y \in S_2 \ \exists x \in S_1 \ P(x, y).$$

Explain what it means for R to be true in terms of $\mathbb{T}$.
(Hint: this is related to Exercise 2.13a).)

c) *Show that $Q \implies R$.*

d) *Show that R* does not imply *Q by giving an example of an open sentence P for which R is true and Q is false.*

EXERCISE 2.15. *Give "good negations" for each of the following statements. (A good negation "brings the $\neg$ as far in as possible." In other words, you are allowed to negate primitive statements but not more complicated logical expressions.)*

a) $(P \wedge Q) \implies R$.

b) $\forall x \in \mathbb{R}, \ \exists y \in \mathbb{Z} \ |x - y| \leq \frac{1}{3}$.

c) *If I'm lying, I'm dying.*

d) $\exists x \in S \ P(x) \iff Q(x)$.
(Here $P(x)$ and $Q(x)$ are open sentences with domain $x \in S$.)

e) $\forall x \in S \ P(x) \implies Q(x)$.
(Here $P(x)$ and $Q(x)$ are open sentences with domain $x \in S$.)

f) *I want you and only you.*[13]

EXERCISE 2.16. *Show that the six binary logical operators of Example 2.32 are the only monotone binary logical operators.*

EXERCISE 2.17. *Let S be a set, and let $\mathcal{F}$ be a family of subsets of S. Show that*

$$\mathcal{F}^{\uparrow} \coloneqq \{T \in 2^S \mid T \supseteq U \text{ for some } U \in \mathcal{F}$$

is an isotone family of subsets of S.

EXERCISE 2.18. *Let S be a set, and let $\mathcal{F}$ be a family of subsets of S.*

a) *Suppose that every element of $\mathcal{F}$ is a finite set. Show: every element of $\mathcal{F}$ contains a minimal element of $\mathcal{F}$.*

b) *Suppose that $\mathcal{F}$ is itself a finite set. Show: every element of $\mathcal{F}$ contains a minimal element of $\mathcal{F}$.*

EXERCISE 2.19. *For $n \in \mathbb{Z}^+$, let $X_n = \mathbb{Z}^{\geq n} = \{n, n+1, n+2, \ldots\}$ be the set of integers that are at least n. Let*

$$\mathcal{F} \coloneqq \{X_n \mid n \in \mathbb{Z}^+\},$$

so that $\mathcal{F}$ is an (infinite) family of (infinite) subsets of $\mathbb{Z}^+$. Show that $\mathcal{F}$ has no minimal elements and thus it is very far from being true that every element of $\mathcal{F}$ is contained in a minimal element.

EXERCISE 2.20. *Complete the proof of Proposition 2.41 by showing that if X and Y are monotone n-ary logical operators, then so is $X \vee Y$.*

EXERCISE 2.21. *Let $X(P_1, \ldots, P_n)$ be a monotone logical operator.*

---

[13] Outside of math, I would say "You're the only one I want" intstead of "I want you and only you." But in mathematics, "only X" means "nothing other than X": it does *not* imply X!

a) *Suppose that X is neither the tautology nor the contradiction. Show that X can be expressed in terms of $P_1, \ldots, P_n$, $\wedge$ and $\vee$.*
(Hint: Use Theorem 2.37 and think about what this claim means in terms of Sperner families. Further hint: start by identifying which Sperner families correspond to operators using $\wedge$ alone.)

b) *Suppose that X is the tautology or the contradiction. Show that X* cannot *be expressed in terms of $P_1, \ldots, P_n$, $\wedge$ or $\vee$.*
(Hint: show that any such expression evaluates to T when $P_1, \ldots, P_n$ are all true and evaluates to F when $P_1, \ldots, P_n$ are all false.)

EXERCISE 2.22. *Let $X(P_1, \ldots, P_n)$ be an n-ary logical operator. We say that X is* ***antitone*** *if for any two length n Boolean lists ℓ_1 and ℓ_2, if $\ell_1 \prec \ell_2$ and $X(\ell_1)$ is false, then $X(\ell_2)$ is also false.*

a) *Show: X is isotone if and only if $\neg X$ is antitone.*
b) *Show: X is antitone if and only if $\neg X$ is isotone.*
c) *Deduce that the number of antitone n-ary logical operators is D_n, the nth Dedekind number.*

EXERCISE 2.23.

a) *You are shown a selection of cards, each of which has a single letter printed on one side and a single number printed on the other side. Then four cards are placed on the table (so that one side is visible and the other is not). On these cards you can see, respectively, D, K, 3 and 7. Here is a rule:*
"Every card that has a D on one side has a 3 on the other."
Your task is to select all those cards, but only those cards, which you must turn over in order to discover whether the rule has been violated. Which cards are these?

b) *You have been hired to watch, via closed-circuit camera, the bouncer at a certain 18-and-over club. In order to be allowed to drink inside the club, a patron must display valid 21-and-over ID to the bouncer, who then gives them a special bracelet. In theory the bouncer should check everyone's ID, but (assume for the purposes of this problem, at least!) it is not illegal for someone who is under 18 to enter the club, so you are not concerned about who the bouncer lets in or turns away, but only who gets a bracelet. You watch four people walk into the club, but because the bouncer is so large, sometimes he obscures the camera. Here is what you can see:*
- *The first person gets a bracelet.*
- *The second person does not get a bracelet.*
- *The third person displays ID indicating they are 21.*
- *The fourth person does not display any ID.*

You realize that you must enter the club and find some of the people to either check their ID's or see whether they got a bracelet. Precisely which people do you need to find to verify that the bouncer is obeying the law?

c) *Briefly compare and contrast parts a) and b).*

CHAPTER 3

Counting Finite Sets

1. Cardinality of a Finite Union

The basic problem we are interested in is the following: suppose we have finite sets $A_1, \ldots, A_N$, and let

$$A \coloneqq \bigcup_{i=1}^{N} A_i$$

be their union. Suppose that we know the sizes $\#A_1, \ldots \#A_N$ of these sets. Can we determine the size of A?

No, we cannot. Suppose A_1 is the set of people in Athens, GA who own a car and A_2 is the set of people in Athens, GA who own a cell-phone, so A is the set of people in Athens, GA who own a car or a cell-phone (or both). If $\#A_1 = n_1$ and $\#A_2 = n_2$, then what we can say is that

$$\max(n_1, n_2) \leq \#A \leq n_1 + n_2.$$

Here, for real numbers x and y, $\max(x, y)$ denotes the larger of x and y.

The first inequality is pretty clear: a subset of a finite set has size at most that of the set itself (Exercise 1.6), which shows that $n_1 \leq \#A$ and $n_2 \leq \#A$; since $\max(n_1, n_2)$ is equal to n_1 or n_2 (or both, if $n_1 = n_2$) it follows that $\max(n_1, n_2) \leq \#A$. As for the inequality $\#A \leq n_1 + n_2$, it can be seen as follows: if we make a list of everyone who has a car in Athens and then we make a list of everyone who has a cell-phone in Athens, and then tape the second list to the bottom of the first list, then we get a list whose associated set is A. However, $\#A$ could be smaller than $n_1 + n_2$, because even assuming that the first two lists were irredundant so that the first list ℓ_1 has size n_1 and the second list ℓ_2 has size n_2, the new list, say ℓ that we get by putting ℓ_1 and ℓ_2 together may be redundant, which by Exercise 1.4 is what would cause the associated set A to have cardinality less than the length of ℓ.[1]

This motivates us to establish several things in more generality and with more care, as we now do. First of all, if ℓ_1 and ℓ_2 are finite lists, it does make perfect sense to "combine them" into another finite list. This could be done in several different ways, but the way we just alluded to is perfectly good: if we have lists

$$\ell_1 : x_1, \ldots, x_m$$

of length m and

$$\ell_2 : y_1, \ldots, y_n$$

[1] In this case we certainly do have $\#A < n_1 + n_2$: not to brag too much, but I live in Athens and I have both a car and a cell-phone.

of length n, we define the **concatenated list**

$$\ell_1 + \ell_2 : x_1, \ldots, x_m, y_1, \ldots, y_n,$$

of length $m + n$. For that matter we can do the same with any finite number of finite lists $\ell_1, \ldots, \ell_N$. This allows us to establish the following:

THEOREM 3.1 (Sum Theorem).
Let $A_1, \ldots, A_N$ be finite sets, and put $A \coloneqq \bigcup_{i=1}^N A_i$. Then:

a) *We have*

$$\max(\#A_1, \ldots, \#A_N) \leq \#A \leq \#A_1 + \ldots + \#A_N.$$

b) *We have that $\#A = \#A_1 + \ldots + \#A_N$ if and only if the sets are pairwise disjoint: for all $1 \leq i \neq j \leq N$ we have $A_i \cap A_j = \varnothing$.*

PROOF. a) Let $1 \leq i \leq N$. Then $A_i \subseteq A$, so by Exercise 1.6 we have $\#A_i \leq \#A$. Again $\max(\#A_1, \ldots, \#A_N) = \#A_i$ for some $1 \leq i \leq N$, so we have

$$\max(\#A_1, \ldots, \#A_N) \leq \#A.$$

Now for $1 \leq i \leq N$, let ℓ_i be an irredundant finite list with associated set A_i, so by Exercise 1.4 the list ℓ_i has length $\#A_i$. Now form the list

$$\ell \coloneqq \ell_1 + \ldots + \ell_N,$$

it has length $\#A_1 + \ldots + \#A_N$ and associated set A since every element of A appears on at least one of the lists, and conversely every entry in each list is an element of A. By Exercise 1.5 this shows that $\#A \leq \#A_1 + \ldots + \#A_N$, completing the proof of part a).
b) Exercise 1.4 shows that $\#A$ is equal to the length of the list ℓ if and only if the list ℓ is irredundant. Since ℓ was contained by concatenating irredundant lists $\ell_1, \ldots, \ell_N$, a repetition in ℓ occurs if and only if there are $1 \leq i \neq j \leq N$ such that some entry on the list ℓ_i is equal to some entry on the list ℓ_j, which happens if and only if $A_i \cap A_j \neq \varnothing$. □

Despite the somewhat heavy notation and pedantic proof, the ideas behind Theorem 3.1 are very simple and familiar. The result says that if we partition a finite set into parts, then the size of the set is equal to the sum of the sizes of the parts: okay! On the other hand, if we express a finite set as a union of *overlapping* subsets, then the sum of the sizes of the subsets will be an overestimate on the size of the set...because elements in overlapping sets get counted at least twice.

Indeed, *overlap* is a key word here. In the motivating example, the extra piece of information we need is the number of Athenians who have both a car and a cell-phone: this number should be subtracted off from $\#A_1 + \#A_2$ (the number of people who have a car plus the number of people who have a cell-phone). So:

PROPOSITION 3.2. *Let A_1 and A_2 be finite sets, and put $A \coloneqq A_1 \cup A_2$. Then:*

$$\#A = \#A_1 + \#A_2 - \#(A_1 \cap A_2).$$

PROOF. Let us give two proofs.
FIRST PROOF: Let's go back to our construction with lists: let ℓ_1 be an irredundant finite list with associated set A_1, let ℓ_2 be a finite list without repeittions with associated set A_2, and let $\ell \coloneqq \ell_1 + \ell_2$. What we said so far is that ℓ is a finite list, of length $\#A_1 + \#A_2$ with associated set A, but if $A_1 \cap A_2 \neq \varnothing$ then it is

redundant. There is just one new thing to say: the number of repetitions in ℓ is $\#A_1 \cap A_2$. More precisely, each element $x \in A_1 \cap A_2$ appears (exactly once) in each of ℓ_1 and ℓ_2; if we remove the second occurrence of each such element x, then we have removed $\#(A_1 \cap A_2)$ elements to get a list ℓ' with no repeated elements and still with associated set A. Thus the size of A is the length of ℓ', which is the length of ℓ minus $\#(A_1 \cap A_2)$, which is $\#A_1 + \#A_2 - \#(A_1 \cap A_2)$.

SECOND PROOF: We define subsets

$$B_1 := A_1 \setminus A_2,$$

$$B_2 := A_1 \cap A_2,$$

$$B_3 := A_2 \setminus A_1.$$

As we can see immediately from a Venn diagram, we have

$$A = B_1 \coprod B_2 \coprod B_3,$$

so Theorem 3.1 gives

$$\#A = \#B_1 + \#B_2 + \#B_3.$$

A Venn diagram also easily shows us that

$$A_1 = B_1 \coprod B_2 \text{ and } A_2 = B_2 \coprod B_3,$$

so applying Theorem 3.1 once again, we get

$$\#A_1 + \#A_2 - \#(A_1 \cap A_2) = (\#B_1 + \#B_2) + (\#B_2 + \#B_3) - \#B_2$$

$$= \#B_1 + \#B_2 + \#B_3 = \#A. \qquad \square$$

Let's look for an analogue of Proposition 3.2 when we have three finite sets A_1, A_2 and A_3 and $A = A_1 \cup A_2 \cup A_3$. Again the interesting case is when the sets A_1, A_2, A_3 are not pairwise disjoint; in this case Theorem 3.1 tells us that

$$\#A < \#A_1 + \#A_2 + \#A_3,$$

but now "overlap" has become more complicated: presumably it should involve some or all of

$$A_1 \cap A_2,\ A_1 \cap A_3,\ A_2 \cap A_3,\ A_1 \cap A_2 \cap A_3.$$

If we guessed that "pairwise overlap was enough" we might try showing that $\#A$ is equal to

$$\#A_1 + \#A_2 + \#A_3 - \#(A_1 \cap A_2) - \#(A_1 \cap A_3) - \#(A_2 \cap A_3). \tag{7}$$

However in the case that $A = A_1 = A_2 = A_3$ the above formula evaluates to 0 rather than $\#A$. This may suggest that just as $\#A_1 + \#A_2 + \#A_3$ is an overestimate for $\#A$ and needs to be corrected by subtracting pairwise intersections, then (7) is an (albeit better) underestimate for $\#A$ and needs to be corrected by adding the triple intersection. This turns out to be true:

PROPOSITION 3.3. *Let A_1, A_2, A_3 be finite sets; put $A := A_1 \cup A_2 \cup A_3$. Then:*

(8)

$$\#A = \#A_1 + \#A_2 + \#A_3 - \#(A_1 \cap A_2) - \#(A_1 \cap A_3) - \#(A_2 \cap A_3) + \#(A_1 \cap A_2 \cap A_3).$$

One way to prove Proposition 7.3 is to use Venn diagrams as above: we break $A_1 \cup A_2 \cup A_3$ into the $2^3 - 1 = 7$ "elementary regions" that comprise the Venn diagram, one of which is $A_1 \cap A_2 \cap A_3$. Breaking A_1, A_2, A_3, $A_1 \cap A_2$, $A_1 \cap A_3$ and $A_2 \cap A_3$ into unions of elementary regions, one arrives a slightly tedious but totally doable calculation, which the reader is asked to do in Exercise 3.1.

This is however neither the most enlightening approach nor the one that is most amenable to general N, in which one would have to deal with a Venn diagram with $2^N - 1$ regions. Here is a different approach. The basic idea is to focus on how each element $x \in A$ contributes to the right hand side of (8). If you ask me to state that formally, I will respond that we want to check that for all $x \in A$, we have

$$1 = \#(A_1 \cap \{x\}) + \#(A_2 \cap \{x\}) + \#(A_3 \cap \{x\})$$

$$-\#(A_1 \cap A_2 \cap \{x\}) - \#(A_1 \cap A_3 \cap \{x\}) - \#(A_2 \cap A_3 \cap \{x\}) + \#(A_1 \cap A_2 \cap A_3 \cap \{x\}),$$

for then summing that over all $x \in A$ gives (8). But after seeing the argument you might want to think about how it makes the most sense to you.

Case 1: Suppose that x lies in exactly one of A_1, A_2, A_3, hence in none of the sets $A_1 \cap A_2$, $A_1 \cap A_3$, $A_2 \cap A_3$ or $A_1 \cap A_2 \cap A_3$, so it contributes 1 to the right hand side of (8).

Case 2: Suppose that x lies in exactly 2 of A_1, A_2, A_3, hence in exactly one of $A_1 \cap A_2$, $A_1 \cap A_3$, $A_2 \cap A_3$ and does not lie in $A_1 \cap A_2 \cap A_3$. Therefore x contributes $1 + 1 - 1 = 1$ to the right hand side of (8).

Case 3: Suppose that x lies in all of A_1, A_2, A_3 and hence also in all of $A_1 \cap A_2$, $A_1 \cap A_3$, $A_2 \cap A_3$ and $A_1 \cap A_2 \cap A_3$. Then overall x contributes $1+1+1-1-1+1 = 1$ to the right hand side of (8).

2. Independent Choices and Cartesian Products

How many cards are in a standard deck of playing cards? 52. Why?

That's a bit of a strange question, but we can answer it in terms of some other familiar facts:

- Every card has exactly one of four suits: clubs, diamonds, hearts or spades.
- Every card has exactly one of thirteen values: two, three, four, five, six, seven, eight, nine, ten, jack, queen, king, or ace.
- Every possible combination of suit and value occurs exactly once.

Given this, the answer is that $4 \cdot 13 = 52$.

This is an example of the **Principle of Independent Choices**. Suppose that we have two choices to make. For each choice, there are finitely many options: say that we have $n_1 \in \mathbb{Z}^+$ options for the first choice and $n_2 \in \mathbb{Z}^+$ options for the second choice. Suppose also that the choices are *independent* in the sense that whatever we choose among the first set of options does not restrict our choice among the second set of options. Then the total number of ways to make these choices is $n_1 \cdot n_2$.

We can state a result with any finite number $N \in \mathbb{Z}^+$ of choices instead of two:

PROPOSITION 3.4 (Principle of Independent Choices). *Let $N \in \mathbb{Z}^+$. Suppose that we have N choices to make. For $1 \leq i \leq N$, suppose that we have $n_i \in \mathbb{Z}^+$ options for the ith choice. Suppose further that the choices are independent, in the sense that all combinations of choices among different options are possible. Then the total number of ways to make these N choices is $n_1 \cdots n_N$.*

What do we think about Proposition 3.4? Is it obvious? Does it require proof?

My answers are "yes" and "yes, if possible." Notice that in the proof of Proposition 1.21 we already used the Principle of Indpendent Choices to argue that a set with N elements has 2^N subsets: we observed that forming a subset of $\{x_1, \ldots, x_N\}$ amounts to successively choosing to include or exclude x_i for $1 \leq i \leq N$; this gives N independent choices with two options each, hence by Proposition 3.4 there are 2^N options all together.

But in theoretical mathematics everything must be proved in terms of some set of assumptions, including basic ones called **axioms**. Although we do not need to have the basic axioms in full view at all (or even most) times, nevertheless in principle there should be a *fixed* set of axioms: we don't add to our list of axioms everytime something sounds obvious and we don't know how to prove it. This would turn mathematics into some kind of second-rate comic book, where the hero resolves their latest scrape by means of a superpower that has never been mentioned before that issue.

In this course we do not consider formal axioms for sets, but rather manipulate them in intuitively plausible ways (that can, and have, been justified by formal axioms). We also do not construct any of the basic number systems or the basic operations on them, and therefore some properties of numbers will not be provable by us because the proofs are too closely bound to the constructions of these systems. Are we in one of these cases now?

Sort of. To clarify what is going on, we observe that Proposition 3.4 is equivalent to the following result.

PROPOSITION 3.5 (Cardinality of Finite Cartesian Products). *Let $N \in \mathbb{Z}^+$, and let $A_1, \ldots, A_N$ be finite sets. Then*

$$\#(A_1 \times \ldots A_N) = (\#A_1) \times \ldots (\#A_N).$$

To see the relationship between Propositions 3.4 and 3.5, we observe that in the setting of Proposition 3.4 we can let A_1 be the set of options for the first choice, A_2 be the set of options for the second choice, and so forth, finally letting A_N be the set of options for the Nth choice. Then a collection of independent choices from these sets is precisely a finite list of length N,

$$\ell : c_1, \ldots, c_N$$

with $c_i \in A_i$. As above, just by adding parentheses: $c_1, \ldots, c_N \mapsto (c_1, \ldots, c_N)$, such lists may be identified with elements of the Cartesian product $A_1 \times \ldots \times A_N$. This shows that Proposition 3.5 implies Proposition 3.4, and the converse implication is

almost the same.[2]

Let us now concentrate on Proposition 3.5 instead, since it is formulated in terms of sets rather than choices. Let us first concentrate on the case $N = 2$: if we have two finite sets A_1 and A_2 then

$$\#(A_1 \times A_2) = \#A_1 \times \#A_2.$$

How would we prove this?

My answer is that this is a key interpretation of multiplication that we learn in elementary school. At an early age we learn that for positive integers m and n, the product $m \times n$ is the number of dots in an m by n rectangular array of dots. This is the content of Proposition 3.5. But that's not all we learn in elementary school about $m \times n$. Another explanation of $m \times n$ is as repeated addition:

$$m \times n = n + n + \ldots n \ (m \text{ times}).$$

Geometrically, this corresponds to regarding the rectangular array of m by n dots as m columns of n dots each. (We could also swap the rows and the columns here, of course.) This gives us a proof idea:

If $\#A_1 = m$ and $\#A_2 = n$, write $A_1 = \{x_1, \ldots, x_m\}$ and $A_2 = \{y_1, \ldots, y_n\}$. Then we can partition $A_1 \times A_2$ into "columns"

$$C_1 \coloneqq \{(x_1, y_1),\ (x_1, y_2), \ldots, (x_1, y_n)\},$$

$$C_2 \coloneqq \{(x_2, y_1),\ (x_2, y_2), \ldots, (x_2, y_n)\},$$

$$\vdots$$

$$C_m \coloneqq \{(x_m, y_1),\ (x_m, y_2), \ldots, (x_m, y_n).$$

Each C_i has n elements. So by Theorem 3.1 we have

$$\#(A_1 \times A_2) = \#C_1 + \ldots + \#C_m = n + \ldots + n \ (m \text{ times}) = m \times n = \#A_1 \cdot \#A_2.$$

This proves the $N = 2$ case of Proposition 3.5 (and hence also of Proposition 3.4). The general case requires no new ideas. For instance, for $N = 3$ we have

$$\#A_1 \times A_2 \times A_3 = \#((A_1 \times A_2) \times A_3),$$

and then using the $N = 2$ case twice we get

$$\#(A_1 \times A_2) \times A_3) = \#(A_1 \times A_2) \cdot \#A_3 = \#A_1 \times \#A_2 \times \#A_3.$$

For general N this is a textbook case for *mathematical induction*. Literally: we will learn this later, and completing the proof of Proposition 3.5 will be an exercise.

[2]There is one small wrinkle: in Proposition 3.5 the sets $A_1, \ldots, A_N$ are allowed to be empty; if $A_i = \varnothing$ for some i, then the Cartesian product $A_1 \times \ldots \times A_N$ is empty (and conversely). On the choosing side, a choice with 0 options sounds ominous. But this is a rather trivial case.

3. Counting Irredundant Lists and Subsets

Here is another set-theoretic counting problem. It is easy but useful, and it serves as a steppingstone to more interesting set-theoretic counting problems.

For positive integers n and k, let $P(n,k)$ be the number of irredundant finite lists of length k drawn from an n element set, say from $[n] = \{1, \ldots, n\}$.

PROPOSITION 3.6. *Let $n, k \in \mathbb{Z}^+$.*

a) *If $n < k$, then $P(n,k) = 0$.*

b) *If $n \geq k$, then $P(n,k) = n(n-1)\cdots(n-k+1)$.*

PROOF. a) If ℓ is an irredundant list of length k with entries in $[n]$, then the associated set S has size k and is a subset of $[n]$, $k = \#S \leq \#[n] = n$. Therefore if $n < k$ there are no such lists, so $P(n,k) = 0$.

b) To build an irredundant list

$$\ell : x_1, \ldots, x_k$$

of length k from $[n] = \{1, \ldots, n\}$, we have n choices for the first entry x_1. Having chosen x_1, we cannot choose it again, so we have $n-1$ choices for the second entry x_2. It continues in this manner: for each entry we have one fewer choice than we did before, so the number of such lists is

$$n \cdot (n-1) \cdots (n-(k-1)). \qquad \square$$

For later use, it is also convenient to define $P(n,k)$ when one or both of n and k are allowed to be zero. Namely, for all $n \geq 0$, we put

$$P(n,0) \coloneqq 1.$$

This corresponds to the fact that there is exactly one list of length zero with elements drawn from any set, namely the empty list. Also, for all $k \geq 1$ we put

$$P(0,k) \coloneqq 0.$$

This corresponds to the fact that there is no list of positive length with elements drawn from the empty set.

The counting problem solved in the proof of Proposition 3.6 may help to clarify what "independent choices" mean, because this time our choices are *dependent*: each choice we make restricts our optons on the next choice.

Once again, that this is the solution to our (simple) counting problem seems very intuitive, but one could ask for an argument from basic principles. For this we can also make the following argument: if

$$\ell : x_1, \ldots, x_k$$

is an irredundant list of length k drawn from $[n] = \{1, \ldots, n\}$, let

$$\ell' : x_2, \ldots, x_k.$$

Then

$$\ell = x_1 + \ell',$$

and ℓ' is an irredundant list drawn from the set $[n] \setminus \{x_1\}$, which has $n-1$ elements. Conversely, for any element $x_1 \in [n]$, if ℓ' is an irredundant list of length $k-1$ with elements drawn from the $n-1$ element set $[n]\setminus\{x_1\}$, then $x_1 + \ell'$ is an irredundant list

of length k drawn from the set $[n]$. Thus we have partitioned the set of irredundant lists of length k into n different subsets – according to the first entry in the list – each of which has $P(n-1,k-1)$ elements. This shows the following:

PROPOSITION 3.7. *For all $1 \leq k \leq n$ we have*

$$P(n,k) = nP(n-1,k-1). \tag{9}$$

The formula (9) allows us to calculate $P(n,k)$ in all cases: e.g. we have

$$P(5,3) = 5P(4,2) = 5 \cdot 4 \cdot P(3,1) = 5 \cdot 4 \cdot 3 \cdot P(2,0) = 5 \cdot 4 \cdot 3 \cdot 1 = 5 \cdot 4 \cdot 3.$$

To formally derive Proposition 3.6b) from (9) is good exercise in mathematical induction, which will occur later in this course.

Here is a useful piece of notation: we put

$$n! \coloneqq P(n,n).$$

Thus

$$0! = 1$$

and for $n \geq 1$ we have

$$n! = n(n-1)\cdots(n-n+1) = n(n-1)\cdots 1.$$

Now for non-negative integers k and n, we denote by $\binom{n}{k}$ the number of k element subsets of an n element set, say of $[n] = \{1,\ldots,n\}$.

It turns out that the quantities $\binom{n}{k}$ and $P(n,k)$ are closely related. Indeed, every k-element subset of $[n]$ is associated to an irredundant list of length k. However, in most cases different irredundant lists will yield the same set, e.g.

$$\ell_1 : 1,3,2,4 \text{ and } \ell_2 : 4,1,2,3$$

both yield the set $\{1,2,3,4\}$.

In order to "fix this," we need to count the number of irredundant lists that yield the same associated set S. But these are the irredundant lists of length k with elements drawn from the k-element set S, so there are precisely $P(k,k) = k!$ of them. This argument shows that

$$P(n,k) = \binom{n}{k} P(k,k). \tag{10}$$

Since $P(k,k) = k!$ is always a positive integer, we deduce:

PROPOSITION 3.8. *For all $n,k \geq 0$ we have*

$$\binom{n}{k} = \frac{P(n,k)}{P(k,k)} = \frac{n(n-1)\cdots(n-k+1)}{k!} = \frac{n!}{k!(n-k)!}. \tag{11}$$

PROOF. Dividing (10) by $P(k,k)$ gives the first equality, and the second equality just uses that $P(k,k) = k!$. Finally, we have

$$n(n-1)\cdots(n-k+1) = \frac{n!}{(n-k)!}.$$

□

In particular $\binom{n}{k}$ is always a non-negative integer. It is positive if and only if $P(n,k)$ is positive, namely when $k \leq n$. Indeed this is the condition for an n-element set to have at least one k-element subset.

The expression $\binom{n}{k}$ is called a "binomial coefficient" for reasons that we explain in the next section.

There are many identities involving binomial coefficients. Here is one of the most basic and useful:

PROPOSITION 3.9. *For $n, k \in \mathbb{Z}^+$, we have*

$$\binom{n}{k} = \binom{n-1}{k-1} + \binom{n-1}{k}. \tag{12}$$

PROOF. We have that $\binom{n}{k}$ is the number of k-element subsets of $[n] = \{1, \ldots, n\}$. Since every subset $S \subseteq [n]$ either contains n or not and not both, if N_1 is the number of k-element subsets containing n and N_2 is the number of k-element subsets not containing n, then

$$\binom{n}{k} = N_1 + N_2.$$

Moreover, if $S \subseteq [n]$ is a k-element subset containing n, then $S \setminus \{n\}$ is a $(k-1)$-element subset of $[n-1]$, and conversely if T is a $(k-1)$-element subset of $[n-1]$ then $T \cup \{n\}$ is a k-element subset of $[n]$ containing n, which shows that

$$N_1 = \binom{n-1}{k-1}.$$

Similarly but more easily, if $S \subseteq [n]$ is a k-element subset not containing n, then S is a k-element subset of $[n-1]$, and conversely. So

$$N_2 = \binom{n-1}{k}.$$

This implies the result. □

PROPOSITION 3.10 (Zhu-Vandermonde). *For all $m, n, r \in \mathbb{N}$, we have*

$$\binom{m+n}{r} = \sum_{k=0}^{r} \binom{m}{k}\binom{n}{r-k}.$$

PROOF. This provides us with an excellent first opportunity for a **combinatorial proof**. This is a method of proving an algebraic identity by showing that both sides count the number of elements of the same finite set.

In this case, the left hand side $\binom{m+n}{r}$ has the more immediate combinatorial interpretation: it is the number of r element subsets of a set S of size $m+n$. To bring the right hand side into the picture, we imagine that the elements of the set S are balls and that we have m red balls and n blue balls. Then for an r-element subset T of S, we can consider the number k of red balls in T. The number of subsets of S that have precisely k red balls is $\binom{m}{k}\binom{n}{r-k}$, because we have k ways of choosing from among the m red balls of S, and then we are left to choose $r-k$ of the n blue balls from S. Since we get all possible r-element subsets of S by letting the number k of red balls range over $0 \leq k \leq r$, this shows that $\sum_{k=0}^{r} \binom{m}{k}\binom{n}{r-k}$ also counts the number of r-element subsets of S and completes the proof. □

4. The Binomial Theorem

The following identities are probably familiar:

$$(x+y)^2 = x^2 + 2xy + y^2,$$

$$(x+y)^3 = x^3 + 3x^2y + 3xy^2 + y^3.$$

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

In these identities it actually doesn't matter much what x and y are, as long as we can add and multiply them subject to some familiar properties for addition and multiplication – certainly including $xy = yx$. In order not to get derailed by abstract algebraic considerations, let us just assume that x and y are real numbers.

We want a general formula for the expansion of the binomial $(x+y)^n$ for a positive integer n. Our main task is to understand the connection to the material of the previous section. To see this, let's look at an example:

$$(x+y)^3 = (x+y)(x+y)(x+y) = xxx + xxy + xyx + xyy + yxx + yxy + yyx + yyy.$$

We have eight terms in the sum. Why? To get a term of $(x+y)(x+y)(x+y)$ we must choose either x or y for each of the three instances of $(x+y)$. By the Principle of Independent Choices we are making three independent binary (i.e., with two different options each: here, x versus y) choices, and this leads to $2 \cdot 2 \cdot 2 = 2^3 = 8$ terms. Because $xy = yx$ it doesn't matter in what order the factors occur: e.g.

$$yxx = xyx = xxy.$$

So every term is going to be of the form $x^{n-k}y^k$ for some $0 \leq k \leq 3$.

For a general n we will have 2^n terms altogether, and every term can be written in the form $x^{n-k}y^k$ for some $0 \leq k \leq n$. The question becomes: for each fixed $0 \leq k \leq n$, how many times do we get $x^{n-k}y^k$?

As we saw in Proposition 1.21, a collection of n independent binary choices can be used to build a subset of $[n] = \{1, \ldots, n\}$ and conversely: here, the elements of the subset are the "indices" of the factors in which we choose y. Perhaps this is most clear if we spell it out completely for $n = 3$:

$$xxx \leftrightarrow \varnothing,$$

$$xxy \leftrightarrow \{3\},$$

$$xyx \leftrightarrow \{2\},$$

$$xyy \leftrightarrow \{2,3\},$$

$$yxx \leftrightarrow \{1\},$$

$$yxy \leftrightarrow \{1,3\},$$

$$yyx \leftrightarrow \{1,2\},$$

$$yyy \leftrightarrow \{1,2,3\}.$$

Notice that the size of the subset is equal to the number of instances of y in the list, which in turn is equal to the k when we rewrite the term as $x^{n-k}y^k$. Therefore the number of terms $x^{n-k}y^k$ is equal to the number of k element subsets of $[n]$, which is $\binom{n}{k}$. Aha! This establishes an important result:

THEOREM 3.11 (Binomial Theorem). *For real numbers x and y and a non-negative integer n we have*

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

The following is an immediate consequence:

COROLLARY 3.12. *For $n \in \mathbb{Z}^+$, we have*

$$\binom{n}{0} - \binom{n}{1} + \binom{n}{2} - \ldots + (-1)^n\binom{n}{n} = 0.$$

PROOF. We apply the Binomial Theorem with $x = 1$ and $y = -1$. □

5. The Inclusion-Exclusion Principle

Finally, we return to the problem that we had only solved in special cases: if we have finite sets $A_1, \ldots, A_N$, we want a formula for the size of

$$A \coloneqq \bigcup_{i=1}^N A_i$$

in terms of the sizes of the sets A_i and their various intersections. Perhaps you know what the formula should be; even so, writing it down is a bit of a chore. Fix $1 \leq k \leq N$. By a **k-wise intersection** of the sets $A_1, \ldots, A_N$, we mean an intersection over any k of the sets. More formally, for each k-element subset $S \subseteq [N]$ we define

$$A_S \coloneqq \bigcap_{i \in S} A_i.$$

Let us also write $\binom{[N]}{k}$ for the set of all k-element subsets of $[N]$. Then we want our formula for $\#A$ to involve the terms

$$S_k(A_1, \ldots, A_N) \coloneqq \sum_{S \in \binom{[N]}{k}} \#A_S.$$

For example, we have

$$S_1(A_1, \ldots, A_N) = \#A_1 + \ldots + \#A_N$$

is the sum of the sizes of the sets themselves,

$$S_2(A_1, \ldots, A_N) = \#(A_1 \cap A_2) + \#(A_1 \cap A_3) + \ldots + \#(A_1 \cap A_N) + \#(A_2 \cap A_N) + \ldots + \#(A_{N-1} \cap A_N),$$

is the sum of the sizes of their pairwise intersections, and

$$S_N(A_1, \ldots, A_N) = \#(A_1 \cap A_2 \cap \ldots \cap A_N)$$

the size of their mutual intersection.

THEOREM 3.13 (Principle of Inclusion-Exclusion). *For finite sets $A_1, \ldots, A_N$ and $A = \bigcup_{i=1}^N A_i$, we have*

$$\#A = S_1(A) - S_2(A) + \ldots + (-1)^{N+1}S_N(A). \tag{13}$$

PROOF. We adapt the proof of Proposition 7.3, which is the $N = 3$ case. Namely, for each $x \in A$ we need to show that x contributes 1 to $S_1(A) - S_2(A) + \ldots + (-1)^N S_N(A)$ in the above sense. Let

$$T := \{i \in [N] \mid x \in A_i.\}.$$

That is, T keeps track of the indices of the subsets in which x lies. We put

$$n := \#T.$$

Then:
• The element x lies in exactly n of the sets $A_1, \ldots, A_N$ so contributes k to $S_1(A)$.
• The pairwise intersections $A_i \cap A_j$ that contain x are indexed by the elements of $\binom{T}{2}$, i.e., the 2-element subsets of T, of which there are $\binom{k}{2}$. So the element x contributes $\binom{n}{2}$ to $S_2(A)$.
• Similarly, for any $1 \leq k \leq n$, the k-fold intersections that contain x are indexed by the elements of $\binom{T}{k}$, i.e., by the k-element subsets of T, of which there are $\binom{n}{k}$. So the element x contributes $\binom{n}{k}$ to $S_k(A)$.
• If $k > n$ then x lies in no k-fold intersection, so contributes nothing to $S_k(A)$.

Therefore the entire contribution of x to the right hand side of (13) is

$$\binom{n}{1} - \binom{n}{2} + \binom{n}{3} - \ldots + (-1)^{n+1}\binom{n}{n} = 1.$$

Here the last equality follows by applying Corollary 3.12: bring all the terms except for $\binom{n}{0}$ to the other side of the equation. This shows that indeed x contributes one to the right hand side of (13); certainly x contributes to 1 to the left hand side, establishing the result. □

Theorem 3.13 is clearly the most complex result we have encountered so far. It is also by far the most interesting and useful. In fact it has been applied thousands of times over in various contexts and also vastly generalized. We will now give one relatively simple, but beautiful, application.

For $N \in \mathbb{Z}^+$, we call elements of $\mathcal{P}(N, N)$ **permutations on N elements**. Thus a permutation on N elements is a finite irredundant list

$$\ell : x_1, \ldots, x_N$$

of length N drawn from $[N] = \{1, \ldots, N\}$ and thus consists of all the integers from 1 to N written in some order. We may thus think of ℓ as giving a "reordering" of the standard permutation $1, 2, 3, \ldots, N$. For $1 \leq i \leq N$ we say that i is a **fixed point** of the permutation ℓ if $x_i = i$. We can think of a fixed point as an element that "keeps its place" under the reodering given by ℓ. For example, the $3! = 6$ elements of $P(3, 3)$ and their fixed points are:

$\ell_1 : 1, 2, 3$. The points 1,2,3 are fixed points.

$\ell_2 : 1, 3, 2$.The point 1 is a fixed point.

$\ell_3 : 2, 1, 3$.The point 3 is a fixed point.

$\ell_4 : 2, 3, 1$.There are no fixed points.

$\ell_5 : 3, 1, 2$.There are no fixed points.

$\ell_6 : 3, 2, 1$.The point 2 is a fixed point.

We notice that two out of the six elements of $P(3,3)$ have no fixed points at all: in the reordering, no element stays in its original place.

A **derangement** is a permutation without fixed points. Let $\mathcal{D}_N$ be the set of permutations on N elements that are derangements. Thus above we computed that $\#\mathcal{D}_3 = 2$. Similarly but more easily we see that $\mathcal{D}_1 = 0$ and $\mathcal{D}_2 = 1$. In Exercise 3.6 you are asked to find all elements of $\mathcal{D}_4$ and thereby show that $\#\mathcal{D}_4 = 9$.

It is not hard to see that $\#\mathcal{D}_N$ increases rapidly with N, so in some ways it is more interesting to consider the fraction $\frac{\#\mathcal{D}_N}{P(N,N)} = \frac{\#\mathcal{D}_N}{N!}$. Using the most basic language of probability, this fraction is the probability that a permtuation on N elements has no fixed points.[3]

We will apply Theorem 3.13 will be to compute $\#\mathcal{D}_N$ for all positive integers N, and then just by dividing by $N!$ we will get a (nicer!) formula for $\frac{\#\mathcal{D}_N}{P(N,N)}$.

Let A be the set of permutations on N elements with a fixed point. Thus $\mathcal{D}_N = \mathcal{P}(N,N) \setminus A$, so

$$\#\mathcal{D}_N = N! - \#A,$$

so it suffices to compute $\#A$. Here's the key observation: for $1 \leq i \leq N$, let A_i be the set of permutations for which i is a fixed point. Since a permutation has a fixed point if and only if i is a fixed point for some $1 \leq i \leq N$, we have

$$A = \bigcup_{i=1}^{N} A_i.$$

Now we will apply Theorem 3.13 to $A_1, \ldots, A_N$. Fix $1 \leq k \leq N$ and let S be a k-element subset of $[N]$. Then $A_S = \bigcap_{i \in S} A_i$ is the set of permutations on N elements in which every i in S is a fixed point. Removing all the elements of S from the list ℓ, we get an irredundant list of length $N-k$ with elements drawn from the $(N-k)$-element subset $[N] \setminus S$, and every irredundant list of length $N-k$ with elements drawn from $[N] \setminus S$ arises exactly once in this way. This shows that

$$\#A_S = P(N-k, N-k) = (N-k)!.$$

Since this holds for every $S \in \binom{[N]}{k}$ and there are $\binom{N}{k}$ such subsets, overall we have

$$S_k := S_k(A_1, \ldots, A_N) = \binom{N}{k}(N-k)! = \frac{N!}{k!}.$$

Therefore Theorem 3.13 gives:

$$\#A = S_1 - S_2 + \ldots + (-1)^{N+1} S_N = \sum_{k=1}^{N} (-1)^{k+1} \frac{N!}{k!}.$$

It follows that

$$\frac{\#\mathcal{D}_N}{P(N,N)} = \frac{N! - \#A}{N!} = 1 + \sum_{k=1}^{N} \frac{(-1)^k}{k!} = \sum_{k=0}^{N} \frac{(-1)^k}{k!}. \tag{14}$$

[3]Although we neither develop nor use probability in this course, there is really nothing up our sleeves: if we have a finite nonempty set T of outcomes that we consider equally likely, for a subset $S \subseteq T$, the probability that the outcome lies in S is defined to be $\frac{\#S}{\#T}$.

Equation (14) is really remarkable. Let us borrow from calculus the fact that for all $x \in \mathbb{R}$ we have

$$e^x = \sum_{k=0}^{\infty} \frac{x^k}{k!}.$$

It then follows that

$$\lim_{N\to\infty} \frac{\#\mathcal{D}_N}{P(N,N)} = \sum_{k=0}^{\infty} \frac{(-1)^k}{k!} = e^{-1} \approx 0.36787944117.$$

This means that as the length N of the permutation increases without bound, the probability that a randomly chosen permutation of length N is a derangement converges to $\frac{1}{e}$. Still borrowing from caclulus, the error estimate that accompanies the Alternating Series Test [**Cl-HC**, Thm. 11.28b)] gives

$$\left| \frac{\#\mathcal{D}_N}{N!} - \frac{1}{e} \right| < \frac{1}{(N+1)!},$$

which implies that the convergence is very rapid indeed. For instance:
• Since $\frac{1}{(N+1)!} < \frac{1}{100}$ for all $N \geq 4$, the probability that a permutation of length N is a derangement lies within .01 of $\frac{1}{e}$ for all $N \geq 4$.
• Since $\frac{1}{(N+1)!} < \frac{1}{10^{10}}$ for all $N \geq 13$, the probability that a permutation of length N is a derangement lies within $\frac{1}{10^{10}}$ of $\frac{1}{e}$ for all $N \geq 13$.

Exercise 3.7 gives a real life application of this fact.

6. The Pigeonhole Principle

The traditional statement of the Pigeonhole Principle is as follows: suppose we have a (finite!) flock of pigeons that return from flight to a mesh of pigeonholes. Each hole is large enough to contain several pigeons. Suppose that the number of pigeons, say P, exceeds the number of holes, say H. Then, once all the pigeons land, there must be at least one hole that contains at least two pigeons.

Why? Well, suppose not, label the holes 1 through H, and for $1 \leq i \leq H$ let p_i be the number of pigeons that land in the ith hole. Then by assumption we have $p_i \in \{0, 1\}$ for all i, and so on the one hand we have

$$p_1 + \ldots + p_H = P$$

and on the other hand we have

$$p_1 + \ldots + p_H \leq 1 + \ldots + 1 = H,$$

so

$$P = p_1 + \ldots + p_H \leq H,$$

contradicting our assumption that $P > H$. (Later we will recognize this as better phrased as a proof by contrapositive.)

A little thought shows that when the number of pigeons is significantly greater than the number of holes, we can make a stronger conclusion.

EXAMPLE 3.14. *Suppose that* 50 *people are waiting to go through airport security, and they can choose any of four different lines, each with a TSA agent waiting at the end. If everyone lines up then at least one line must contain at least* 13

people: if not, all four lines would contain at most 12 *people, and then there could be at most* 48 *people altogether.*

To enunciate the stronger conclusion is not so hard, but to do so it is useful to introduce the following notation: for a real number x, the **ceiling of x** $\lceil x \rceil$ is the least integer n such that $n \geq x$. In other words, $\lceil x \rceil = x$ if x is already an integer, and if not then we round up to the nearest integer.

In Exercise 3.8 you are asked to show: if $x \in \mathbb{Z}$, $y \in \mathbb{R}$ and $x \geq y$, then also $x \geq \lceil y \rceil$.

THEOREM 3.15 (Strong Pigeonhole Principle). *Let $P, H \in \mathbb{Z}^+$. If P pigeons fly into H holes, then at least one hole contains at least $\lceil \frac{P}{H} \rceil$ pigeons.*

PROOF. Again label the holes 1 through H, and for $1 \leq i \leq H$ let p_i be the number of pigeons that land in the ith hole, so we have

$$p_1 + \ldots + p_H = P.$$

Suppose that we had $p_i < \frac{P}{H}$ for all $1 \leq i \leq H$. Then

$$P = p_1 + \ldots + p_H < \frac{P}{H} + \ldots + \frac{P}{H} \ (H \text{ times}) \ = P,$$

which is a contradiction. Therefore for at least one i we have $p_i \geq \frac{P}{H}$. Since p_i is the size of a finite set, it is an integer, so by Exercise 3.8 we conclude $p_i \geq \lceil \frac{P}{H} \rceil$. □

It may be surprising that in Theorem 3.15 the hypothesis that the number of pigeons was greater than the number of holes did not appear. However if $P \leq H$ then $0 < \lceil \frac{P}{H} \rceil \leq 1$, so the conclusion is that at least one hole contains at least one pigeon: true, but not profound.

It should be clear that the Pigeonhole Principle is not really about pigeons. It is really a statement about functions from one finite set to another. We will speak about functions later on in this text. For now we observe that it has a formulation in terms of finite lists: indeed, the basic form of the Pigeonhole Princple is equivalent to the following assertion: if $P > H$ are positive integers and

$$\ell : x_1, \ldots, x_P$$

is a finite list of length P with elements drawn from a finite set S of size H, then the list ℓ is redundant. Indeed, if ℓ were irredundant than the associated set $T \coloneqq \{x_1, \ldots, x_P\}$ would on the one hand have P elements and on the other hand be a subset of S hence should have at most $\#S = H < P$ elements, a contradiction. In Exercise 3.9 you are asked to interpret Theorem 3.15 as a statement involving finite lists and then prove it directly using lists. The proof is however not much different from the one we gave.

We are about to give a (famous) example of a statement proved via the Pigeonhole Principle. First, we recall that two positive integers x and y are **coprime** if they have no common divisor $d > 1$. Any two consecutive positive integers N and $N+1$ must be relatively prime, since if N is exactly divisble by $d > 1$, then $N+1$ leaves a remainder of 1 upon division by d and thus cannot be exactly divisible by d.

PROPOSITION 3.16. *Let $n \in \mathbb{Z}^+$. If S is a subset of $[2n] = \{1, \ldots, 2n\}$ of size at least $n+1$, then S contains coprime integers x and y.*

PROOF. We let the pigeons be the elements of S and the pigeonholes be

$$H_1 \coloneqq \{1,2\},\ H_2 \coloneqq \{3,4\}, \ldots, H_i \coloneqq \{2i-1, 2n\}, \ldots, H_n \coloneqq \{2n-1, 2n\}.$$

The pigeon $x \in S$ flies into the hole H_i if and only if $x \in H_i$: then each pigeon goes into exactly one hole because $\{H_1, \ldots, H_n\}$ is a partition of $[2n]$. By assumption we have more pigeons than holes, so two pigeons must fly in the same hole: for some $1 \leq i \leq n$ we have $2i-1 \in S$ and also $2i \in S$. As said above, any two consecutive positive integers are coprime, so $2i-1$ and $2i$ are coprime elements of S. □

In Proposition 3.16, if $S \subsetneq [2n]$ had size at most n, then the Pigeonhole Principle argument would not work: we need more pigeons than holes. And indeed the conclusion can fail for sets with at most n elements: take e.g.

$$S = \{2, 4, 6, \ldots, 2n\}.$$

Then $\#S = n$ and every element of S is divisible by 2.

In Exercise 3.11 you are asked to show that under the hypotheses of Proposition 3.16 there are always $x, y \in S$ such that y is a multiple of x.

PROPOSITION 3.17. *A family $\{X_i\}_{i \in I}$ is called **pairwise intersecting** if for all $i, j \in I$ we have $X_i \cap X_j \neq \varnothing$. Let $n \in \mathbb{Z}^+$, and let $\mathcal{F} \subseteq 2^{[n]}$ be a pairwise intersecting set of subsets of $[n]$. Then we have $\#\mathcal{F} \leq 2^{n-1}$.*

PROOF. We will show the contrapositive: let $\mathcal{F}$ be a family of subsets of $2^{[n]}$ with $\#\mathcal{F} > 2^{n-1}$; we will show that $\mathcal{F}$ is not pairwise intersecting.

To see this we consider the partition $\mathcal{P}$ of $2^{[n]}$ into 2^{n-1} parts, where each part is of the form $\{S, S^c\}$ for a subset $S \subseteq [n]$ and its complement $S^c = [n] \setminus S$. We can view the elements of $\mathcal{F}$ as pigeons and map $S \in \mathcal{F}$ to the unique element of $\mathcal{P}$ that contains it as an element: namely $\{S, S^c\}$. Because we have more than 2^{n-1} pigeons and 2^{n-1} pigeonholes, there must be a subset $S \subseteq [n]$ such that S, S^c are both elements of $\mathcal{F}$, but then $S \cap S^c = \varnothing$, so $\mathcal{F}$ is not pairwise intersecting. □

In Exercise 3.14 you are asked to show that Proposition 3.17 is *sharp*: for all $n \in \mathbb{Z}^+$ there are pairwise intersecting families of subsets of $[n]$ with exactly 2^{n-1} elements.

7. Exercises

EXERCISE 3.1. *Use the Venn diagram for $A_1 \cup A_2 \cup A_3$ to prove Proposition 7.3.*

EXERCISE 3.2. *In the proof of Proposition 3.6b), we did not explicitly use the assumption that $n \geq k$. Did we in fact use it somehow? If not, and the argument is correct, this means that the conclusion still holds when $n < k$: is that true?*

EXERCISE 3.3. *Earlier we gave a set-theoretic counting argument to show (11). Show (11) algebraically: i.e., prove that*

$$\frac{n!}{k!(n-k)!} = \frac{(n-1)!}{(k-1)!(n-k)!} + \frac{(n-1)!}{k!(n-1-k)!}.$$

EXERCISE 3.4. *The goal of this exercise to establish*

$$\text{For all } k, n \in \mathbb{N},\ \text{we have } \binom{n}{k} = \binom{n}{n-k}. \tag{15}$$

a) *Show (15) by direct algebraic manipulation.*
b) *Show (using sets) that the number of k-element subsets of $[n]$ is equal to the number of $(n-k)$-element subsets of $[n]$.*

EXERCISE 3.5. *Let $k, n \in \mathbb{Z}$ with $0 \leq k \leq n$.*

a) *Show: we have $\binom{n}{k} \leq \binom{n}{k+1} \iff k \leq \frac{n-1}{2}$.*
b) *Suppose $n \geq 2$ is even. Show:*

$$\binom{n}{0} < \binom{n}{1} < \ldots < \binom{n}{n/2} > \ldots > \binom{n}{n-1} > \binom{n}{n}.$$

c) *Suppose $n \geq 3$ is odd. Show:*

$$\binom{n}{0} < \binom{n}{1} < \ldots \leq \binom{n}{\frac{n-1}{2}} = \binom{n}{\frac{n+1}{2}} > \ldots > \binom{n}{n-1} > \binom{n}{n}.$$

EXERCISE 3.6. *Write down all elements of $\mathcal{D}_4$ – i.e., the permutations on four elements without fixed points – and thereby show that $\#\mathcal{D}_4 = 9$.*

EXERCISE 3.7. *Consider the following game: you and I will each take a standard deck of 52 playing cards. We will exchange decks and shuffle until we are convinced they are random. Then, one by one, we will each turn over our cards at the same time. If we ever turn over the same card (both rank and suit) at the same time, I win the game and you pay me 1 dollar. If we always turn over different cards, you win the game and you pay me 1 dollar.*

a) *Do you want to play this game with me? (Hint: no, you don't. Why?)*
b) *What is the correct, fair price you should pay when I win the game?*

EXERCISE 3.8. *Show: if $x \in \mathbb{Z}$, $y \in \mathbb{R}$ and $x \geq y$, then $x \geq \lceil y \rceil$.*

EXERCISE 3.9. *Finite Lists and the Strong Pigeonhole Principle:*

a) *Give a statement of the Strong Pigeonhole Principle (Theorem 3.15) in terms of limiting the number of repetitions in a finite list of length P with elements drawn from a finite set of size H.*
b) *Give a direct proof of your statement from part a). Do you like this any better than the given proof of Theorem 3.15?*

EXERCISE 3.10. *Let X and I be nonempty sets. A **I-pseudopartition** of X is an I-indexed family $\widetilde{\mathcal{P}} = \{Y_i\}_{i \in I}$ of X such that:*

(i) *We have $\bigcup_{i \in I} Y_i = X$, and*
(ii) *For all $i \neq j \in I$, we have $Y_i \cap Y_j \neq \varnothing$.*

A pseudopartition of X is an I-pseudopartition for some nonempty index set I.

a) *Let $\tilde{P}$ be an I-pseudopartition of X such that for all $i \in I$ we have $Y_i \neq \varnothing$. Show that $\mathcal{P} \coloneqq \{Y_i \mid i \in I\}$ is a partition of X and $\#\mathcal{P} = \#I$.*
(In other words, every pseudopartition consisting of nonempty subsets of X induces a partition.)
b) *Formulate a version of Theorem 3.15 using pseudopartitions, and show that this version is equivalent to the Strong Pigeonhole Principle (in the sense that each can easily be deduced from the other).*

EXERCISE 3.11. *Show that under the hypotheses of Proposition 3.16 there are always $x, y \in S$ such that $x \mid y$. (The definition of $x \mid y$ is given in §4.3.)*

EXERCISE 3.12. *Show: at a party with at least* 2 *(and finitely many!) people, there are at least two different people at the party that have the same number of friends at the party. (Assume that if* X *is friends with* Y *then* Y *is friends with* X *and that no one is friends with themself.)*

EXERCISE 3.13. *Show: given any five points on a sphere, there is a closed hemisphere containing at least four of them.*

EXERCISE 3.14.
a) *Let* $n \in \mathbb{Z}^+$. *Find a pairwise intersecting set* $\mathcal{F}$ *of subsets of* $[n]$ *of size* 2^{n-1}.
(Hint: you can even find such a family such that $\bigcap_{S \in \mathcal{F}} S \neq \varnothing$.*)*
b) *In fact, find* n *different such families.*
(If you did part a), this should be easy.)
c) *Are there any other pairwise intersecting families of subsets of* $[n]$ *of size* 2^{n-1} *other than the ones you found in part b)?*

EXERCISE 3.15. *Someone's kitchen is a* 5 *by* 5 *square of lineoleum tiles. Late at night* 25 *roaches come out, and each chooses a different square tile as their hangout. The kitchen's owner arrives, turns on the light, shrieks and reaches for a can of roach spray. The roaches skitter – each moves from their tile to an orthogonally adjacent tile (i.e., a tile which shares a side with their present tile). Show that, no matter how they do this, after the skittering, two roaches will occupy the same tile.*

CHAPTER 4

Numbers, Inequalities and Rings

1. Field Axioms for $\mathbb{R}$

Two basic structures of the real numbers are the addition and multiplication operations. These are given as functions

$$+ : \mathbb{R} \times \mathbb{R} \to \mathbb{R}$$

and

$$\cdot : \mathbb{R} \times \mathbb{R} \to \mathbb{R}.$$

This means that given any two real numbers x, y we have a well-defined real number $x + y$ and a well-defined real number $x \cdot y$. (Functions are studied in detail in Chapter F.) We can, and will, define subtraction and division in terms of addition and multiplication, using certain familiar properties these operations satisfy called **field axioms**. They go as follows:

(A1) (Commutativity of Addition) $\forall\ x, y \in \mathbb{R},\ x + y = y + x$.
(A2) (Associativity of Addition) $\forall\ x, y, z \in \mathbb{R},\ (x + y) + z = x + (y + z)$.
(A3) (Identity for Addition) $\exists\ 0 \in \mathbb{R},\ 0 + x = x = x + 0$.
(A4) (Inverses for Addition) $\forall\ x \in \mathbb{R}, \exists\ y \in \mathbb{R}\ x + y = 0 = y + x$.
(M1) (Commutativity of Multiplication) $\forall x, y \in \mathbb{R},\ x \cdot y = y \cdot x$.
(M2) (Associativity of Multiplication) $\forall x, y, z \in \mathbb{R},\ (x \cdot y) \cdot z = x \cdot (y \cdot z)$.
(M3) (Identity for Multiplication) $\exists 1 \in \mathbb{R},\ \forall x \in \mathbb{R},\ x \cdot 1 = x = 1 \cdot x$.
(M4) (Inverses for Multiplication) $\forall x \in \mathbb{R} \setminus \{0\},\ \exists y \in \mathbb{R}\ x \cdot y = 1 = y \cdot x$.
(D) (Distributivity) $\forall x, y, z \in \mathbb{R},\ x \cdot (y + z) = (x \cdot y) + (x \cdot z)$.
(ND) (Nondegeneracy) We have $1 \neq 0$.

The above formulation is slightly sloppy in that the additive and multiplicative identities are asserted to exist in (A3) and (M3) and then (A4), (M4) and (ND) are phrased as though we are *given* an additive identity 0 and a multiplicative identity 1. The following result shows that there is actually no ambiguity, since the axioms imply that there is a *unique* additive identity, which we may therefore denote by 0, and a *unique* multiplicative identity, which we may therefore denote by 1.

PROPOSITION 4.1. *Let $(F, +, \cdot)$ be a number system satisfying the field axioms. Then:*

a) *For all $x \in F$, we have $0 \cdot x = 0$.*
b) *There is a unique identity element for addition, which we may therefore denote by 0.*
c) *There is a unique identity element for multiplication, which we may therefore denote by 1.*

d) *For all $x \in F$, the additive inverse of x is unique. If -1 is the additive inverse of 1, then the additive inverse of x is $(-1) \cdot x$.*
e) *For all $x \in F$, the multiplicative inverse of F is unique; we may therefore denote it by x^{-1}.*

PROOF. a) We have

$$0 \cdot x = (0 + 0) \cdot x = (0 \cdot x) + (0 \cdot x).$$

Let y be the additive inverse of $0 \cdot x$. Adding y to both sides, we get

$$0 = 0 \cdot x.$$

b) Let a and a' be two elements of F satisfying (A3). Then we have

$$a = a + a' = a'.$$

c) Let m and m' be two elements of F satisfying (M3). Then we have

$$m = m \cdot m' = m'.$$

d) Let $x \in F$, and let $y, z \in F$ be two additive inverses of x, so we have

$$y + x = 0 = x + z.$$

Adding y to both sides, we get

$$y = y + 0 = y + (x + z) = (y + x) + z = 0 + z = z.$$

Moreover we have

$$(-1) \cdot x + x = (-1) \cdot x + 1 \cdot x = (-1 + 1) \cdot x = 0 \cdot x = 0.$$

e) Let $x \in F \setminus \{0\}$, and let $y, z \in F$ be two multiplicative inverses of x, so we have

$$y \cdot x = 1 = x \cdot z.$$

Multiplying both sides by y, we get

$$y = y \cdot 1 = y \cdot (x \cdot z) = (y \cdot x) \cdot z = 1 \cdot z = z. \qquad \square$$

Notice that Proposition 4.1 implies that 0 does not have a multiplicative inverse. Thus the axioms for addition and multiplication are very similar but not identical.

Our take on subtraction is that it is an operation defined in terms of addition: by $x - y$ we mean $x + (-y)$: i.e., x plus the additive inverse of y. More fundamentally, in any number system F satisfying the field axioms, we have

$$x - y = z$$

if and only if

$$x = y + z.$$

Thus $x - y$ is the unique element of F such that when we add y to $x - y$ we get x.

A similar discussion holds for division. For $x, y \in F$ with $y \neq 0$, by $\frac{x}{y}$ we mean xy^{-1}, where y^{-1} is the multiplicative inverse of y. Again the equation

$$\frac{x}{y} = z$$

holds if and only if

$$x = y \cdot z.$$

Thus $\frac{x}{y}$ is the unique element of F such that when we multiply $\frac{x}{y}$ by y we get x.

PROPOSITION 4.2. *Let $(F, +, \cdot)$ be a number system satisfying the field axioms. For $x, y \in F$, the following are equivalent:*

(i) *We have $x = 0$ or $y = 0$.*
(ii) *We have $xy = 0$.*

PROOF. (i) $\implies$ (ii): If $x = 0$, then by Proposition 4.1 we have $xy = 0 \cdot y = 0$. Similarly, if $y = 0$, then $x \cdot 0 = 0 \cdot x = 0$.
(ii) $\implies$ (i): We argue by contrapositive: suppose that $x \neq 0$ and $y \neq 0$, so by (M4) there are $x^{-1}, y^{-1} \in F$ such that

$$xx^{-1} = yy^{-1} = 1.$$

Then

$$(y^{-1}x^{-1})(xy) = y^{-1}(x^{-1}x)y = y^{-1} \cdot 1 \cdot y = 1,$$

so xy has a multiplicative inverse. Proposition 4.1 shows that 0 has no multiplicative inverse, so $xy \neq 0$. □

In Exercise 7.21 you are asked to show that for any n elements $x_1, \ldots, x_n$ of a number system satisfying the field axioms, we have $x_1 \cdots = x_n = 0$ if and only if $x_i = 0$ for at least one $1 \leq i \leq n$.

2. Ordered Field Axioms

In addition to the operations of $+$ and $\cdot$, the real numbers are also endowed with an order relation $\leq$. This means that for $x, y \in \mathbb{R}$ it is either true that $x \leq y$ or it is false that $x \leq y$. This notation comes with some standard variations: we write

- $x < y$ if $x \leq y$ and $x \neq y$,
- $x \geq y$ if $y \leq x$,
- $x > y$ if $y < x$.

This relation satisfies the following properties:
(O1) (Reflexivity) For all $x \in \mathbb{R}$, $x \leq x$.
(O2) (Anti-Symmetry) For all $x, y \in \mathbb{R}$, if $x \leq y$ and $y \leq x$, then $x = y$.
(O3) (Transitivity) For all $x, y, z \in \mathbb{R}$, if $x \leq y$ and $y \leq z$, then $x \leq z$.

Finally, we give three more properties describing the interactions among $\leq$, $+$ and $\cdot$:

(OF1) (Trichotomy) For all $x \in \mathbb{R}$, exactly one of the following holds:

$$x = 0, \quad x > 0, \quad -x > 0.$$

(OF2) For all $x, y, z \in \mathbb{R}$, if $x \leq y$, then $x + z \leq y + z$.
(OF3) For all $x, y \in \mathbb{R}$, if $x \geq 0$ and $y \geq 0$, then $xy \geq 0$.

Given a set F endowed with binary operations $+$ and $\cdot$ and a binary relation $\leq$ satisfying the ten field axioms and the six order axioms as above, we call it an **ordered field**. Thus first of all we are saying that the real numbers $\mathbb{R}$ are an ordered field. Also the rational numbers $\mathbb{Q}$ form an ordered field with the same operations restricted to $\mathbb{Q}$.

PROPOSITION 4.3. *Let F be an ordered field, and let $x, y, z \in F$.*

a) *If $x < y$, then $x + z < y + z$.*

b) *If $x > 0$ and $y > 0$, then $xy > 0$.*
c) *Exactly one of the following holds: (i) $x < y$; (ii) $x = y$; (iii) $y < x$.*

PROOF. a) If $x < y$, then $x \leq y$, so by (OF2) we have $x + z \leq y + z$. Moreover, if $x + z = y + z$ then adding the additive inverse of z to both sides shows $x = y$, contradicting $x < y$. So we must have $x + z < y + z$.
b) If $x > 0$ and $y > 0$, then $x \geq 0$ and $y \geq 0$, so by (OF3) we have $xy \geq 0$. If $xy = 0$, then by Proposition 4.2 we have $x = 0$ or $y = 0$; either one is a contradiction.
c) First we show that if any of (i), (ii), (iii) hold then the other two cannot hold: If $x < y$ then certainly we do not have $x = y$; if $y < x$ then in particular we have $x \leq y$ and $y \leq x$, so then (O2) implies $x = y$, a contradiction. If $x = y$, then certainly we do not have $x < y$ or $y < x$. The case of $y < x$ is obtained from the case of $x < y$ just by interchanging x and y.

Next we apply (OF1) to $y - x$: if $y - x = 0$, then $x = y$. If $y - x > 0$, then equivalently $0 < x - y$; adding y to both sides and using part a) we get $y < x$. If $x - y = -(y - x) > 0$, then equivalently $0 < y - x$; arguging as above with x and y interchanged shows $x < y$. □

PROPOSITION 4.4. *Let F be an ordered field.*

a) *We have $1 > 0$ and $-1 < 0$.*
b) *For all $x \in F \setminus \{0\}$, we have $x^2 > 0$.*
c) *For all $x \in F$, we have $x^2 \geq 0$.*

PROOF. a) By Nondegeneracy (ND), we have $1 \neq 0$. So by Trichotomy (OF1), we have either $-1 > -0$ or $1 > 0$ and not both. If $-1 > 0$, then by (OF3) we have that $1 = (-1) \cdot (-1) \geq 0$, and since $1 \neq 0$ we get $1 > 0$, a contradiction. So it must be that $1 > 0$ and $-1 < 0$.
b) Since $x \neq 0$, by Trichotomy we have either $x > 0$ or $-x > 0$. In the former case, using (OF3) we get $x^2 \geq 0$. By Proposition 4.2 we know that $x^2 \neq 0$, so indeed we have $x^2 > 0$. The other case is similar: if $-x > 0$, then $x^2 = (-x)(-x) \geq 0$, and since $x^2 \neq 0$, in fact $x^2 > 0$.
c) Once we make the observation that $0^2 = 0$, this follows from part b). □

Proposition 4.4 has an interesting consequence. The complex numbers satisfy the field axioms, but they do not come endowed with an ordering $\leq$. One might try to *find* an order relation $\leq$ on $\mathbb{C}$. But don't try too hard: since $\mathbb{C}$ contains an element i with square -1, it follows from parts a) and c) of Proposition 4.4a) that there is no way to endow $\mathbb{C}$ with an ordering $\leq$ that satisfies the six order axioms.

Proposition 4.4 also gives a useful technique for proving inequalities: for $x, y \in \mathbb{R}$ (or any ordered field), we have

$$x \leq y \iff y - x \geq 0$$

(Exercise 4.1). Though there is no real content to this observation, it often does simplify things: now instead of dealing with two numbers x and y, we are dealing with the one number $y - x$. But Proposition 4.4 allows us to take things further: in order to show that $y - x \geq 0$ it is sufficient to show that it is the square of something else. This holds in any ordered field. In some ordered fields, like $\mathbb{Q}$, this is not necessary: e.g. $2 = 1 + 1 > 0$ in $\mathbb{Q}$, but – as we will see later on! – 2 is not the square of any *rational* number. However, as a consequence of the Intermediate Value Theorem, *every* non-negative real number is the square of another real number, so

when working in the real numbers (as we almost always will be), the technique of showing that a number is non-negative by finding a square root will in principle always work. Here is a simple example.

PROPOSITION 4.5. *For all $x, y \in \mathbb{R}$ we have*

$$x^2 - 2xy + y^2 \geq 0.$$

PROOF. For all $x, y \in \mathbb{R}$ we have

$$x^2 - 2xy + y^2 = (x - y)^2 \geq 0$$

by Proposition 4.4. (In fact this argument works in any ordered field.) □

PROPOSITION 4.6. *Let x, y, z be elements of an ordered field F.*

a) *We have $x < 0 \iff -x > 0$.*
b) *If $x > 0$ then $\frac{1}{x} > 0$. If $x < 0$, then $\frac{1}{x} < 0$.*
c) *If $x > 0$ and $y < 0$, then $xy < 0$.*
d) *If $x < 0$ and $y < 0$, then $xy > 0$.*

PROOF. a) Suppose $x < 0$. By Proposition 4.3c) we do not have $x > 0$, so by (OF1) we must have $-x > 0$. Conversely, if $-x > 0$ then $x \neq 0$. If $x > 0$, then adding $-x$ to both sides and using Proposition 4.3a) we get $0 < -x$, a contradiction. So we must have $x < 0$.
b) Case 1: Suppose $x > 0$. Then $\frac{1}{x} \neq 0$ (because $xx^{-1} = 1$), so by Trichotomy (OF1) we must have $\frac{1}{x} > 0$ or $-\frac{1}{x} > 0$. We will rule out the latter: if $-\frac{1}{x} > 0$ then

$$-1 = x(-\frac{1}{x}) > 0,$$

but then Trichotomy implies we *cannot* have $1 > 0$, contradicting Proposition 4.4a). Case 2: Suppose $x < 0$, so, by part a) we have $-x > 0$, and by Case 1 we have $-\frac{1}{x} = \frac{1}{-x} > 0$ and then $\frac{1}{x} < 0$ by part a) again.
c) If $x > 0$ and $y < 0$, then $-y > 0$, so by Proposition 4.3b) we have

$$-(xy) = x(-y) > 0,$$

which means that $xy < 0$.
d) If $x < 0$ and $y < 0$, then $-x > 0$ and $-y > 0$, so by Proposition 4.3b) we have

$$xy = (-x)(-y) > 0.$$ □

We define the **sign** of an element x of an ordered field F as follows:

$$\operatorname{sgn}(x) = \begin{cases} 1 & x > 0 \\ 0 & x = 0 \\ -1 & x < 0 \end{cases}.$$

Some simple properties of the sign function are established in Exercise 4.2.

PROPOSITION 4.7. *Let F be an ordered field, let $a, b \in F \setminus \{0\}$, and suppose $a < b$. Then:*

a) *If $\operatorname{sgn}(a) \neq \operatorname{sgn}(b)$, then $\frac{1}{a} < \frac{1}{b}$.*
b) *If $\operatorname{sgn}(a) = \operatorname{sgn}(b)$, then $\frac{1}{a} > \frac{1}{b}$.*

PROOF. a) For all $x \in F \setminus \{0\}$, we have $\operatorname{sgn}(x^{-1}) = \operatorname{sgn}(x)$. So if $a < 0 < b$ then $\frac{1}{a} < 0 < \frac{1}{b}$, while similarly if $b < 0 < a$ then $\frac{1}{b} < 0 < \frac{1}{a}$.
b) We have $\frac{1}{a} - \frac{1}{b} = \frac{b-a}{ab} \neq 0$ (snce $a \neq b$). Using properties of the sign function established in Exercise 4.2 we have

$$\operatorname{sgn}(\frac{1}{a} - \frac{1}{b}) = \operatorname{sgn}(b-a)\operatorname{sgn}(a^{-1})\operatorname{sgn}(b^{-1})$$

$$= \operatorname{sgn}(b-a)\operatorname{sgn}(a)\operatorname{sgn}(b) = \operatorname{sgn}(b-a)\operatorname{sgn}(a)^2 = \operatorname{sgn}(b-a) = 1,$$

so $\frac{1}{a} > \frac{1}{b}$. □

3. Well-Ordering

Let $S \subseteq \mathbb{R}$. We say that $x \in S$ is a **minimum element of S** if it is less than every other element of S: that is,

$$\forall y \in S,\ x \leq y.$$

We say that $x \in S$ is a **maximum element of S** if it is greater than every other element of S: that is,

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

A subset may have no minimum or one minimum, but it cannot have more than one minimum: if x and y are each less than every other element of S, then $x \leq y$ and $y \leq x$, so $x = y$. In other words, the minimum may not *exist* but if it does it must be *unique*. Exactly the same goes for the maximum.

EXAMPLE 4.8. *Some subsets of* $\mathbb{R}$*:*

a) *The empty set* $\varnothing$ *has neither maximum nor minimum elements. Indeed it has no elements at all!*
b) *The set* $\mathbb{Z}$ *has no maximum element: for any integer* n*, we have a larger integer* $n+1$ *and a smaller integer* $n-1$*.*
c) *The set* $\mathbb{N}$ *of non-negative integers has a minimum element,* 0*, but no maximum. The set* $\mathbb{Z}^+$ *of positive integers has a minimum element,* 1*, but no maximum.*
d) *If* $S \subseteq \mathbb{R}$ *is finite and nonempty, then it has a maximum and minimum element. To see that a minimum exists: let* $N := \#S$*, and let* $x_1 \in S$ *be any element. If* x_1 *is not a minimum, there is* $x_2 < x_1$*. If* x_2 *is not a minimum, there is* $x_3 < x_2$*. By the transitivity of strict inequality we have* $x_3 < x_2 < x_1$*, so these are three different elements. We can continue this procedure as many times as we like.*[1] *Once we perform it* N *times, we get*

$$x_1 > x_2 > \ldots > \ldots > x_N > x_{N+1}.$$

Thus we have an irredundant list of length $N+1$ *drawn from elements of* S*, which contradicts* $\#S = N$*. The argument for the maximum is very similar and left to the reader as Exercise 5.1.*
e) *For any* $a \leq b$*, the closed bounded interval*

$$\{[a,b] := \{x \in \mathbb{R} \mid a \leq x \leq b\}$$

has a *as a minimum element and* b *as a maximum element.*

[1] But please read Remark 4.10.

In Exercise 5.2 you are asked to show that no infinite subset of $\mathbb{Z}$ has both a minimum and a maximum element.

A subset S of $\mathbb{R}$ is **well-ordered** if every nonempty subset $T \subseteq S$ has a minimum element.

THEOREM 4.9 (Well-Ordering Principle). *The natural numbers $\mathbb{N}$ are well-ordered.*

Here is an argument for the Well-Ordering Principle: Let $S \subseteq \mathbb{N}$ be a nonempty subset. Then S has an element $x_1 \geq 0$. Consider the set

$$T := S \cap [0, x_1] = \{x \in S \mid x \leq x_1\}.$$

The set T is finite: indeed, each of its elements is an integer x with $0 \leq x \leq x_1$, so $\#T \leq x_1 + 1$. By Example 4.8, the set T has a minimum m. Since $x_1 \in T$, we have in particular that $m \leq x_1$.

We claim that this element m is also a minimum for S: indeed, let $x \in S$. If $x \leq x_1$ then $x \in T$, so $m \leq x$. If $x \notin S$ then $m \leq x_1 < x$.

REMARK 4.10. *The careful reader will notice that above we spoke of an "argument," not a "proof." Later in these notes we will discuss Mathematical Induction, the most important of all proof techniques, and we will justify it using the Well-Ordering Principle. One can also justify the Well-Ordering Principle in terms of Mathematical Induction. Believe it or not, we already have: in Example 4.8d) when we said "we can continue this procedure as many times as we like," we are actually appealing to Mathematical Induction.*

The truth is that Well-Ordering and Mathematical Induction are equivalent foundational *properties of the integers: this means that they are so basic that they can only be justified by a formal construction of $\mathbb{Z}$, which would in turn rely on certain set-theoretic axioms. We will not be so formal in this course (and in fact most mathematicians do not want to be this formal, but rely on certain other mathematicians who have thought these things through).*

In Exercise 4.6, you are asked to show that a nonempty subset $S \subseteq \mathbb{Z}$ is well-ordered if and only if it has a minimum. In particular, for all $N \in \mathbb{Z}$ the subset

$$\mathbb{Z}^{\geq N} := \{n \in \mathbb{N} \mid n \geq N\}$$

is well-ordered.

COROLLARY 4.11 (Principle of Infinite Descent). *There is no infinite strictly decreasing sequence*

$$x_1 > x_2 > \ldots > x_n > \ldots \tag{16}$$

of positive integers.

PROOF. From the strictly decreasing sequence (16) we form the associated set

$$S := \{x_n \mid n \in \mathbb{Z}^+\} \subseteq \mathbb{Z}^+.$$

This is a nonempty subset of $\mathbb{Z}^+$ but it cannot have a minimum: every element is of the form x_n for some $n \in \mathbb{Z}^+$, and then $x_{n+1} < x_n$. □

A small variation on this is also useful. For $a, b \in \mathbb{Z}$ we say that **a divides b** if there is $c \in \mathbb{Z}$ such that $ac = b$. We say that **a properly divides b** if a divides b but b does *not* divide a. We will study divisibility more carefully in §5.2; in particular we

will prove in Proposition 5.5 that for nonzero integers a and b, if a properly divides b then $|a| < |b|$.

COROLLARY 4.12 (Principle of Infinite Divisibility). *There is no infinite sequence*

$$x_1, x_2, x_3, \ldots, x_n, \ldots$$

such that each x_n is an integer and x_{n+1} properly divides x_n for all $n \in \mathbb{Z}^+$.

PROOF. Seeking a contradiction, suppose there is such an infinite sequence of integers. If for some $n \in \mathbb{Z}^+$ we have $x_{n+1} = 0$, then 0 properly divides x_n, contradicting Proposition 5.5b). Therefore all terms of the sequence except possibly the first are nonzero integers. Then Proposition 5.5d) implies that

$$|x_2| > |x_3| > \ldots > |x_n| > \ldots,$$

so we have an infinite strictly descending sequence of positive integers, contradicting Corollary 4.11. □

THEOREM 4.13 (Infinite Subsets of $\mathbb{N}$). *Let $S \subseteq \mathbb{N}$ be an infinite subset. Then:*

a) *For all $n \in \mathbb{Z}^+$ there is a non-negative integer a_n such that:*
   (i) *We have $a_1 < a_2 < \ldots < a_n < \ldots$, and*
   (ii) *$S = \{a_n \mid n \in \mathbb{Z}^+\}$.*

b) *The infinite list $a_1, a_2, \ldots$ of part a) is unique: that is, if $b_1, b_2, \ldots$ is another infinite list of non-negative integers satisfying the conditions (i) and (ii) of part a), then $a_n = b_n$ for all $n \in \mathbb{Z}^+$.*

PROOF. a) The basic idea is that since S is infinite, it is nonempty and remains nonempty (indeed infinite) after removing any finite number of elements. First, S is a nonempty subset of $\mathbb{N}$, so by Well Ordering it has a least element a_1. Then $S_1 := S \setminus \{a_1\}$ is still an infinite subset of $\mathbb{N}$, so it has a least element a_2, which is then the second smallest of all the elements of S. We continue in this way: for $n \in \mathbb{Z}^+$, after having defined a_n and observed that it is the nth smallest element of S, we let a_{n+1} be the least element of $S \setminus \{a_1, \ldots, a_n\}$ and observe that a_{n+1} is the $(n+1)$st smallest element of S. Because S is infinite, this argument works for all $n \in \mathbb{Z}^+$, and it is clear that the resulting infinite list $a_1, a_2, \ldots, a_n, \ldots$ satisfies property (i) of part a). Now let $N \in S$. Since every a_n is an integer and

$$0 \leq a_1 < a_2 < \ldots < a_N < a_{N+1}$$

we must have $a_1 \geq 0$, $a_2 \geq 1$, and so forth, hence finally $a_{N+1} \geq N$. This shows that N is no larger than the $(N+1)$st smallest element of S, so we must have $N = a_n$ for some $0 \leq n \leq N+1$. In particular every element of S is of the form a_n for some $n \in \mathbb{Z}^+$, showing (ii).

b) Above we chose a_n to be the nth smallest element of S and then checked that this infinite list satisfies properties (i) and (ii), but going the other way around is even easier: if $S = \{b_n \mid n \in \mathbb{Z}^+$ and $b_1 < b_2 < b_3 < \ldots$ then clearly b_1 is the least element of S, b_2 is the second smallest element of S, and so forth: in general b_n is the nth smallest element of S, so $b_n = a_n$. □

4. The Rational Numbers

Let x be a rational number. Thus we may write $x = \frac{a}{b}$ with $a \in \mathbb{Z},\ b \in \mathbb{Z} \setminus \{0\}$. We say that the expression $\frac{a}{b}$ is **in lowest terms** if a and b are not both multiples

of some integer $d > 1$. Thus for instance $\frac{2}{3}$ is in lowest terms and $\frac{10}{15}$ is not, since 10 and 15 are both divisible by 5.

(Strictly speaking we should not say that the *fraction* $\frac{a}{b}$ is in lowest terms but rather that the ordered pair (a, b) is in lowest terms, because after all $\frac{2}{3} = \frac{10}{15}$. But having been this pedantic once, we will not say it again.)

PROPOSITION 4.14. *Every rational number can be written in lowest terms.*

PROOF. We can write 0 in lowest terms as $\frac{0}{1}$. Moreover, we have that $\frac{a}{b}$ is in lowest terms if and only if $\frac{-a}{b}$ is in lowest terms, so it suffices to show that every positive rational number $\frac{a}{b}$ is in lowest terms.

So suppose that $a, b \in \mathbb{Z}^+$ are such that $\frac{a}{b}$ is *not* in the lowest terms: this means there is an integer $d_1 > 1$ such that

$$a = d_1 a_1,\ b = d_1 b_1$$

for $a_1, b_1 \in \mathbb{Z}^+$. Then we have

$$\frac{a}{b} = \frac{d_1 a_1}{d_1 b_1} = \frac{a_1}{b_1}.$$

If $\frac{a_1}{b_1}$ is in lowest terms, we're done. If not, there is an integer $d_2 > 1$ such that

$$a_1 = d_2 a_2,\ b_1 = d_2 b_2$$

for $a_2, b_2 \in \mathbb{Z}^+$. Then we have

$$\frac{a}{b} = \frac{d_1 a_1}{d_1 b_1} = \frac{a_1}{b_1} = \frac{d_2 a_2}{d_2 b_2} = \frac{a_2}{b_2}.$$

And so forth. If this procedure terminates at some stage, then we have written $\frac{a}{b}$ in lowest terms. If not, then we get an infinite sequence of positive integers

$$a_1 = d_1 a_2,\ a_2 = d_2 a_3, \ldots, a_n = d_n a_{n+1}$$

with $d_n > 1$ for all $n \in \mathbb{Z}^+$. Thus $a_n > a_{n+1}$ for all n, so $a_1, a_2, a_3, \ldots, a_n, \ldots$ is an infinite decreasing sequence of positive integers, contradicting the Principle of Infinite Descent (Corollary 4.11). □

5. Exercises

EXERCISE 4.1. *Let F be an ordered field, and let x, y be elements of F. Show: $x \leq y \iff y - x \geq 0$.*

EXERCISE 4.2. *Let F be an ordered field, and let x, y be elements of F.*

a) *We define* $|x| \coloneqq \begin{cases} x & x \geq 0 \\ -x & x < 0 \end{cases}$. *Show:*

$$x = \operatorname{sgn}(x)|x|.$$

b) *Show:* $\operatorname{sgn}(xy) = \operatorname{sgn}(x)\operatorname{sgn}(y)$.

EXERCISE 4.3. *Let F be a field. Show:* $(-1) \cdot (-1) = 1$.

EXERCISE 4.4. *Let $x_1 \leq \ldots \leq x_n$ be elements of an ordered field F.*

a) *Show: if $x_1 = x_n$, then $x_1 = \ldots = x_n$.*

b) *Show: if $x_i < x_{i+1}$ for some $1 \leq i \leq n-1$ then $x_1 < x_n$.*

EXERCISE 4.5. *Let $S \subseteq \mathbb{R}$ be well-ordered. Show: every subset $T \subseteq S$ is well-ordered.*

EXERCISE 4.6. *In this exercise we will determine the well-ordered subsets of $\mathbb{Z}$.*

a) *Show: the empty set $\varnothing$ is well-ordered.*

b) *Let $N \in \mathbb{Z}$. Show that the set*

$$\mathbb{Z}^{\geq N} := \{n \in \mathbb{Z} \mid n \geq N\}$$

is well-ordered.
(Suggestion: for $S \subseteq \mathbb{Z}^{\geq N}$, consider $S' := \{s - N \mid s \in S\}$. Show that $S' \subseteq \mathbb{N}$ and that S has a minimum if and only if S' has a minimum.)

c) *For a nonempty subset $S \subseteq \mathbb{Z}$, show that the following are equivalent:*
   (i) *The set S is well-ordered.*
   (ii) *The set S has a minimum.*
   (Suggestion: for the harder direction (ii) $\implies$ (i), use part b) and Exercise 4.5.)

EXERCISE 4.7. *Let X and Y be subsets of $\mathbb{R}$ such that the symmetric difference $X \Delta Y$ is finite. Show: X is well-ordered if and only if Y is well-ordered.*

EXERCISE 4.8. *For a subset $X \subseteq \mathbb{R}$, show that the following are equivalent:*

(i) *X is well-ordered.*
(ii) *There is no strictly decreasing infinite sequence in X: that is, there is no sequence*

$$x_1 > x_2 > \ldots > x_n > \ldots$$

with $x_n \in X$ for all $n \in \mathbb{Z}^+$.

(Hint: for (i) $\implies$ (ii): adapt the proof Corollary 4.11. Hint for (ii) $\implies$ (i): prove the contrapositive by showing that a nonempty subset without a minimum gives rise to a strictly decreasing sequence in X.)

EXERCISE 4.9. *We say a subset $X \subseteq \mathbb{R}$ has the **Finite Predecessors Property (FPP)** if for every $x \in X$, the set*

$$P_X(x) := \{y \in X \mid y < x\}$$

of elements of X that are strictly less than x is finite.

a) *Suppose $X \subseteq \mathbb{R}$ satisfies (FPP). Show that X is well-ordered.*
(Suggestion: use Exercise 4.8.)

b) *In our "explanation" that $\mathbb{N}$ is well-ordered, we used that the set $\mathbb{N}$ satisfies (FPP), which we regarded as a "foundational fact" about $\mathbb{N}$. Assuming this, part a) gives another proof that $\mathbb{N}$ is well-ordered that may be a bit cleaner. Do you agree?*

c) *Let $N \in \mathbb{Z}$, and put*

$$\mathbb{Z}^{\geq N} := \{n \in \mathbb{Z} \mid n \geq N\}$$

be the set of integers that are at least N. For any $n \in \mathbb{Z}^{\geq N}$, give an irredundant finite list of the elements of $P_{\mathbb{Z}^{\geq N}}(n)$ and thereby show that it has exactly $n - N$ elements. Deduce that $\mathbb{Z}^{\geq N}$ is well-ordered.

d) *Let*

$$X := \left\{ 1 - \frac{1}{n} \;\middle|\; n \in \mathbb{Z}^+ \right\} \cup \{1\}.$$

Show that X does not *satisfy (FPP) but is nevertheless well-ordered.*[2]

[2]In a certain precise sense, this set is the simplest well-ordered subset of $\mathbb{R}$ that is more complicated than $\mathbb{N}$ (whereas the sets $\mathbb{Z}^{\geq N}$ are equally complicated as $\mathbb{N}$). There is in fact an astoundingly rich hierarchy of well-ordered subsets of $\mathbb{R}$, which (un?)fortunately we will not meet in this course. Chapter 12 of the course text (which we will **not** cover) addresses this topic.

CHAPTER 5

Number Theory

1. The Division Theorem

In elementary school, when one divides one whole number by another, the answer is again a whole number together with a remainder, e.g.:

$$7 \div 3 = 2 \quad r\ 1.$$

As one goes on in mathematics, one learns fractions and then loses one's distaste for "improper fractions." If you ask a fully grown person what is $7 \div 3$ they are liable to answer: $\frac{7}{3}$.

Both answers are correct, and each has its merits. The idea of turning a "division problem" into a new kind of number – i.e., a rational number – is absolutely magnificent. Yet division with remainder has its uses in higher mathematics. The following result, which is a theoretical abstraction of the elementary school division process, is extremely useful in the deeper study of integers.

THEOREM 5.1 (Division Theorem). *Let $a \in \mathbb{Z}$, and let $b \in \mathbb{Z}^+$.*

a) *There are $q, r \in \mathbb{Z}$ such that $a = qb + r$ and $0 \leq r < b$.*
b) *The integers q and r are unique, subject to the properties they are asserted to have: that is, if $q_1, q_2, r_1, r_2 \in \mathbb{Z}$ are such that $0 \leq r_1, r_2 < b$ and $q_1 b + r_1 = a = q_2 b + r_2$, then $q_1 = q_2$ and $r_1 = r_2$.*
c) *If $a \geq 0$ then also $q \geq 0$.*

PROOF. a) Consider the set

$$S := \{a + nb \mid n \in \mathbb{Z}\} \cap \mathbb{N}.$$

That is, S is the set of all non-negative integers of the form $a + nb$ for some integer n. First we observe that $S \neq \varnothing$: indeed, if $a \geq 0$ then $a + 0 \cdot b = a \in S$, while if $a < 0$ then $a + (-a)b = a(1 - b) \geq 0$ since $a \leq 0$ and $1 - b \leq 0$, and thus in this case we have $a + (-a)b \in S$.

By the Well-Ordering Principle, the set S has a minimum: say $a + Nb$. Put

$$q := -N,\ r := a + Nb.$$

By our choice of N we have $r = a + Nb \geq 0$. If we had $r \geq b$ then

$$0 \leq r - b = (a + Nb) - b < a + Nb$$

and thus $r - b$ would be a smaller element of S than $a + Nb$, a contradiction. Moreover we have

$$qb + r = -Nb + (a + Nb) = a.$$

b) Let $q_1, q_2, r_1, r_2 \in \mathbb{Z}$ be such that $0 \leq r_1, r_2 < b$ and

$$q_1 b + r_1 = a = q_2 b + r_2.$$

Then

$$(q_1 - q_2)b = (r_2 - r_1),$$

so $b \mid r_2 - r_1$. Since $0 \leq r_1, r_2 < b$ we have

$$r_2 - r_1 \geq -r_1 > -b$$

and

$$r_2 - r_1 \leq r_2 < b.$$

The only multiple of b that lies strictly in between $-b$ and b is 0, so $r_2 - r_1 = 0$, i.e., $r_1 = r_2$. It now follows that $q_1 b = q_2 b$, and since $b \neq 0$, we conclude $q_1 = q_2$.
c) In the setting of part a), we suppose that $q < 0$. Then $q+1 \leq 0$; since also $b > 0$, we have

$$a = qb + r < qb + b = (q+1)b \leq 0.$$

This establishes part c) by contraposition. □

We call q the **quotient** obtained by dividing a by b, and we call r the **remainder** obtained by dividing a by b. (The uniqueness assertion of Theorem 5.1b) is what lets us speak of "the" quotient and "the" remainder.)

Here is a first application of the Division Theorem: let $n \in \mathbb{Z}$. Applying the Division Theorem with $a = n$ and $b = 2$ we get that exactly one of the following is true: either $n = 2q$ for some $q \in \mathbb{Z}$ or $2n = 2q + 1$ for some $q \in \mathbb{Z}$. In the former case we say that n is **even**; in the latter case we say that n is **odd**. We thus get a partition of $\mathbb{Z}$ into two parts, the even integers and the odd integers.

2. Divisibility

Let a and b be integers. We say that **a divides b** and write $a \mid b$ if there is an integer c such that $ac = b$.

• Suppose $a \neq 0$. We claim that $a \mid b$ if and only if $\frac{b}{a} \in \mathbb{Z}$. Indeed, if $\frac{b}{a} = c \in \mathbb{Z}$ then multiplying through by a gives $b = ac$ and thus $a \mid b$. Conversely, if $a \mid b$ then there is an integer c such that $ac = b$, and then dividing through by a gives $\frac{b}{a} = c \in \mathbb{Z}$.

• Suppose $a = 0$. We claim that $0 \mid b$ if and only if $b = 0$. Indeed, for any $c \in \mathbb{Z}$ we have $0 \cdot c = 0$, showing that $0 \mid 0$. Conversely, if $0 \mid b$ then there is $c \in \mathbb{Z}$ such that $b = 0 \cdot c = 0$, so $b = 0$.

Thus, while the definition of divisibility is "there is an integer c such that $ac = b$," if $a \mid b$ and $a \neq 0$ then the *unique* $c \in \mathbb{Z}$ such that $ac = b$ is $c = \frac{b}{a}$. (On the other hand every $c \in \mathbb{Z}$ satisfies $0 \cdot c = 0$.)

When $a \mid b$ we may also say that **b is divisible by a**. Once upon a time it was also common to say "b is evenly divisible by a." The following result explains why and gives another good way to think about divisibility when $a > 0$.

PROPOSITION 5.2. *Let $a \in \mathbb{Z}^+$ and $b \in \mathbb{Z}$. The following are equivalent:*

(i) *We have $a \mid b$.*
(ii) *When we write $b = qa + r$ with $q, r \in \mathbb{Z}$ and $0 \leq r < a$ as in Theorem 5.1, we have $r = 0$.*

PROOF. (i) $\implies$ (ii): Suppose $a \mid b$ and let $c \in \mathbb{Z}$ be such that $ac = b$. Then

$$qa + r = b = ac,$$

so

$$r = a(c - q),$$

so

$$|r| = |a||c - q|.$$

Since $c, q \in \mathbb{Z}$, if $c - q \neq 0$, then $|c - q| \geq 1$ and thus

$$a = |a| \leq |a||c - q| = |r| = r,$$

contradicting the fact that $r < a$. So we must have $c - q = 0$, i.e., $q = c$, and thus $r = b - qa = b - ca = b - b = 0$.
(ii) $\implies$ (i): If $b = qa + 0 = qa$ then certainly $a \mid b$. □

In fact Proposition 5.2 is more useful for understanding divisbility then it may first appear, because whether one integer divides another is "independent of sign":

PROPOSITION 5.3. *For integers $a, b \in \mathbb{Z}$, the following are equivalent:*

(i) *We have $a \mid b$.*
(ii) *We have $-a \mid b$.*
(iii) *We have $a \mid -b$.*
(iv) *We have $-a \mid -b$.*

PROOF. (i) $\implies$ (ii): If $a \mid b$, there is $c_1 \in \mathbb{Z}$ such that $ac_1 = b$. Then $-a(-c_1) = b$, showing that $-a \mid b$.
(ii) $\implies$ (iii): If $-a \mid b$, there is $c_2 \in \mathbb{Z}$ such that $-ac_2 = b$. Then $ac_2 = -b$, showing that $a \mid -b$.
(iii) $\implies$ (iv): If $a \mid -b$, then there is $c_3 \in \mathbb{Z}$ such that $ac_3 = -b$. then $(-a)(-c_3) = -b$, showing that $-a \mid -b$.
(iv) $\implies$ (i): If $-a \mid -b$, then there is $c_4 \in \mathbb{Z}$ such that $-ac_4 = -b$. Then $ac_4 = b$, showing that $a \mid b$. □

PROPOSITION 5.4. *Let $a, b, c \in \mathbb{Z}$.*

a) *If $a \mid b$ and $b \mid c$, then $a \mid c$.*
b) *If $a \mid b$ and $a \mid c$, then for all $d, e \in \mathbb{Z}$ we have $a \mid db + ec$.*
c) *The following are equivalent:*
   (i) *We have $a \mid b$ and $b \mid a$.*
   (ii) *We have $|a| = |b|$.*

PROOF. a) Since $a \mid b$, there is $x \in \mathbb{Z}$ such that $ax = b$. Since $b \mid c$, there is $y \in \mathbb{Z}$ such that $by = c$. Then

$$c = by = (ax)y = a(xy),$$

so $a \mid c$.
b) Since $a \mid b$, there is $x \in \mathbb{Z}$ such that $ax = b$. Since $a \mid c$, there is $z \in \mathbb{Z}$ such that $az = c$. Then

$$db + ec = d(ax) + e(az) = a(dx + ez),$$

so $a \mid db + ec$.
c) (i) $\implies$ (ii): If $a \mid b$ and $b \mid a$ there are $x, y \in \mathbb{Z}$ such that $a = xb$ and $b = ya$. Then

$$a = xb = x(ya) = a(xy),$$

so

$$a(xy-1)=0.$$

So either $a=0$ or $xy=1$. If $a=0$ then $b=y\cdot 0=0$, so $|a|=0=|b|$. If $xy=1$, then since $x,y\in\mathbb{Z}$ we have either $x=y=1$ or $x=y=-1$ (see Exercise 5.7). If $x=y=1$ then $a=b$, while if $x=y=-1$ then $a=-b$; either way, $|a|=|b|$.
(ii) $\implies$ (i): If $|a|=|b|$ then $a=b$ or $a=-b$. Either way, $a\mid b$ and $b\mid a$. □

Recall that for $a,b\in\mathbb{Z}$ we say that **a properly divides b** if $a\mid b$ and $b\nmid a$.

PROPOSITION 5.5. a) *For all $a\in\mathbb{Z}\setminus\{0\}$, we have that a properly divides 0.*
b) *The integer 0 does not properly divide any integer.*
c) *For $a,b\in\mathbb{Z}\setminus\{0\}$, the following are equivalent:*
(i) *The integer a properly divides the integer b.*
(ii) *There is $c\in\mathbb{Z}$ with $|c|>1$ such that $ac=b$.*
d) *If $a,b\in\mathbb{Z}\setminus\{0\}$ and a properly divides b, then $|a|<|b|$.*

PROOF. a) If a is a nonzero integer, then $a\mid 0$ and $0\nmid a$, so a properly divides 0.
b) Let $b\in\mathbb{Z}$. Then $b\mid 0$, so 0 does not properly divide b.
c) Let a and b be nonzero integers.
(i) $\implies$ (ii) Suppose that a properly divides b. Then there is a unique integer c such that $ac=b$. Since $b\neq 0$ we have $c\neq 0$. Since $b\nmid a$ we have $c\neq\pm 1$. It follows that $|c|>1$.
(ii) $\implies$ (i): If there is $c\in\mathbb{Z}$ such that $|c|>1$ and $ac=b$, then certainly $a\mid b$. Moreover, we have $a=\frac{1}{c}b$, and since $|c|>1$, we have $\frac{1}{c}\notin\mathbb{Z}$, so $b\nmid a$. It follows that a properly divides b.
d) By part c), if a and b are nonzero integers such that a properly divides b, there is $c\in\mathbb{Z}$ with $|c|>1$ such that $ac=b$. It follows that $|b|=|c||a|>|a|$. □

3. Prime and Composite Numbers

A **prime number** is an integer $p>1$ such that if a positive integer d divides p, then either $d=1$ or $d=p$.

For example, 2, 3, 5 and 7 are primes, while 4 and 6 are not.

We call an integer $n>1$ **composite** if n is *not* prime. In n is composite, then it has a divisor a with $1<a<n$. Then we also have $\frac{n}{a}\mid n$ and $1<\frac{n}{a}<n$. Writing $n=a\cdot\frac{n}{a}$, we have established the following characterization of composite numbers.

PROPOSITION 5.6. *An integer $n>1$ is composite if and only if there are integers a,b with $1<a,b<n$ such that $n=ab$.*

PROPOSITION 5.7. *Every integer $n>1$ is divisible by some prime number.*

PROOF. Let $n\in\mathbb{Z}^{\geq 2}$. If n is itself prime, then we are done. If n is not prime, then there are integers a_1,b_1 with $1<a_1,b_1<n$ such that $n=a_1b_1$. But since $a_1\in\mathbb{Z}^{\geq 2}$, the same reasoning applies: either a_1 is prime, in which case since $a_1\mid n$ we are done, or $a_1=a_2b_2$ for integers a_2,b_2 $1<a_2,b_2<a_2$. Since $a_2\in\mathbb{Z}^{\geq 2}$, the same reasoning applies: either a_2 is prime, in which case $a_2\mid a_1\mid n$ and we're done, or $a_2=a_3b_3$ for integers a_3,b_3 with $1<a_3,b_3<a_2$. Continuing in this manner,

we either eventually reach a prime number a_n, in which case $a_n \mid a_{n-1} \mid \ldots \mid n$ and we are done, or we generate and infinite sequence of such integers a_n. In the latter case, since $a_n = a_{n+1}b_{n+1}$ with $b_{n+1} \in \mathbb{Z}^{\geq 2}$ we have that a_{n+1} properly divides a_n, and our sequence $a_1, a_2, \ldots, a_n, \ldots$ contradicts the Principle of Infinite Divisibility (Corollary 4.12). So the latter case cannot occur: for some $n \in \mathbb{Z}^+$ we have that a_n is a prime divisor of n. □

4. Greatest Common Divisors

Early on in school one studies the greatest common divisor of two positive integers. This treatment misses some subleties that we want to discuss here, so let us make the following distinction: if a and b are integers, not both 0, then the **biggest common divisor** $\text{bcd}(a, b)$ is the largest integer d such that $d \mid a$ and $d \mid b$. In other words, we are asserting that the set

$$D(a, b) \coloneqq \{n \in \mathbb{Z} \mid n \mid a \text{ and } n \mid b\}$$

has a maximum, and we call that $\text{bcd}(a, b)$. To see why this is true: first of all $1 \in D(a, b)$, so the set is nonempty.

Suppose first that $a \neq 0$. If $d \mid a$, then there is $c \in \mathbb{Z} \setminus \{0\}$ such that $cd = a$, so $|c| \geq 1$ and thus

$$|d| = \frac{|a|}{|c|} \leq |a|.$$

The set of integers of absolute value less than $|a|$ is $\{-|a|, -|a|+1, \ldots, 0, 1, \ldots, |a|\}$, so $D(a, b)$ is nonempty and finite and thus has a maximum. Since $1 \in D(a, b)$, that maximum must be a postive integer. If instead $b \neq 0$ then the argument is identical.

In contrast, because every integer divides 0, we have

$$D(0, 0) = \{n \in \mathbb{Z} \mid n \mid 0 \text{ and } n \mid 0\} = \mathbb{Z},$$

which does not have a maximum. So there is no such thing as $\text{bcd}(0, 0)$.

Now let $a, b \in \mathbb{Z}$. We say that an integer d is a **greatest common divisor** of a and b if for every $e \in \mathbb{Z}$ such that $e \mid a$ and $e \mid b$, we also have $e \mid d$. In other words, here the optimization is not using the usual $\leq$ relation on $\mathbb{Z}$ but using the divisibility relation. Otherwise put, while above we saw that for any $a, b \in \mathbb{Z}$ not both 0 the set $D(a, b)$ of common divisors has a maximum, we now want to know whether $D(a, b)$ contains a "maximally divisible element," i.e., an element that is divisible by every other element.

Proposition 5.8. *An integer d is a greatest common divisor of 0 and 0 if and only if $d = 0$.*

Proof. To see that 0 is a common divisor of 0 and 0, the key observation is that every integer divides 0. This shows that for any $e \in \mathbb{Z}$ we have that e is a common divisor of 0 and 0 and (for the third time!) $e \mid 0$, so 0 is a greatest common divisor of 0 and 0.

Conversely, let d be a nonzero integer. Then $2d \mid 0$ and $2d \mid 0$ (yes, the same thing twice), but $2d \nmid d$, so d is not a greatest common divisor of 0 and 0. □

Thus, while $\text{bcd}(0, 0)$ does not exist, we have $\gcd(0, 0) = 0$.

The next case to look at is when one of a and b is 0 and the other is nonzero. Without loss of generality we may suppose that $a \neq 0$ and $b = 0$.

PROPOSITION 5.9. *Let a be a nonzero integer.*

a) *We have* $\operatorname{bcd}(a, 0) = |a|$.

b) *An integer d is a greatest common divisor of a and 0 if and only if $d = a$ or $d = -a$.*

PROOF. a) Since every integer divides 0, an integer d is a common divisor of a and 0 iff $d \mid a$. As we saw above, if $d \mid a$ then $|d| \leq |a|$. Since $|a| \mid a$, it follows that $|a|$ is the biggest among all divisors of a.
b) The assertion that a is a greatest common divisor of a and a unwinds to: if $d \mid a$ and $d \mid 0$ then $d \mid a$. That is certainly true. Also, for any integers d and a we have $d \mid a \iff d \mid -a$, so $-a$ is also a greatest common divisor of a and a. If d is any divisor of a other than a or $-a$, then $|d| < |a|$, so $a \nmid d$. □

So we may assume that a and b are both nonzero integers: this is by far the most interesting case.

Essentially the same argument as in the proof of Proposition 5.9b) shows that if d is a greatest common divisor of a and b, then so too is $-d$. You are asked to confirm this in Exercise 5.9.

PROPOSITION 5.10. *Let a and b be nonzero integers.*

a) *If d is a greatest common divisor of a and b, then d is also a greatest common divisor of $-a$ and b, of a and $-b$, and of $-a$ and $-b$.*

b) *If d is a greatest common divisor of a and b then the greatest common divisors of a and b are precisely d and $-d$.*

PROOF. a) You are asked to show this in Exercise Y.Y.
b) If d is a greatest common divisor of a and b, then by Exercise 5.9, so is $-d$. Now suppose that $f \in \mathbb{Z}$ is a greatest common divisor of a and b. In particular:
• Since f is a common divisor of a and b and d is a greatest common divisor of a and b, we have $f \mid d$.
• Since d is a common divisor of a and b and f is a greatest common divisor of a and b, we have $d \mid f$.
This gives that $|f| \leq |d|$ and $|d| \leq |f|$, so $|d| = |f|$ and thus $f = \pm d$. □

Proposition 5.10b) shows that if nonzero integers a and b have any greatest common divisors at all, then they have precisely two greatest common divisors, a positive one d and a negative one $-d$. Because of this, for integers a and b, not both 0, by convention we write $\gcd(a, b) = d$ to mean the unique positive greatest common divisor of a and b...assuming one exists.

PROPOSITION 5.11. *Let a and b be integers, not both zero, and suppose that a and b have a greatest common divisor. Then:*

$$\operatorname{bcd}(a, b) = \gcd(a, b).$$

PROOF. As above, let d be the positive integer that is a greatest common divisor of a and b. Let $e \in D(a, b)$, i.e., e is a common divisor of a and b. Then $e \mid d$, so

$$e \leq |e| \leq |d| = d.$$

This shows that d is the maximum of $D(a,b)$. □

Thus after clearing away a certain amount of smoke, our core task remains: show that any two positive integers a and b have a greatest common divisor. If they do, then the unique positive greatest common divisor *is* $\mathrm{bcd}(a,b)$.

PROPOSITION 5.12. *Let $a, b \in \mathbb{Z}$. Then:*

a) *For any $n \in \mathbb{Z}$, we have*
$$D(a,b) = D(b, a-nb).$$

b) *Suppose $b \in \mathbb{Z}^+$. Then there are $q, r \in \mathbb{Z}$ with $a = qb + r$ and $0 \leq r < b$, and we have*
$$D(a,b) = D(b,r). \tag{17}$$

PROOF. a) Let $e \in D(a,b)$, so $e \mid a$ and $e \mid b$. Thus there are $c_1, c_2 \in \mathbb{Z}$ such that $ec_1 = a$ and $ec_2 = b$. It follows that $a - nb = ec_1 - n(ec_2) = e(c_1 - nc_2)$, so $e \mid a - nb$. Thus e is also a common divisor of b and $a - nb$. Conversely, if $e \mid b$ and $e \mid a - nb$ then there are $c_3, c_4 \in \mathbb{Z}$ such that $ec_3 = b$ and $ec_4 = a - nb$, and then
$$a = nb + (a - nb) = n(ec_3) + (ec_4) = e(nc_3 + c_4).$$
Thus e is also a common divisor of a and b.
b) The first assertion is precisely the Division Theorem. By part a) we get $D(a,b) = D(b, a - qb) = D(b,r)$. □

If you have never seen (17) before, your first reaction may be "Okay, it's true, but who cares?" The point is that writing $D(a,b) = D(b,r)$ is making progress on the problem of not only showing that a and b have a greatest common divisor but also on computing it.

Indeed, if $r = 0$ then $D(a,b) = D(b,0)$ is the set of all divisors of b, and then as seen above we would have $\gcd(a,b) = \gcd(b,0) = b$. For reasons that will become clear shortly, let us put
$$r_{-1} := a,\ r_0 := b,\ r_1 := r.$$
If r_1 is again positive, then we can apply the Division Algorithm again: there are $q_2, r_2 \in \mathbb{Z}$ with
$$b = q_2 r_1 + r_2 \text{ and } 0 \leq r_2 < r$$
and we have
$$D(a,b) = D(b,r) = D(r,r_2).$$
If $r_2 = 0$ then we have
$$\gcd(a,b) = \gcd(b,r) = \gcd(r,0) = r.$$
If $r_2 > 0$, we can apply the Division Algorithm yet again: there are $q_3, r_3 \in \mathbb{Z}$ with
$$r = q_3 r_2 + r_3 \text{ and } 0 \leq r_3 < r_2$$
and
$$D(a,b) = D(b,r) = D(r,r_2) = D(r_2,r_3).$$
Can this process go on forever? No, it cannot: that would yield an infinite sequence
$$b = r_0 > r_1 > r_2 > r_3 > \dots$$
of positive integers, contradicitng the Principle of Infinite Descent (Corollary 4.11).

This shows that for any $a \in \mathbb{Z}$ and any $b \in \mathbb{Z}^+$ (above we reduced the case of

b any integer to this case), not only does $\gcd(a,b)$ exist, but it can actually be computed by the above procedure of repeated division with remainder: we get a finite sequence

$$r_0, r_1 r_2, \ldots, r_n, 0$$

with $b = r_0 > r_1 > r_2 > \ldots r_n > r_{n+1} = 0$, and then $\gcd(a,b) = r_n$.

This procedure for computing the gcd of two positive integers is called the **Euclidean Algorithm**.

In Exercise 5.4 you are asked to show that for $a \in \mathbb{Z}^+$ and $b \in \mathbb{Z}$, if $a \mid b$ then $\gcd(a,b) = a$.

5. The GCD as a Linear Combination

THEOREM 5.13. *Let $a, b \in \mathbb{Z}$. Then there are integers x and y such that*

$$\gcd(a,b) = xa + yb.$$

PROOF. First: as we saw above, $\gcd(0,0) = 0$. Since for any $x, y \in \mathbb{Z}$ we have $0 = x \cdot 0 + y \cdot 0$, we are done in this case.
Next: If $b = 0$, then

$$\gcd(a,b) = |a| = (\pm 1)a + 0 \cdot b,$$

so we are done in this case. The same argument handles the case of $a = 0$. So we may assume that $a, b \in \mathbb{Z} \setminus \{0\}$. Since $\gcd(-a,b) = \gcd(-a,-b) = \gcd(a,-b) = \gcd(a,b)$, we reduce to the case a and b being positive: for instance, since

$$\gcd(-3,-5) = 1 = 3 \cdot (-3) + (-2) \cdot (-5),$$

we have

$$\gcd(3,5) = 1 = (-3) \cdot 3 + 2 \cdot 5.$$

So suppose $x, y \in \mathbb{Z}^+$. Now we consider the set

$$L^+(a,b) \coloneqq \{xa + yb \mid x, y \in \mathbb{Z}\} \cap \mathbb{Z}^+,$$

i.e., the set of positive integers that are integer linear combinations of x and y. This set contains $a = 1 \cdot a + 0 \cdot b$ (and also $b = 0 \cdot a + 1 \cdot b$), hence is nonempty. By the Well-Ordering Princple, the set $L^+(a,b)$ has a minimum element

$$d = xa + yb.$$

We claim that $d = \gcd(a,b)$. Because $d \in \mathbb{Z}^+$, to see this we must show that d is a divisor of both a and b and that for any integer e, if $e \mid x$ and $e \mid y$ then $e \mid d$.

To see that $d \mid a$ we apply the Division Theorem: there are $q, r \in \mathbb{N}$ such that $a = qd + r$ and $0 \leq r < d$. If $r = 0$ then $d \mid a$ and we're done. If $r > 0$, then

$$r = a - qd = x - q(xa + yb) = (1 - qx)a + yb,$$

so r is an element of $L^+(a,b)$ that is smaller than d, contradicting the fact that d is the least element of $L^+(a,b)$. The proof that $d \mid b$ is nearly identical.

Moreover, if $e \in \mathbb{Z}$ is such that $e \mid a$ and $e \mid b$, then $e \mid xa + yb = d$. □

The proof of Theorem 5.13 is a little theoretical. Here – unusually for this text! – we do not intend that as a complement. If one is given specific $a, b \in \mathbb{Z}^+$, it would be nice to *actually find* $x, y \in \mathbb{Z}$ such that $\gcd(a,b) = xa + by$. The proof we gave does not do this: it tells us that the minimum of a certain set $L^+(x,y)$ of positive

integers does the job but does not tell us how to find that minimum.

Euclid himself knew how to do better: in fact the Euclidean algorithm, which provides an efficient procedure for finding the gcd of two positive integers, can be tweaked to express this gcd as a linear combination of them.

EXAMPLE 5.14. *Let $a = 37$, $b = 27$. We perform the Euclidean algorithm, at each stage solving the resulting equation for the remainder*

$$37 = 1 \cdot 27 + 10, \text{ so } 10 = 37 - 1 \cdot 27,$$

$$27 = 2 \cdot 10 + 7, \text{ so } 7 = 27 - 2 \cdot 10,$$

$$10 = 1 \cdot 7 + 3, \text{ so } 3 = 10 - 1 \cdot 7,$$

$$7 = 2 \cdot 3 + 1, \text{ so } 1 = 7 - 2 \cdot 3.$$

The last step of the algorithm is

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

which shows that $\gcd(37, 27) = \gcd(1, 0) = 1$. *Now we reverse the steps:*

$$\gcd(37, 27) = 1 = 7 - 2 \cdot 3$$

$$= 7 - 2 \cdot (10 - 1 \cdot 7) = -2 \cdot 10 + 3 \cdot 7$$

$$-2 \cdot 10 + 3 \cdot (27 - 2 \cdot 10) = 3 \cdot 27 - 8 \cdot 10$$

$$= 3 \cdot 27 - 8 \cdot (37 - 1 \cdot 27) = 11 \cdot 27 - 8 \cdot 37.$$

It is not hard to see that this procedure of reversing the Euclidean Algorithm works in general for $a, b \in \mathbb{Z}^+$ to find $x, y \in \mathbb{Z}$ such that $\gcd(a, b) = xa + yb$: each step of the Euclidean Algorithm expresses the next remainder as an explicit $\mathbb{Z}$-linear combination of the previous two remainders, hence upon repeated substitution, of $r_{-1} = a$ and $r_0 = b$.

THEOREM 5.15. *For integers a, v, let*

$$L(a, b) := \{xa + yb \mid x, y \in \mathbb{Z}\}$$

be the set of $\mathbb{Z}$-linear combinations of a and b. For $n \in \mathbb{Z}$, the following are equivalent:

(i) *We have $n \in L(a, b)$.*
(ii) *We have $\gcd(a, b) \mid n$.*

PROOF. (i) $\implies$ (ii): If $n \in L(a, b)$ then $n = xa + yb$ for some $x, y \in \mathbb{Z}$. Since $\gcd(a, b)$ divides both a and b, it also divides $xa + yb$.
(ii) $\implies$ (i): If $\gcd(a, b) \mid n$ then we may write $n = c \gcd(a, b)$ for some $c \in \mathbb{Z}$. By theorem 5.13 there are $x, y \in \mathbb{Z}$ such that

$$\gcd(a, b) = xa + yb.$$

Multiplying through by c, we get that

$$n = c \gcd(a, b) = (cx)a + (cy)b \in L(a, b). \qquad \square$$

An important consequence of Theorem 5.15 is that for integers a and b, not both 0, the least positive integer that is a $\mathbb{Z}$-linear combination of a and b is $\gcd(a, b)$. This observation is enough to prove the following useful result.

THEOREM 5.16 (Scaling Property of GCDs). *Let $a, b \in \mathbb{Z}$ and let $d \in \mathbb{Z}^+$. Then we have*

$$\gcd(da, db) = d \gcd(a, b). \tag{18}$$

PROOF. We have that $\gcd(da, db)$ is the least positive element of $L(da, db)$. Every element of $L(da, db)$ is a multiple of d, and for all $n \in \mathbb{Z}$ we have $n \in L(da, db)$ if there are $x, y \in \mathbb{Z}$ such that $n = x(da) + y(db)$ if and only if there are $x, y \in \mathbb{Z}$ such that $\frac{n}{d} = xa + yb$ if and only if $\frac{n}{d} \in L(a, b)$. This shows that

$$L(da, db) = \{dn \mid n \in L(a, b)\},$$

i.e., the elements of $L(da, db)$ are obtained from the elements of $L(a, b)$ by multiplying by d. Since the least positive element of $L(a, b)$ is $\gcd(a, b)$, the least positive element of $L(da, db)$ is $d \gcd(a, b)$. It follows that $d \gcd(a, b) = \gcd(da, db)$. □

COROLLARY 5.17. *For integers a and b, not both 0, we have*

$$\gcd\left(\frac{a}{\gcd(a, b)}, \frac{b}{\gcd(a, b)}\right) = 1.$$

PROOF. There are $a', b' \in \mathbb{Z}$ such that $a = \gcd(a, b)a'$ and $b = \gcd(a, b)b'$. Using Theorem 5.16 we get

$$\gcd(a, b) = \gcd(\gcd(a, b)a', \gcd(a, b)b') = \gcd(a, b) \gcd(a', b').$$

Sicne $\gcd(a, b) \neq 0$, we conclude that

$$1 = \gcd(a', b') = \gcd\left(\frac{a}{\gcd(a, b)}, \frac{b}{\gcd(a, b)}\right). \qquad \square$$

6. Euclid's Lemma

We say that $a, b \in \mathbb{Z}$ are **coprime** if $\gcd(a, b) = 1$. e integer $p > 1$ whose only positive integer divisors are 1 and p.

PROPOSITION 5.18. *Let p be a prime number, and let $n \in \mathbb{Z}$.*

a) *If $p \mid n$, then $\gcd(p, n) = p$.*
b) *If $p \nmid n$, then $\gcd(p, n) = 1$.*

PROOF. a) This is a case of Exercise 5.6.
b) Let $d = \gcd(p, n)$, so $d \mid p$ and $d \mid n$. By definition of a prime number, we have either $d = 1$ or $d = p$. If $d = p$ then $p \mid n$, contrary to our hypothesis. So $d = 1$. □

We can now prove an important number-theoretic result.

THEOREM 5.19 (Euclid's Lemma). *Let p be a prime number, and let $a, b \in \mathbb{Z}$. If $p \mid ab$, then $p \mid a$ or $p \mid b$.*

PROOF. A good way to show $A \implies (B \vee C)$ is to assume A and $\neg B$ and deduce C. So in this case: we may assume that $p \mid ab$ and $p \nmid a$ and show that $p \mid b$.

Since p is prime and $p \nmid a$, by Proposition 5.18 we have $\gcd(p, a) = 1$, and then by Theorem 5.13 there are integers x and y such that

$$xp + ya = 1.$$

Multiplying this through by b, we get

$$xpb + yab = b.$$

Since $p \mid ab$, we get $p \mid xpb + yab = b$. □

The following is a generalization:

THEOREM 5.20 (Generalized Euclid's Lemma). *Let $a, b, c \in \mathbb{Z}$. If $a \mid bc$ and $\gcd(a, b) = 1$, then $a \mid c$.*

You are asked to prove Theorem 5.20 in Exercise 5.12. (Happily, the proof of Theorem 5.19 carries over easily.)

We also need the following extension of Euclid's Lemma to products with more than two terms.

THEOREM 5.21 (n-fold Euclid's Lemma). *Let p be a prime number, let $k \in \mathbb{Z}^+$, and let $n_1, \ldots, n_k \in \mathbb{Z}$. If $p \mid n_1 \cdots n_k$ then $p \mid n_i$ for some $1 \leq i \leq k$.*

Theorem 5.21 will be easy to deduce from Theorem 5.19 when we have the proof technique of induction available. We will return then to this in Chapter 6.

7. The Least Common Multiple

For positive integers a and b, along with the greatest common divisor $\gcd(a, b)$ one also learns early on about their **least common multiple**. We want to give a discussion of this parallel to that of greatest common divisors given above. So whereas above for $a, b \in \mathbb{Z}$ we considered the set

$$D(a, b) = \{n \in \mathbb{Z} \mid n \mid a \text{ and } n \mid b\}$$

of common divisors of a and b, now we wish to consider the set

$$M(a, b) \coloneqq \{n \in \mathbb{Z} \mid a \mid n \text{ and } b \mid n\}$$

of **common multiples of a and b.**

Once again we begin by disposing of the cases in which a or b is 0. Since the only integer that 0 divides is 0, the only multiple of 0 is 0, and therefore if either a or b is 0 we have

$$M(a, b) = \{0\}.$$

In this case 0 is the *only* common multiple.

So suppose now that $a, b \in \mathbb{Z} \setminus \{0\}$. If a and b have the same sign, then ab is a positive element of $M(a, b)$, while if a and b have opposite signs, then $-ab$ is a positive element of $M(a, b)$. Either way the set $M(a, b) \cap \mathbb{Z}^+$ is a nonempty set of positive integers, so has a least element. We call this least element $\operatorname{scm}(a, b)$ and call it the **smallest common multiple** of a and b.

The situation here is analogous to the one considered above for common divisors: for $a, b \in \mathbb{Z}^+$ what we have called the smallest common multiple $\operatorname{scm}(a, b)$ is what in pre-university mathematics is simply called the least common multiple. However, again we wish to reserve that name for a divisor that is "minimal with respect to divisibility," not merely minimal with respect to size. So we make the following definition: for $a, b \in \mathbb{Z}$, an integer m is a **least common multiple** of a and b if for every $n \in \mathbb{Z}$ such that $a \mid n$ and $b \mid n$ we have $m \mid n$. Thus a least common multiple of a and b is any integer in $M(a, b)$ that divides every element of $M(a, b)$.

PROPOSITION 5.22. *Let $a, b \in \mathbb{Z}$.*

a) *If one of a, b is 0, then 0 is the unique least common multiple of a and b.*

b) *If both a and b are nonzero integers, then* 0 *is* not *a least common multiple of a and b.*

You are asked to prove Proposition 5.22 in Exercise 5.13.

PROPOSITION 5.23. *Let $a, b \in \mathbb{Z} \setminus \{0\}$.*

a) *The set of common multiples of a and b does not depend upon the signs of a and b: we have*

$$M(a,b) = M(-a,b) = M(a,-b) = M(-a,-b).$$

b) *If m is a least common multiple of a and b, then the least common multiples of a and b are precisely m and $-m$.*

You are asked to prove Proposition 5.23 in Exercise 5.14. In view of this result, if two nonzero integers have a least common multiple, they have precisely one *positive* least common multiple, which we denote by $\operatorname{lcm}(a,b)$. If one of a and b is 0 then Proposition 5.22 justifies our putting $\operatorname{lcm}(a,b) = 0$, and we will (though this is a rather degenerate case).

PROPOSITION 5.24. *Let $a, b \in \mathbb{Z} \setminus \{0\}$. If a and b have a least common multiple, then we have*

$$\operatorname{scm}(a,b) = \operatorname{lcm}(a,b).$$

PROOF. If a and b have a least common multiple, then the set $M(a,b)$ of all common multiples is the set $\{n \operatorname{lcm}(a,b) \mid n \in \mathbb{Z}\}$ of all integer multiples of $\operatorname{lcm}(a,b) \in \mathbb{Z}^+$. It follows that $\operatorname{lcm}(a,b)$ is the least positive element of $M(a,b)$. □

PROPOSITION 5.25. *Let $a, b \in \mathbb{Z}$ be coprime: i.e., $\gcd(a,b) = 1$. Then ab is a least common multiple of a and b.*

PROOF. For any integers a and b, certainly ab is a common multiple of a and b. Now let n be any common multiple of a and b; we must show that $ab \mid n$. Since $b \mid n$, there is $e \in \mathbb{Z}$ such that $n = be$, and thus we have $a \mid n = be$. Since $\gcd(a,b) = 1$, by Theorem 5.20 we have $a \mid e$, so there is $f \in \mathbb{Z}$ such that $af = e$. It follows that $n = be = baf$, so $ab \mid n$. □

PROPOSITION 5.26 (Scaling Property of LCMs). *Let a and b be two nonzero integers that have a least common multiple. Then for all $d \in \mathbb{Z}^+$, the integers d and db have a least common multiple, and moreover we have*

$$\operatorname{lcm}(da, db) = d \operatorname{lcm}(a,b).$$

PROOF. Under the assumption that a and b have a least common multiple, the set $M(a,b)$ is the set of all integer multiples of $\operatorname{lcm}(a,b)$. For any $n \in M(da,db)$, we have $da \mid n$ and $db \mid n$. Certainly then $d \mid n$. Moreover there are $x, y \in \mathbb{Z}$ such that $n = dax$ and $n = dby$, so $\frac{n}{d} = ax$ and $\frac{n}{d} = by$, which shows that $a \mid \frac{n}{d}$ and $b \mid \frac{n}{d}$, so $\frac{n}{d} \in M(a,b)$. Similarly, if $\frac{n}{d} \in M(a,b)$ then $n \in M(da,db)$, and we conclude that

$$M(da,db) = \{dn \mid n \in M(a,b)\} = \{dn \operatorname{lcm}(a,b) \mid n \in \mathbb{Z}\},$$

and thus every element of $M(da,db)$ is a multiple of the positive integer $d \operatorname{lcm}(a,b)$. It follows that $\operatorname{lcm}(da,db)$ exists and equals $d \operatorname{lcm}(a,b)$. □

Finally we can prove that every pair of nonzero integers has a least common multiple and give a formula for it at the same time.

THEOREM 5.27. *Let $a, b \in \mathbb{Z} \setminus \{0\}$. Then a and b a least common multiple, and moreover we have*

$$\operatorname{lcm}(a,b) = \frac{|ab|}{\gcd(a,b)}.$$

PROOF. In view of Proposition 5.23a), we may assume that a and b are positive and show that $\operatorname{lcm}(a,b)$ exists and is equal to $\frac{ab}{\gcd(a,b)}$.

Put

$$a' = \frac{a}{\gcd(a,b)},\ b' = \frac{b}{\gcd(a,b)}.$$

By Corollary 5.17 we have $\gcd(a',b') = 1$, so by Proposition 5.25 the least common multiple of a' and b' exists and is equal to $a'b'$. Finally, Proposition 5.26 implies that $a = \gcd(a,b)a'$ and $b = \gcd(a,b)b'$ have a least common multiple, which is equal to $\gcd(a,b)\operatorname{lcm}(a',b') = \gcd(a,b)a'b'$. Thus

$$\operatorname{lcm}(a,b) = \gcd(a,b)a'b' = \gcd(a,b)\frac{a}{\gcd(a,b)}\frac{b}{\gcd(a,b)} = \frac{ab}{\gcd(a,b)}. \qquad \square$$

8. The Fundamental Theorem of Arithmetic

THEOREM 5.28 (Fundamental Theorem of Arithmetic). *Let $n > 1$ be an integer.*

a) *There is $k \in \mathbb{Z}^+$ and prime numbers $p_1, \ldots, p_k$ such that*

$$n = p_1 \cdots p_k. \tag{19}$$

b) *The factorization of (19) is unique, up to the order of the prime factors. That is: suppose that*

$$n = p_1 \cdots p_k = q_1 \cdots q_l,$$

are two factorizations of n into primes, with $p_1 \leq \ldots \leq p_k$ and $q_1 \leq \ldots \leq q_l$. Then $k = l$ and $p_i = q_i$ for all $1 \leq i \leq k$.

PROOF. a) Let S be the set of integers $n > 1$ that *cannot* be written as a product of prime numbers as in 19. Our task is to show that the subset $S \subseteq \mathbb{Z}^+$ is empty, so assume it isn't: by the Well-Ordering Principle there is then a minimum element $n \in S$. The element n cannot be prime, for otherwise it would be of the form (19) with $k = 1$. (In other words, though we speak of "products of primes," we allow there to be just one factor, in which case nothing is actually being multiplied.) Therefore we may write $n = ab$ with $1 < a, b < n$. Since a and b are each smaller than the minimum of S, they cannot lie in S, so they are each products of primes: there are $k_1, k_2 \in \mathbb{Z}^+$ and primes $p_1, \ldots, p_{k_1}, q_1, \ldots, q_{k_2}$ such that

$$a = p_1 \cdots p_{k_1},\ b = q_1 \cdots q_{k_2}.$$

But then we have

$$n = ab = p_1 \cdots p_{k_1} q_1 \cdots q_{k_2},$$

showing that n is a product of primes and thus not an element of S: contradiction. It follows that $S = \varnothing$.

b) Similarly to the proof of part a) above, let T be the set of integers $n > 1$ that admit at least two different factorizations into primes: that is, we can write

$$n = p_1 \cdots p_k = q_1 \cdots q_l$$

with $p_1 \leq \ldots \leq p_k$ and $q_1 \leq \ldots \leq q_l$ and such that the two finite lists $p_1, \ldots, p_k$ and $q_1, \ldots, q_l$ are not the same. If T is nonempty, let n be its minimum, and write

$$n = p_1 \cdots p_k = q_1 \cdots q_l.$$

Then $p_1 \mid n = q_1 \cdots q_l$, so by the n-fold Euclid's Lemma (Theorem 5.21) we have $p_1 \mid q_i$ for some $1 \leq i \leq l$. This may hold for several indices i, and we choose the *least* one for which this is the case. Since q_i is a prime number, its only divisor among integers greater than one is q_i itself, whereas the prime number p_1 is an integer greater than 1 that divides q_i. So we have $p_1 = q_i$, and we can write

$$\frac{n}{p_1} = p_2 \cdots p_k = q_1 \cdots q_{i-1} q_{i+1} \cdots q_l.$$

Since $\frac{n}{p_1} < n$, we cannot have $\frac{n}{p_1} \in T$. This forces $k = l$ and $p_2 = q_1$, $p_3 = q_2$, and so forth. Since $p_1 \leq p_i$ for all $1 \leq i \leq k$, it follows that $q_j \geq p_1$ for all $1 \leq j \leq l$; since also $q_i = p_1$, we conclude that $q_1 = p_1$ and therefore $i = 1$. That is, we have

$$q_1 = p_1, \ q_2 = p_2, \ldots, q_l = q_k = p_k$$

so in fact the two prime factorizations of n were the same. This contradiction shows that T is empty and completes the proof. □

9. Exercises

EXERCISE 5.1. *Let S be a finite nonempty subset of $\mathbb{R}$ of size N.*

a) *Show: S has a maximum element.*
b) *Show: we may write $S = \{x_1, \ldots, x_N\}$ with $x_1 < x_2 < \ldots < x_N$.*

EXERCISE 5.2. *Let S be an infinite subset of $\mathbb{Z}$. Show: S cannot have both a minimum and a maximum.*
(Suggestions: (i) you may assume that S has a minimum element x_1. The logic is as follows: to prove $A \implies (B \vee C)$, it suffices to assume $A \wedge (\neg B)$ and prove C. (ii) Show that $S \setminus \{x_1\}$ has an element $x_2 \geq x_1 + 1$. (iii) Repeat.)

EXERCISE 5.3. *For a subset $S \subseteq \mathbb{Z}$, show that the following are equivalent:*

(i) *The set S is well-ordered.*
(ii) *The set S has a minimum element.*

EXERCISE 5.4. *Show that for $a \in \mathbb{Z}^+$ and $b \in \mathbb{Z}$, if $a \mid b$ then $\gcd(a, b) = a$.*

EXERCISE 5.5. *Let $a, b, c \in \mathbb{Z}$ with $c \neq 0$. Show: $a \mid b \iff ac \mid bc$.*

EXERCISE 5.6 (The Frog Got Wet). *Suppose a frog is jumping its way down a linear road, always from left to right. The length of each jump can vary, but there is a maximum distance $b > 0$ that the frog can cover in a single jump. Suppose there is a puddle in the road of length $\ell \geq b$. If after a certain number of jumps the frog lies to the left of the puddle and after a larger number of jumps the frog lies to the right of the puddle, then we claim that **the frog got wet**: that is, after at least one jump the frog must land* in *the puddle.*

We formalize this as follows: let $0 < b \leq \ell$ be real numbers. We consider a strictly increasing sequence

$$x_1 < x_2 < \ldots < x_n < \ldots$$

of real numbers. Let $a \in \mathbb{R}$. We suppose:
(i) For all $n \in \mathbb{Z}^+$, we have $x_{n+1} - x_n \leq b$.

(ii) There is $n_1 \in \mathbb{Z}^+$ such that $x_{n_1} < a$ and $n_2 \in \mathbb{Z}^+$ such that $x_{n_2} > a + \ell$.
CLAIM*: There is an integer n_3 with $n_1 < n_3 < n_2$ such that $x_{n_3} \in [a, a + \ell]$.*

a) *Explain how the mathematical formalism models the jumping frog. What aspects of the "frog story" correspond to conditions (i) and (ii)? What subset of $\mathbb{R}$ corresponds to the puddle?*
b) *Prove the* CLAIM.
(Suggestion: consider the least $n \in \mathbb{Z}^+$ such that $x_n > a + \ell$.)

EXERCISE 5.7. *Let $x, y \in \mathbb{Z}$ be such that $xy = 1$. Show: either $x = y = 1$ or $x = y = -1$.*

EXERCISE 5.8. *Let $a \in \mathbb{Z}^+$ and $b \in \mathbb{Z}$. Show: if $a \mid b$, then $\gcd(a, b) = a$.*

EXERCISE 5.9. *Let $a, b, d \in \mathbb{Z}$. Show: if d is a greatest common divisor of a and b, then so is $-d$.*

EXERCISE 5.10. *A subset $I \subseteq \mathbb{Z}$ is an **ideal** if $I \neq \varnothing$, for all $x, y \in I$ we have $x + y \in I$ and for all $n \in \mathbb{Z}$ and $x \in I$ we have $nx \in I$.*

a) *Show: if I is an ideal, then $0 \in I$.*
b) *Show: $\{0\}$ is an ideal. If $\{0\} \subsetneq I$, then I contains a positive integer.*
c) *For integers $a_1, \ldots, a_n$, we put*

$$\langle a_1, \ldots, a_n \rangle \coloneqq \{x_1 a_1 + \ldots + x_n a_n \mid x_1, \ldots, x_n \in \mathbb{Z}\}.$$

*Show: $\langle a_1, \ldots, a_n \rangle$ is an ideal. An ideal of the form $\langle a_1 \rangle = \{n a_1 \mid n \in \mathbb{Z}\}$ is called **principal**.*
d) *Explain why it follows from Theorems 5.13 and 5.15 that every ideal of the form $\langle a, b \rangle$ is principal.*
e) *Show that in fact every ideal is principal.*
(Suggestion: Since $\{0\} = \langle 0 \rangle$, we may assume that $\{0\} \subsetneq I$ and thus I contains positive elements. Show that if d is the least positive element of I then $I = \langle d \rangle$.

EXERCISE 5.11. *For the sake of this exercise alone, we call an integer n **Euclidean** if it satisfies the conclusion of Euclid's Lemma: that is: for all $a, b \in \mathbb{Z}$ if $n \mid ab$ then $n \mid a$ or $n \mid b$. With this new terminology, Euclid's Lemma can be restated as: every prime number is Euclidean.*
Determine exactly which integers are Euclidean (and prove your answer!).

EXERCISE 5.12.

a) *Prove the **Generalized Euclid's Lemma** (Theorem 5.20).*
(Suggestion: Adapt the proof of Theorem 5.19.)
b) *Explain why the Generalized Euclid's Lemma implies Euclid's Lemma.*

EXERCISE 5.13. *Prove Proposition 5.22.*

EXERCISE 5.14. *Prove Proposition 5.23.*

EXERCISE 5.15. *For $a, b, n \in \mathbb{Z}$ show that the following are equivalent:*

(i) *We have $a \mid n$ and $b \mid n$.*
(ii) *We have $\operatorname{lcm}(a, b) \mid n$.*

EXERCISE 5.16. *Let $N > 1$, and let $a_1, \ldots, a_N \in \mathbb{Z}$. Show that there is a nonempty subset $J \subseteq [N]$ such that*

$$N \mid \sum_{i \in J} a_i.$$

(Suggestion: use the Pigeonhole Principle.)

CHAPTER 6

Fundamentals of Proof

1. Vacuously True and Trivially True Implications

As mentioned in §2.8, here is by far the most common form of results that we want to prove: we have a nonempty (usually infinite) set S, open sentences $P(x)$ and $Q(x)$ with domain $x \in S$, and we wish to show the universally quantified implication

$$\forall x \in S,\ P(x) \implies Q(x). \tag{20}$$

We normally conceptualize and prove this by establishing a *logical relationship* between $P(x)$ and $Q(x)$: even the way we read the symbol $\implies$, "implies" suggests a *logical entailment* between $P(x)$ and $Q(x)$. However, as we discussed when defining the symbol $\implies$, this is not strictly speaking the case. In fact, there are two extreme cases in which (20) holds even though $P(x)$ and $Q(x)$ may have *nothing* to do with each other.

EXAMPLE 6.1. *Consider the following implication:*

For all $x \in \mathbb{Z}$, if $6x$ is a prime number, then the Riemann Hypothesis holds.

Now the Riemann Hypothesis is the most famous open problem in mathematics.[1] *If you solve it correctly you will receive one million dollars, but I would say that's an underbilling of its importance. However, the above implication is clearly true even though we don't know whether the Riemann Hypothesis holds. This is because $6x$ is never a prime number: x needs to be positive in order for $6x$ to even be a positive integer, and then it is divisible by 1, 2, 3, and 6. Since an implication is true whenever its hypothesis is false, this implication is true for all $x \in \mathbb{Z}$.*

We say (the terminology here is rather standard) that the quantified implication (20) is **vacuously true** if the hypothesis is never true: that is, for no $x \in S$ does $P(x)$ hold. What is notable here – and may seem at terms silly or distressing – is that a vacuously true implication is indeed true, but *not* because of any logical relationship between $P(x)$ and $Q(x)$, as the above example (I hope) makes clear.

Here is an example of a similar phenomenon.

EXAMPLE 6.2. *Consider the following implication:*

For all $k \in \mathbb{Z}^+$, if $\sum_{n=1}^{\infty} \frac{1}{n^{2k+1}}$ is irrational, then $2k+1$ is odd.

We learn in calculus that for a real number p, the series $\sum_{n=1}^{\infty} \frac{1}{n^p}$ converges if and only if $p > 1$ [**Cl-HC**, Thm. 11.15]*. Whether it converges to a* rational number *is a much more delicate question. Taking $k = 1$ we are asking whether $\sum_{n=1}^{\infty} \frac{1}{n^3}$ is irrational. In one of the more surprising mathematical developments of the 20th century, this was shown in 1979 by Apéry* [**Ap79**]*. Taking $k = 2$, we are*

[1]For the problem statement by the organization that has endowed the prize, see `https://www.claymath.org/millennium-problems/riemann-hypothesis`.

asking whether $\sum_{n=1}^{\infty} \frac{1}{n^5}$ *is irrational. This remains unknown. On the other hand, look at the conclusion:* $2k+1$ *is odd. Well, yes, of course it is. So whereas the hypothesis of the implication is* believed *to always be true but is extremely difficult to establish even for particular values of* k*, the conclusion is (obviously) always true, and therefore the implication is true.*

In [**CPZ**], Chartrand Polimeni and Zhang call a universally quantified implication **trivially true** if $Q(x)$ is true for all $x \in S$. This terminology is *not* standard, but I like it and will use it. The implication of Example 6.2 is trivially true.

To me it is clear that the concepts of vacuous truth and trivial truth have some pedagogical value in understanding the rules of an implication. One can ask whether they have any real use in mathematical practice. As for vacuous truth, the merit of an implication $P \implies Q$ is that if you are in a situation where you know that P is true, you can deduce Q, and among all true implications the vacuously true ones are precisely those for which this will never occur. For a trivially true implication: okay, there may be values of x for which $P(x)$ is true, in which case you can deduce $Q(x)$ (in Example 6.2 this holds for $k = 1$), but...wouldn't it be much more useful just to record $\forall x \in S, Q(x)$ in this case rather than (20)?

The answer to this if that if you *know* that (20) is vacuously true or trivially true, then expressing it as an implication is not very helpful for precisely the reasons just mentioned. However, in practice we may be able to show the implication (20) by the usual means of finding a logical entailment, and then we may ask whether the implication is in fact vacuously or trivially true. Both of these possibilities arise in normal mathematical practice.

That a proven implication may turn out to be vacuously true is a sort of "grad student's bane" – there are many horror stories of students who spend months or years proving a statement of the form (20) only to be asked at their thesis defense of an example of $x \in S$ for which $P(x)$ is true, and the conclusion is either they don't know any or (worse) a professor on their committee tells them that there is no such $x \in S$: i.e., the implication is vacuously true and therefore tells us nothing we didn't already know. Like most horror stories, this almost never happens in the exact form in which it was told, but the *fear* is of something real. In general, one can work very hard to prove a quantified implication (20) and it may look like important work, but the question of how often the hypothesis is true is important in evaluating the merit of the result...and can be hard to evaluate.

In mathematics we often prove statements that we *think* are trivially true but do not *know* are trivially true. That is, we want to show that $Q(x)$ holds for all $x \in S$ but we can only show this based on some other hypothesis $P(x)$ that we do not know to be true for all $x \in S$. Such a result in mathematics is called **conditional**. Often these conditions are major conjectures that are widely believed but seem difficult to prove. In fact, notice the second word of "Riemann Hypothesis": indeed this assertion (and also certain generalizations, which go under the name "GRH") is often used to prove other results. In this case one says that one has proved a result **conditionally on GRH**, and in my branch of mathematics there are many results like this (I have a few).

I will admit though that in the context of coures like this – i.e., undergraduate

courses introducing logic and proof – it is common for instructors to ask students to prove quantified implications (20) in situations where the implication is either vacuously true or trivially true. So beware!

In the following sections we discuss the three basic proof formats: direct proof, contrapositive and proof by contradiction.

2. Direct Proof

To directly prove

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

we need to show that for every $x \in S$ such that $P(x)$ is true, we also have that $Q(x)$ is true. Such an argument begins and ends as follows:

"Let $x \in S$ and suppose that $P(x)$ is true. [...] Then $Q(x)$ is true."

(Obviously it's what's in the middle that counts!) Here it is understood that we are not *choosing* an element of S; rather, the argument supplied in the [...] must apply to *every* $x \in S$ such that $P(x)$ is true. There two ways to do this:

I. We make an argument that is *sufficiently general* in character that it applies simultaneously to all $x \in X$.

This is somehow touching on the essence of pure mathematics: we want to find *general* arguments that explain why something is *always* true. However, sometimes this aspiration is not quite attained, and we may need to argue differently for different $x \in S$. To be formal about it, this works as follows:

II. We find an indexed family $\{Y_i\}_{i \in I}$ of subsets $Y_i \subseteq S$ such that $\bigcup_{i \in I} Y_i = S$ (when this happens, we say that the family $\{Y_i\}_{i \in I}$ **covers** S) and then for each $i \in I$ we show separately that

$$\forall x \in Y_i,\ P(x) \implies Q(x).$$

Let us see some simple examples.

PROPOSITION 6.3. *For all $x \in \mathbb{Z}$, if x is even then $x^2 + 2021$ is odd.*

PROOF. Here we can follow strategy I.:

Let x be an even integer. Then we may write $x = 2a$ for some (unique!) $a \in \mathbb{Z}$, and we find that

$$x^2 + 2021 = (2a)^2 + 2021 = 4a^2 + 2021 = 2(2a^2 + 1010) + 1$$

is odd. □

PROPOSITION 6.4. *For all $x \in \mathbb{Z}$, the integer $x^2 + x$ is even.*

Before doing the proof, we observe that $x^2 + x = x(x + 1)$. Here the result seems more clear in a certain case: if x is even, then $x(x + 1)$ is also even. If we can also handle the case in which x is odd, then we'll be done, and once one has this idea of dividing into cases in this way, it is not hard to finish the proof.

PROOF. Let $x \in \mathbb{Z}$. We consider cases:
Case 1: Suppose x is even, so $x = 2a$ for some $a \in \mathbb{Z}$. Then

$$x(x+1) = 2a(x+1) = 2(a(x+1))$$

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

$$x(x+1) = x(2a+1+1) = x(2a+2) = 2(x(a+1))$$

is even. □

One might notice that the statement of Proposition 6.4 doesn't quite adhere to the template $\forall x \in S,\ P(x) \implies Q(x)$ because there is no hypothesis $P(x)$. Rather we are being asked to show a statement of the form

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

This did not trouble us. If we just want to "put things as they were," we could say that $P(x)$ is the statement "$x \in S$." This converts any statement of the form

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

to an implication of the form

$$\forall x \in S, (x \in S) \implies Q(x)$$

which is a bit repetitive but certainly equivalent.

In general, when setting up mathematical statements there is some leeway between *domain* and *hypothesis*.

3. Contrapositive

''Everything will be okay in the end. If it's not okay, it's not the end." – John Lennon

Supppose as usual that we are trying to prove a statement of the form

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

This is logically equivalent to the contrapositive form

$$\forall x \in S,\ \neg Q(x) \implies \neg P(x),$$

so, if helpful, we can prove that instead. Here is a first example.

PROPOSITION 6.5. *Show: for all $x \in \mathbb{Z}$, if x^2 is even, then x is even.*

PROOF. Here direct reasoning is not so helpful: if we assume that x^2 is even, then $x^2 = 2a$ for some $a \in \mathbb{Z}$. So then what do we know about x? We have $x = \frac{2a}{x}$...and now I'm not sure what to do.

The contrapositive form of the statement is: "For all $x \in \mathbb{Z}$, is x is odd, then x^2 is odd." If $x \in \mathbb{Z}$ is odd, then $x = 2a+1$ for some $a \in \mathbb{Z}$, so

$$x^2 = (2a+1)^2 = 4a^2 + 4a + 1 = 2(2a^2 + 2a) + +1$$

is odd. □

In this case, the contrapositive was helpful because in matters of divisibility, we would rather know something about x and prove something about some multiple of x then vice versa.

PROPOSITION 6.6. *For $x \in \mathbb{Z}$, we have $3 \mid x \iff 3 \mid x^2$.*

PROOF. FIRST PROOF: Let $x \in \mathbb{Z}$. We will show that $(3 \mid x) \implies (3 \mid x^2)$ and also that $(3 \mid x^2) \implies (3 \mid x)$.
• $(3 \mid x) \implies (3 \mid x^2)$: Suppose that $3 \mid x$, so $x = 3a$ for some $a \in \mathbb{Z}$. Then $x^2 = (3a)^2 = 9a^2 = 3(3a^2)$, so $3 \mid x^2$.
• $(3 \mid x^2) \implies (3 \mid x)$: We will show the contrapositive: $(3 \nmid x) \implies (3 \nmid x^2)$. Suppose $3 \nmid x$. By the Divsion Theorem we may write $x = 3a + r$ for $0 \leq r < 3$. Indeed we cannot have $r = 0$, for then $3 \mid x$, so either $r = 1$ or $r = 2$.
Case 1: If $x = 3a + 1$, then $x^2 = (3a + 1)^2 = 9a^2 + 6a + 1 = 3(3a^2 + 2a) + 1$ is not divisible by 3.
Case 2: If $x = 3a + 2$, then $x^2 = (3a + 2)^2 = 9a^2 + 12a + 4 = 3(3a^2 + 4a + 1) + 1$ is not divisible by 3.
SECOND PROOF: We partition $\mathbb{Z}$ into three sets according to the remainder upon division by 3:

$$\mathcal{R}_0 \coloneqq \{3a \mid a \in \mathbb{Z}\},\ \mathcal{R}_1 \coloneqq \{3a + 1 \mid a \in \mathbb{Z}\},\ \mathcal{R}_2 \coloneqq \{3a + 2 \mid a \in \mathbb{Z}\}.$$

We claim that for all $x \in \mathcal{R}_0$ both $3 \mid x$ and $3 \mid x^2$ hold, whereas for all $x \in \mathcal{R}_1 \cup \mathcal{R}_2$, neither $3 \mid x$ nor $3 \mid x^2$ holds. This suffices. Since $\mathcal{R}_0$ is precisely the set of integers that are divisible by 3 and $\mathcal{R}_1 \cup \mathcal{R}_2 = \mathbb{Z} \setminus \mathcal{R}_0$, it is clear that $3 \mid x$ holds for all $x \in \mathcal{R}_0$ and for no $x \in \mathcal{R}_1 \cup \mathcal{R}_2$. So we need to check the divisibility of x^2 by 3.
Case 1: If $x \in \mathcal{R}_0$, then $x = 3a$, so $x^2 = 9a^2 = 3(3a^2)$ is divisible by 3.
Case 2: If $x \in \mathcal{R}_1$, then $x = 3a+1$, so $x^2 = (3a+1)^2 = 9a^2+6a+1 = 3(3a^2+2a)+1$ is not divisible by 3.
Case 3: If $x \in \mathcal{R}_2$, then $x = 3a + 2$, so $x^2 = (3a + 2)^2 = 9a^2 + 12a + 4 = 3(3a^2 + 4a + 1) + 1$ is not divisible by 3. □

It is interesting to compare the first and second proofs of Proposition 6.6. Notice that the calculations are identical! What differs is the logic. In particular, the second proof manages to avoid the contrapositive so is perhaps a bit logically more straightforward.

PROPOSITION 6.7. *Let $a \in \mathbb{Z}^+$. If $2^a - 1$ is prime, then a is prime.*

PROOF. This is a natural candidate for a proof by contrapositive, since we would rather use properties of a to study $2^a - 1$ then properties of $2^a - 1$ to study a. So we will show: for all $a \in \mathbb{Z}^+$, if a is not prime, then $2^a - 1$ is not prime.

First of all, since a prime number is a positive integer $a > 1$ that is only divisible by 1 and itself, if a is not prime then either $a = 1$ or $a = bc$ with $1 < b, c$.
Case 1: Indeed if $a = 1$, then $2^a - 1 = 1$ is not prime.
Case 2: Suppose that $a = bc$ with $1 < b, c$, what we need to do is factor $2^{bc} - 1$. Recall that for any $r \in \mathbb{R} \setminus \{1\}$ and $N \in \mathbb{Z}^+$, we have

$$1 + r + \ldots + r^N = \frac{r^{N+1} - 1}{r - 1} :$$

indeed if we put

$$S \coloneqq 1 + \ldots + r^N,$$

then

$$rS = r + \ldots + r^N + r^{N+1},$$

so

$$(r - 1)S = r^{N+1} - 1$$

and thus

$$1 + r + \ldots + r^N = S = \frac{r^{N+1} - 1}{r - 1}.$$

Applying this with $r = 2^b$ and $N = c - 1$ we get

$$1 + 2^b + \ldots + (2^b)^{c-1} = \frac{(2^b)^c - 1}{2^b - 1},$$

so

$$(2^b - 1)(1 + 2^b + \ldots + (2^b)^{c-1} = 2^{bc} - 1.$$

Since $b > 1$, we have $2^b - 1 > 2^1 - 1 = 1$, and since $c > 1$, we have $1 + 2^b + \ldots + (2^b)^{c-1} \geq 1 + 2^b > 1$, so this is a nontrivial factorization of $2^{bc} - 1$, which is accordingly not prime. □

A prime number of the form $2^p - 1$ is called a **Mersenne prime**. At first it seems like it might be true that conversely, if p is a prime, then $2^p - 1$ is prime: this holds for $p = 2, 3, 5, 7$. However

$$2^{11} - 1 = 2047 = 23 \cdot 89.$$

As of January 2023, there are precisely 51 known Mersenne primes; the largest is

$$2^{82,589,933} - 1,$$

which has $24,862,048$ digits. As we know, there are infinitely many primes. It is generally believed that there should be infinitely many Mersenne primes, but it would be extremely surprising if this were proved (or disproved!) in the near future.

4. Contradiction

4.1. Logical Basics. The logical kernel of a proof by contradiction is the following basic and familiar observation: if an implication $P \implies Q$ holds and the conclusion Q is false, then the hypothesis P must be false. In this form this is very closed to the contrapositive, but let's twist it around a bit:

Variant: if $(\neg P) \implies Q$ holds and Q is false, then $\neg P$ is false...so P is true.

This gives a way of proving P: assume that P is false, and from that deduce a false result. In order to argue correctly and reach a false conclusion, we must have a false premise, namely $\neg P$. So P is true.

This is a very common line of reasoning, both inside and outside of mathematics. We have certainly used this type of argument many times already in this text. It is perhaps more interesting to ask why this is called a *proof by contradiction*. Recall that in logic we use "contradiction" both for a logical expression in $P_1, \ldots, P_n$ that evluates to false for all truth values of $P_1, \ldots, P_n$ and also for a statement of the form $P \wedge (\neg P)$. The latter is a specific instance of the former, since $P \wedge (\neg P)$ is false whether P is true or false. On the other hand if we deduce any statement Q that is known to be false, then that means that we also know that $\neg Q$ is true, and thus we can deduce $Q \wedge (\neg Q)$. So the two meanings are not really very different.

Let us again consider our favorite form of a statement to prove:

$$\forall x \in S,\ P(x) \implies Q(x). \tag{21}$$

In order to prove this by contradiction, we go as follows: let $x \in S$. Assume $\neg(P(x) \implies Q(x))$ and deduce a contradiction. That is, we assume $P(x) \wedge (\neg Q(x))$ and deduce a contradiction.

By comparing with our previous two logical approaches to proving (21) we can see the power of proof by contradiction.

Direct Proof: Let $x \in S$. Assume $P(x)$ and deduce $Q(x)$.
Contrapositive: Let $x \in S$. Assume $\neg Q(x)$ and deduce $\neg P(x)$.
Contradiction: Let $x \in S$. Assume $P(x)$ and $\neg Q(x)$, and deduce a contradiction.

Thus in the other two approaches we assume one thing, whereas in a proof by contradiction we get to assume two things. That makes it seem much more powerful. In practice, this is often true.

One comment on the power of proof by contradiction: we can use it in place of the proof by contrapositive at all times. Namely, suppose that we are trying to prove (21), and for $x \in S$ we have proved that

$$\neg Q(x) \implies \neg P(x).$$

Now we assume that $P(x) \implies Q(x)$ is false, so $P(x)$ is true and $Q(x)$ is false. By our just proved implication, we know that $\neg P(x)$ is true. Thus $P(x)$ and $\neg P(x)$ are both true: contradiction.

This may explain why proof by contradiction is a popular way to argue even outside of mathematics and logic, whereas the typical intelligent private citizen does not have "contrapositive" in their vocabulary. Having said that:

Pro Tip: Do *not* cast proofs by contrapositive as proofs by contradiction. A proof by contrapositive is a "positive" argument for $\neg Q \implies \neg P$, while a proof by contradiction is "subjunctive" or "counterfactual": the proof begins with a false premise, so *other* than establishing the premise is false, a proof by contradiction cannot show anything, while in arguing $\neg Q \implies \neg P$ you may go

$$\neg Q \implies X_1 \implies X_2 \implies \dots X_n \implies P$$

for some intermediate statements X_1 through X_n, and then that all of these are implied by $\neg Q$ are facts that you can use outside of the context of the given proof.

4.2. Pythagoras and Euclid. We now give the two most famous proofs by contradiction. In the second case, although it is easier to present the argument as a proof by contradiction, we will explain how arguing a bit differently allows us to retain information established in the proof rather than just the statement itself.

THEOREM 6.8. *The square root of* 2 *is irrational.*

PROOF. A more careful formulation of the statement is:

$$\text{"For all } x \in \mathbb{R}, \text{ if } x \in \mathbb{Q} \text{ then } x^2 \neq 2."$$

Seeking a contradiction, let $x \in \mathbb{Q}$ be such that $x^2 = 2$. Then $x \neq 0$, so we may

write $x = \frac{a}{b}$ with a and b relatively prime integers. In particular, we may assume that a and b are not both even. Now we have

$$\frac{a}{b} = \sqrt{2},$$

and squaring both sides, we get

$$\frac{a^2}{b^2} = 2.$$

Simplifying, we get

$$a^2 = 2b^2.$$

This shows that $2 \mid a^2$, so by Proposition 6.5 we deduce that $2 \mid a$. Thus we may write $a = 2A$ for $A \in \mathbb{Z}$; making that substitution we get

$$2b^2 = a^2 = (2A)^2 = 4A^2,$$

which simplifies to

$$b^2 = 2A^2.$$

This shows that $2 \mid b^2$, so $2 \mid b$. But we assumed that a and b are not both even, so this is a contradiction! Thus there is no $x \in \mathbb{Q}$ such that $x^2 = 2$. □

In this proof (due to a mathematician from the Pythagorean school, an ancient Greek society of thinkers and mystics) the key is that we get to assume *both* that $x \in \mathbb{Q}$ and that $x^2 = 2$. Notice that a direct proof would involve just assuming that $x \in \mathbb{Q}$ and trying to deduce that $x^2 \neq 2$, which seems prohibitively difficult to do (directly). A proof by contrapositive would take a real number x such that $x^2 = 2$ and then show that x is not rational. If we do not wish to proceed by assuming that x is rational (in which case we are doing our proof by contradiction), then again it seems hopeless to proceed directly.

After a bit of reflection on this proof we see that the key fact that makes it work is Proposition 6.5: for all $x \in \mathbb{Z}$, if $2 \mid x^2$ then $2 \mid x$. In Exercise 6.7 you are asked to show that if for a positive integer N we have that for all $x \in \mathbb{Z}$, $N \mid x^2 \implies N \mid x$, then the above argument adapts to show that $\sqrt{N}$ is irrational. The question is which $N \in \mathbb{Z}^+$ have this property. By Proposition 6.6 we know that 3 has this property, so $\sqrt{3}$ is irrational. In turn $2^2 = 4$ so $\sqrt{4}$ is *not* irrational, so 4 cannot have this property. Indeed we have that $4 \mid 2^2$ but $4 \nmid 2$. It takes some doing to determine exactly which positive integers N have this property. In parts a) and b) of Exercise 6.8 you are asked to show that a positive integer N has this property if and only if it is *squarefree* (not divisible by the square of any prime number). Even if N fails to have this property it may still be that $\sqrt{N}$ is irrational. For instance, 12 does not have this property: $12 \mid 6^2$ but $12 \nmid 6$. However, since $(2\sqrt{3})^2 = 12$, if $2\sqrt{3} = \frac{a}{b}$ were rational, then $\sqrt{3} = \frac{a}{2b}$ would also be rational. In Exercise 6.8c) you are asked to show that for $N \in \mathbb{Z}^+$, we have that $\sqrt{N}$ is rational if and only if N is a square: i.e., $N = M^2$ for some integer M. These are some of the more challenging exercises in this text.

THEOREM 6.9 (Euclid). *There are infinitely many prime numbers.*

PROOF. Seeking a contradiction, we suppose that there are only finitely many prime numbers. This means either that there are no prime numbers at all or there are precisely n prime numbers $p_1, \ldots, p_n$ for some $n \in \mathbb{Z}^+$. The first alternative is

ruled out by the observation that 2 is a prime number.

So suppose that $p_1, \ldots, p_n$ are all the prime numbers, and consider

$$N = (p_1 \cdots p_n) + 1.$$

We have $N = (p_1 \cdots p_n) + 1 \geq p_1 + 1 \geq 2 + 1 = 3$, so by Proposition the integer N has at least one prime divisor. However, for all $1 \leq i \leq n$ we have $p_i \nmid N$: indeed, N is of the form $A_i p_i + 1$ so leaves remainder of 1 when divided by p_i. So on the one hand N has a prime divisor, but on the other hand the only primes are $p_1, \ldots, p_n$ and N is not divisible by any of them: contradiction! □

This proof works, but what did we learn about the prime numbers other than that there are infinitely many of them? Nothing – the entire proof took place "in a world where there are only finitely many primes" and then we learn that such a world is contradictory so the proof ends in a puff of logical smoke. This is a shame, since the above argument is very close to giving us an *algorithm* that, when given a finite list of primes, returns a new prime that is not on our list, and that's better.

Let's try this instead: again, we know that there is at least one prime, say 2. So to prove Theorem 6.9 it is enough to show that given any list $p_1, \ldots, p_n$ of distinct primes, we can produce another prime q that is not already on our list. This can be done as follows: consider

$$N = (p_1 \cdots p_n) + 1.$$

Just as above, N is at least 3 and it is not divisible by any of $p_1, \ldots, p_n$. So by Proposition 4.2 it is divisible by some *new* prime number q, and thus $p_1, \ldots, p_n, q$ is a list of $n + 1$ distinct primes.

If we iterate this construction, then we will produce an infinite sequence of prime numbers. Let us actually see some of the terms of such a sequence. For this, we have some choices to make. First of all we must choose a "seed," i.e., a finite nonempty list $p_1, \ldots, p_n$ of primes (it is not actually necessary for these primes to be distinct, but for definiteness let us assume this is the case). We may as well take the simplest choice:

$$p_1 := 2.$$

Next, when we form $N = p_1 \cdots p_n + 1$, we know that N has at least one prime factor and that any prime factor gives a new prime not already on our list, but in general N may have more than one prime factor, so we have to pecify which factor we are appending to our list (we could append more than one factor to our list if we wanted to). For definiteness, and to give the best possible chance of building small primes, let us agree that we will always add to our list the smallest prime divisor of N (and no others).

Now we can see the sequence in action:

$$p_1 := 1,$$

so

$$N_1 := p_1 + 1 = 3.$$

This time N_1 is prime, so we put

$$p_2 := 3$$

and

$$N_2 := 2 \cdot 3 + 1 = 7.$$

Again N_2 is prime, so we put

$$p_3 := 7$$

and

$$N_3 := 2 \cdot 3 \cdot 7 + 1 = 43.$$

Once again N_3 is prime, so we put

$$p_4 := 43$$

and

$$N_4 := 2 \cdot 3 \cdot 7 \cdot 43 + 1 = 1807.$$

But this time N_4 is not prime: indeed $1807 = 13 \cdot 139$. This is an important example, since it shows us that "the N" in Euclid's proof of Euclid's Theorem need not be prime. Many students seem to be under the impression that N is necessarily prime. The first proof certainly *does not* say that, but rather it ends before the question can be grappled with, so it is not helpful in giving the *right* idea.

So we have an infinite sequence of distinct primes. It begins as follows:

$$2, 3, 7, 43, 13, 53, 5, 6221671, 38709183810571, 139, 2801, 11, 17, 5471 \ldots$$

As of now 51 terms of the sequence are known. Finding the $n+1$st term of the sequence given the first n terms involves factoring larger and larger numbers, so this quickly gets difficult. To find the 52nd term would involve finding the least prime divisor of a 335 digit number, and this is beyond current computational reach.

This sequence is called the **Euclid-Mullin** sequence after Euclid and the American engineer and mathematician Albert A. Mullin,[2] who studied it in 1963 [**Mu63**]. One natural question is whether every prime appears eventually in this sequence. This is wide open. The smallest prime that does not appear in the 51 known terms of the sequence is 41.

4.3. A Warning About Proofs by Contradiction. We have already mentioned a drawback about proofs by contradiction: since they proceed from a false assumption, the *only* conclusion that one can draw from the argument is that the initial assumption was false. This is of course all we need for the proof, but many "positive" proofs can establish other useful things *en route*.

There is another possible drawback that we only barely wish to mention. Namely, there is a school of mathematicians and logicians that do not like proving that P is true by proving that $\neg P$ is false; they reject the "law of the excluded middle," i.e., the tautology $P \vee (\neg P)$. Thus they do not accept proofs by contradiction as logically valid. But such schools of thought – of which there are several, going under names like **intuitionism** and **constructivism** – are at odds with not just proofs by contradiction but the entire edifice of Boolean logic on which standard mathematics is founded.[3] These matters have been studied over the years, and it turns out that it is not just matter of finding better proofs: with weaker or different

[2] `https://en.wikipedia.org/wiki/Albert_A._Mullin`

[3] The article `https://en.wikipedia.org/wiki/Constructivism_(philosophy_of_mathematics)` gives a good introduction to constructivism.

logical foundations, many of the standard theorems of mathematicas become false. Some can be repaired via rephrasing, and some cannot. So all we can do is reiterate that we our building our foundations on standard Boolean logic, as is used by the overwhelming majority of working mathematicians (and the others explicitly identify their alternate foundations).

Now we come to our real warning about proofs by contradiction: **they are significantly more prone to error than any other proof**. Suppose that in a proof you make an error of calculation: say an algebra mistake, or you write $<$ when it should be $>$. First of all, it is more likely that this will impede your progress than falsely help you. E.g. if you are trying to factor a polynomial expression in a certain way, then if you have the wrong expression it probably won't factor at all. If you are trying to compute $\lim_{x\to 3} 4x^2 + 1$ directly from the (ϵ, δ) definition and you think the limit is 35, then it just won't work: you won't be able to factor

$$|f(x) - 35| = |4x^2 - 34| = 2|2x^2 - 17|$$

the way you can factor

$$|f(x) - 37| = |4x^2 - 36| = 4|x^2 - 9| = 4|x + 3||x - 9|.$$

If the error does help you – or if you make some unwarranted logical leap – then in a "positive proof" at the end you have a chain of reasoning that you can examine one by one for any weak links. If you do four routine steps where it seems that little progress is made and then one giant step forward that had a short argument, you had really better go back and look carefully at that giant step.

In a proof by contradiction, any mistake – no matter how small – should lead to a contradiction, ending the proof. So instead of having your progress impeded, you may falsely believe that you've won. That's much worse!

Because a succesful proof by contradiction ends in a logical contradiction, there is an inherent "weirdness" to it. In our proof of the irrationality of $\sqrt{2}$, the contradiction comes from the apparently innocuous assumption that we have written a fraction in lowest terms. Isn't that a weird contradiction to reach? Well, yes, a bit – and this argument is correct.

Certain types of contradictions that students derive seem more likely to be derived in error than to be genuine contradictions. For instance if you derive $2x = 0$, deduce that $x = 0$ and derive a contradiction because you know that your x is not zero, then I am already nervous, because

$$x + x = 2x = 0 = x - x,$$

and I wonder if you have just made the sign mistake of replacing $x + x$ by $x - x$. Or if you reach a contradiction like $3 = 5$ after doing some algebra, be extra careful in checking that algebra.

So proof by contradiction is a very powerful technique, but it is not for the faint of heart. You should reach for direct proof and proof by contrapositive first, and then go slowly and carefully if you do attempt a proof by contradiction. Conversely, if you are reading someone else's proof by contradiction, the first thing to ask is if it's really by contradiction. If e.g. they are really arguing by contrapositive but just

don't want to say that word, you will understand their argument more clearly by viewing it that way.

EXAMPLE 6.10. *Walter Rudin's* Principles of Mathematical Analysis [**R**] *is one of the most famous and widely read of all higher mathematics textbooks.*[4] *Consider Theorem 2.37 therein:*

"If E is an infinite subset of a compact set K, then E has a limit point in K." (This asserts that a compact metric space is limit point compact: cf. [**Cl-GT**, Thm. 2.78]*.) The proof begins "If no point of K were a limit point of $E \ldots$" Note the subjunctive mood. It ends "This contradicts the compactness of K." So it is a proof by contradiction. However, a little thought shows that the argument is a "positive proof" of the statement "If K has an infinite subset E without a limit point in K, then K is not compact," which is the contrapositive of the statement of Theorem 2.37. Here it seems that Rudin just did not want to say the word* contrapositive.

5. Without Loss of Generality

In this section we discuss the *least* powerful proof technique of this text, an argument that is usually called "without loss of generality." In principle, a without loss of generality argument does not allow us to complete any proof that we did not otherwise know how to complete. Rather, the point of such arguments is that often proofs contain arguments that are divided up into several cases, and "without loss of generality" is an argument that reduces the number of cases considered by explaining why some of the cases can be deduced from other cases. (Above we said "in principle." In practice, we are limited by our patience and by time and space as to how many cases we can consider individually. If we use a computer, this limit may increase considerably but still exists. If a "without loss of generality" argument allows us to reduce the number of cases from above the threshold that can practically be done to below this threshold, then indeed it may allow us to complete a proof that we otherwise could not.)

5.1. Exploiting Symmetry. Here is a basic idea: let $P(x, y)$ be an open sentence with domain $(x, y) \in S \times S$ for some nonempty set S. Suppose that moreover for all $x, y \in S$ we have $P(x, y)$ and $P(y, x)$ are logically equivalent: we say that the sentence $P(x, y)$ is **symmetric** in x and y. Then the truth locus of P is a symmetric subset of $S \times S$: that is, for all $x, y \in S \times S$, we have that $P(x, y)$ is true if and only if $P(y, x)$ is true.

EXAMPLE 6.11. *Suppose that $P(x, y)$ is a symmetric open sentence with domain $\mathbb{R} \times \mathbb{R}$ and we wish to prove that $P(x, y)$ holds for all $x, y \in \mathbb{R}$. Then it is enough to prove that $P(x, y)$ holds for all $x \leq y$, because if $x < y$ then $P(x, y)$ holds if and only if $P(y, x)$ holds, and $y \leq x$.*

EXAMPLE 6.12. *Let $P > 0$. Among all rectangles with perimeter P, we are interested in the side lengths x and y that yield the maximum area. That is, we wish to maximize xy subject to the constraint $2x + 2y = P$.*

Suppose that we somehow know that this maximum exists and is attained for a unique ordered pair (x, y). Then we must have $x = y$ and thus $x = y = \frac{P}{2}$ and $xy = \frac{P^2}{4}$. Do you see why? It is because both xy and $2x+2y$ are symmetric in x and

[4]Which is not to say it's easy: I bought my copy when I was 18, and I was almost 30 before I felt I could read and understand it with only modest effort.

y, and therefore the set of values at which the maximum occurs must be symmetric under $(x, y) \mapsto (y, x)$. We are assuming moreover that this set of values consists of exactly one value (x, y), so we have $(x, y) = (y, x)$ and thus $x = y$. Since also $2x + 2y = P$, we get $x = y = \frac{P}{2}$.

The catch here is that to actually solve the maximization problem we need to show *that the maximum exists and is attained for a unique ordered pair. It is not difficult to do so in this case (even without calculus), but I think we need to start over: if $2x + 2y = P$, then $y = \frac{P}{2} - x$, so we are maximizing*

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

The last expression is at most $\frac{P^2}{4}$, with equality if and only if $x = \frac{P}{2}$.

EXAMPLE 6.13. *Let's show:*

$$\forall\ x, y \in \mathbb{Z},\ 3 \mid x^2 + y^2 \iff (3 \mid x \text{ and } 3 \mid y).$$

One direction is easy: if $3 \mid x$ and $3 \mid y$ then since $x \mid x^2$ and $y \mid y^2$ we have $3 \mid x^2$ and $3 \mid y^2$ and thus finally $3 \mid x^2 + y^2$.

To prove the other direction we will use the contrapositive: suppose that it is not *the case that $3 \mid x$ and $3 \mid y$; we will show that $3 \nmid x^2 + y^2$. Our general strategy is to divide into cases according to the remainders of x and y upon division by 3. This is nine cases overall, and the only one that is excluded is when x and y are both divisible by 3, so apparently we have eight cases left. That's a lot: let's try to reduce the number. This can be done using that the statement in question is symmetric in x and y, so whenever we can get between two different cases by swapping x and y we only have to do one of the cases. Let's see how this works out:*
Case 1: $x = 3X$ and $y = 3Y + 1$ swaps with Case 2: $x = 3X + 1$ and $y = 3Y$.
Case 3: $x = 3X$ and $y = 3Y + 2$ swaps with Case 4: $x = 3X + 2$ and $y = 3Y$.
Case 5: $x = 3X + 1$ and $y = 3Y + 1$ "swaps with itself".
Case 6: $x = 3X+1$ and $y = 3Y+2$ swaps with Case 7: $x = 3X+2$ and $y = 3Y+1$.
Case 8: $x = 3X + 2$ and $y = 3Y + 2$ "swaps with itself."

So instead of doing all eight cases we only have to do five of them: this is enough savings to be worth doing.
Case 1: We have $x^2 + y^2 = (3X)^2 + (3Y + 1)^2 = 3(3X^2 + 3Y^2 + 2Y) + 1$ is not divisible by 3.
Case 3: We have $x^2 + y^2 = (3X)^2 + (3Y + 2)^2 = 3(3X^2 + 3Y^2 + 4Y + 1) + 1$ is not divisible by 3.
Case 5: We have $x^2 + y^2 = (3X + 1)^2 + (3Y + 1)^2 = 3(3X^2 + 2X + 3Y^2 + 3Y) + 2$ is not divisible by 3.
Case 6: We have $x^2+y^2 = (3X+1)^2+(3Y+2)^2 = 3(3X^2+2X+3Y^2+4Y+1)+2$ is not divisible by 3.
Case 8: We have $x^2+y^2 = (3X+2)^2+(3Y+2)^2 = 3(3X^2+4X+3Y^2+4Y+2)+2$ is not divisible by 3.

5.2. Changing Labels / Colors. We want to discuss a slightly different kind of "without loss of generality" argument. It still exploits symmetry, just not a symmetry that comes precisely from switching variables x and y in the statement we are trying to prove. We introduce it by giving two proofs of the following result.

PROPOSITION 6.14. *If six people are at a party, then there is some trio of them such that either: (i) every person in the trio knows the other two people in the trio, or (ii) no person in the trio knows either of the other two people in the trio.*

PROOF. Label the people P_1 through P_6. Consider the set S of people that person P_1 knows. If any two people in S know each other, then those two people together with person P_1 form a trio all of whom know each other, so we may assume that no two people in S know each other. If $\#S \geq 3$ then we're done because we have three or more people no two of which know each other, so we may assume that $\#S \leq 2$. Since there are six people all together, there must be a trio of people that are disjoint from $P_1 \cup S$. If all of these people know each other, then we're done. If any two of the members of this trio don't know each other, then together with P_1 they form a trio none of whose members know each other. □

Here is a SECOND PROOF: we draw six dots in the plane P_1 through P_6. Between each pair of distinct dots we either use a blue marker to draw a line between them or a red marker to draw a line between them and not both. (It's okay if the lines cross each other!) Then we want to show that there is a **monochromatic triangle**: i.e., three dots all connected by lines of the same color. Now stop and think that this is equivalent to Proposition 6.14: we may model the party situation by connecting two people who know each other by a red line and two people who don't know each other by a blue line. Now here's the new proof:

Thre are five lines coming out of P_1, so either at least three of them are red or at least three of them are blue. However, because we do not disturb the truth of the statement by changing the colors of all red lines to blue and conversely, we may as well assume that P_1 has at least three red lines connecting it to other dots; let this set of dots be called S. If any two of the dots in S, say P_i and P_j, are connected by a red line, then P_1, P_i, P_j is a red triangle. Otherwise all the lines connecting the dots in S are blue, and since $\#S \geq 3$, we get (at least) a blue triangle.

The second proof is a bit shorter than the first, for an interesting reason: in the first proof we had to consider the case that P_1 knew at least 3 people separately from the case in which P_1 knew at most 2 people, but once we reinterpret knowing vs. not knowing as "colors" and that switching all the colors is okay, we only need to consider the first case.

In Exercise 6.14 you are asked to show that for five people at a party, it is possible that for every trio of them some two of the three know each other but all three do not know each other. In Section §9.3.3 we will introduce an entire branch of mathematics – Ramsey Theory – that springs from Proposition 6.14.

We give another classic example in which colors are explicitly present. For a set X, a **two-coloring of X** is a function $\mathfrak{c} : X \to \{\text{red}, \text{blue}\}$. That is, to every element x in X we assign either the color red or the color blue...and not both! In our application, X will be the set $\mathbb{Z}^+$ of positive integers. So to give a 2-coloring of $\mathbb{Z}^+$ means to start at 1, color it either red or blue, move on to 2 and color it either read or blue, and so forth, down the number line. (Evidently there are are a lot of 2-colorings of $\mathbb{Z}^+$: if we wanted, we could identify them with base 2 expansions of real numbers in the interval $[0, 1]$.)

Next we define a **Schur triple** to be an ordered triple (x, y, z) of positive integers such that $z = x + y$. This is not the toughest definition in the text: e.g. $(1, 1, 2)$ and $(2, 3, 5)$ are Schur triples, while $(3, 3, 3)$ is not. If we are given a 2-coloring $\mathfrak{c}$ of $\mathbb{Z}^+$, a Schur triple $(a, b, c) \in (\mathbb{Z}^+)^3$ is **monochromatic** if $\mathfrak{c}(a) = \mathfrak{c}(b) = \mathfrak{c}(c)$, or in other words if a, b, c are either all colored red or all colored blue.

We will now investigate the following question: what is the smallest $N \in \mathbb{Z}^+$ such that for every 2-coloring $\mathfrak{c}$ of $\mathbb{Z}^+$, we have a monochromatic Schur triple (a, b, c) with $c \leq N$? (Notice that if $c \leq N$, then since $a, b, c \geq 0$, if $a + b = c$ then also $a, b \leq c \leq N$.)

a, b, c are positive integers with $c = a + b$, then $c = a + b \geq 1 + 1 \geq 2$, so we need $N \geq 2$.

• $N = 2$: The unique Schur triple with $c = 2$ is $(1, 1, 2)$. If we color 1 red and 2 blue, this is not monochromatic. So we move on.

• $N = 3$: The two Schur triples with $c = 3$ are $(1, 2, 3)$ and $(2, 1, 3)$. Notice that if (a, b, c) is a monochromatic Schur triple with $c \leq N$ then also (b, a, c) is a monochromatic Schur triple with $c \leq N$. This means that in order to investigate whether a given 2-coloring of $\mathbb{Z}^+$ has a monochromatic Schur triple with $c \leq N$, then *without loss of generality* we need only consider Schur triples with $a \leq b \leq c$. With $N = 3$, the unique such Schur triple is $(1, 2, 3)$. But again, if we color 1 red and color 2 blue, then *no matter how we color* 3 neither $(1, 1, 2)$ nor $(1, 2, 3)$ is monochromatic.

• $N = 4$: The new triples to look at are $(1, 3, 4)$ and $(2, 2, 4)$. Sticking with coloring 1 red and 2 blue, if we color 4 red then $(2, 2, 4)$ is not monochromatic, and then if we color 3 blue then $(1, 3, 4)$ is not monochromatic.

• $N = 5$: I claim however that for every 2-coloring $\mathfrak{c}$ of $\mathbb{Z}^+$ there is a monochomratic Schur triple (a, b, c) with $a \leq b \leq c \leq 5$. To see this, observe that for any 2-coloring $\mathfrak{c} : \mathbb{Z}^+$ there is another 2-coloring $\overline{\mathfrak{c}}$ in which we flip all the colors of $\mathfrak{c}$: that is, for all $n \in \mathbb{Z}^+$, $\overline{\mathfrak{c}}$ is red if and only if $\mathfrak{c}$ is blue. Under passage from $\mathfrak{c}$ to $\overline{\mathfrak{c}}$, monochromatic Schur triples (a, b, c) with $a \leq b \leq c \leq N$ are preserved. So by switching from $\mathfrak{c}$ to $\overline{\mathfrak{c}}$ if necessary, *without loss of generality* we may assume that 1 is colored red (as we did above). Then if 2 is also colored red, then $(1, 1, 2)$ is a red Schur triple with $2 \leq 5$ and we're done, so we may assume that 2 is colored blue. If 4 is also colored blue, then $(2, 2, 4)$ is a blue Schur triple with $4 \leq 5$, so we may assume that 4 is colored red. If 5 is also colored red, then $(1, 4, 5)$ is a red Schur triple with $5 \leq 5$, so we may assume that 5 is colored blue. And now for the dramatic conclusion: what color is 3? If it's red, then $(1, 3, 4)$ is a red Schur triple with $4 \leq 5$, while if it's blue then $(2, 3, 5)$ is a blue Schur triple with $5 \leq 5$. We conclude:

PROPOSITION 6.15. *For any 2-coloring of $\mathbb{Z}^+$ there is a monochromatic Schur triple (a, b, c) with $a \leq b \leq c \leq 5$.*

Just like Proposition 6.14, Proposition 6.15 is a special case of a more general result, **Schur's Theorem**, that we will discuss in §9.3.4.

5.3. Debriefing. It is somewhat ironic that while a "without loss of generality" argument is an *a priori* claim that certain cases in a proof can be reduced to others, it is challenging to give an *a priori* description of exactly what sorts of arguments will effect this kind of case reduction. In the last two sections we considered two kinds of such arguments, both involving the exploitation of a kind of symmetry.

6. Equivalences

If the most common logical form of a mathematical theorem is

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

then the second most common form is probably

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

Equivalences $P \iff Q$ are important in mathematics, because they allow us to *freely exchange* P for Q in all arguments. Moreover in modern mathematics key definitions and concepts often come in several equivalent forms, to the extent that it is often much more important to know the equivalence(s) than to know any particular definition.

6.1. Proving Equivalences by Reversible Arguments. As we saw in Propostion 2.9, the equivalence $P \iff Q$ is itself logically equivalent to the conjuction $(P \implies Q) \wedge (Q \implies P)$, and indeed the most common way to prove $P \iff Q$ is to separately prove $P \implies Q$ and $Q \implies P$. The alternative is to make an argument, each step of which is manifestly "logically reversible."

EXAMPLE 6.16. *For all* $x \in \mathbb{R}$, $x = 17$ *iff* $3x - 6 = 45$. *Indeed,*

$$x = 17 \iff 3x = 51 \iff 3x - 6 = 51 - 6 = 45.$$

This works because of the of the following two observations:
(i) For all $x, y, a \in \mathbb{R}$ *we have* $x = y \iff x + a = y + a$.

That is, two real numbers are equal if and only the numbers obtained by adding any number a *to both of them are equal. The reason for this is that every real number has an additive inverse, and the* reverse *of adding* a *is adding the additive inverse* $-a$ *of* a. *Indeed, here* $\mathbb{R}$ *can be replaced by any number system satisfying the field axioms.*
(ii) For all $x, y \in \mathbb{R}$ *and* $a \in \mathbb{R} \setminus \{0\}$, *we have* $x = y \iff ax = ay$.
That is, two real numbers are equal if and only if the numbers obtained by multiplying them both by any number a *are equal. The reason for this is because every* nonzero *real number has a multiplicative inverse, and the* reverse *of multiplying by* $a \neq 0$ *is multiplying by the reciprocal* a^{-1} *of* a. *Again this holds in any number system satisfying the field axioms.*

On the other hand it is not *true that* $x = 17$ *iff* $0x - 6 = -6$. *In fact for any* $x \in \mathbb{R}$ *we have* $0x - 6 = -6$. *This is because multiplying by* 0 *is not reversible: if* $x = y$, *then* $0 \cdot x = 0 \cdot y$, *but of course since* $0 \cdot x = 0 = 0 \cdot y$, *that* $0 \cdot x = 0 \cdot y$ *is true whether* x *and* y *were equal or not.*

EXAMPLE 6.17.

a) *For $x, y \in \mathbb{R}$, if $x = y$ then $x^2 = y^2$. Is the converse true? No – e.g. $-1 \neq 1$ but $(-1)^2 = 1 = 1^2$. However for x, y in any number system F satisfying the field axioms, if $x^2 = y^2$ then*

$$0 = x^2 - y^2 = (x + y)(x - y),$$

so by Proposition 4.2 we have $x + y = 0$ or $x - y = 0$, that is $x = \pm y$ (we understand "$x = \pm y$" to be an abbreviation for "$x = y$ or $x = -y$"). That is one kind of fix. The following is also often useful: for $x, y \in [0, \infty)$ if $x = -y$ then $x = y = 0$, so we get:

$$\forall x, y \in [0, \infty),\ x = y \iff x^2 = y^2.$$

In fact this holds in any number system satisfying the ordered field axioms.

b) *For $x, y \in \mathbb{R}$, I claim that we have $x = y \iff x^3 = y^3$.*

Namely, let $x, y \in \mathbb{R}$ and suppose $x^3 = y^3$. Then

$$0 = x^3 - y^3 = (x - y)(x^2 + xy + y^2).$$

Seeking a contradiction, we suppose that $x \neq y$. Then $x - y \neq 0$, so we can divide through by it, getting

$$0 = x^2 + xy + y^2 = \left(x + \frac{y}{2}\right)^2 + \frac{3}{4}y^2.$$

Because each of the two terms in the sum is non-negative, the only way the sum can be zero is if

$$\left(x + \frac{y}{2}\right)^2 = \frac{3}{4}y^2 = 0.$$

The second equality implies $y = 0$, and plugging this into the first inequality gives $x^2 = 0$ and thus $x = 0$. So $x = 0 = y$: contradiction.

6.2. Proving that N Statements are Equivalent, for $N \geq 3$. Often in mathematics a result has the form "The following are equivalent:" followed by a finite list of $N \geq 2$ different statements. When $N = 2$ this is the usual $P \iff Q$ we have already discussed, so we are now interested in the case $N \geq 3$.

First let us be sure that we understand the meaning of the assertion that the statements $P_1, \ldots, P_N$ (say) are logically equivalent. It means that all of these N statements have the same truth value: either all are true or all are false. Perhaps more usefully, it also means that each of these statements implies all of the others: for all $1 \leq i, j \leq N$ we have $P_i \implies P_j$.

Notice that the above formulation gives N^2 different implications. This is clearly more than necessary, since it includes the N implications $P_i \implies P_i$ for $1 \leq i \leq N$, which are certainly true no matter what the P_i's may be and thus do not require proof. That gives us $N^2 - N$ different implications.

• When $N = 2$ we have $2^2 - 2 = 2$ and we are just saying that to prove $P_1 \iff P_2$ it suffices to prove $P_1 \implies P_2$ and $P_2 \implies P_1$.

• When $N = 3$ we have $3^2 - 3 = 6$ and we are proposing to show that P_1, P_2, P_3 are equivalent by proving

(22) $P_1 \implies P_2$, $P_1 \implies P_3$, $P_2 \implies P_1$, $P_2 \implies P_3$, $P_3 \implies P_1$, $P_3 \implies P_2$.

Well, this will certainly suffice...but it is clearly too much work. We are not taking into account the **transitivity of implication** (Exercise 2.7). Thus e.g. from the first five implications of (22) we know that $P_3 \implies P_1$ and $P_1 \implies P_2$, so we deduce that $P_3 \implies P_2$. So we can prove that equivalence using five implications.

But in fact we can prove that equivalence usng just four implications: $P_1 \implies P_2$, $P_2 \implies P_1$, $P_2 \implies P_3$ and $P_3 \implies P_2$. Otherwise put, it suffices to show

$$P_1 \iff P_2 \iff P_3$$

and then break each $\iff$ into $\implies$ and $\impliedby$. Evidently this strategy will work for any $N \geq 2$: we can show

$$P_1 \iff P_2 \iff \ldots \iff P_{N-1} \iff P_N,$$

which is $N-1$ equivalences and thus $2N-2$ implications.

But *in fact* we can do better yet: it suffices to show

$$P_1 \implies P_2 \implies P_3 \implies P_1.$$

We visualize this by putting P_1, P_2 and P_3 as points on a circle: starting at any point and travelling counterclockwise via our implications, we get to the other three points. In general, to prove the equivalence of $P_1, \ldots, P_N$ we can prove

$$P_1 \implies P_2 \implies \ldots \implies P_N \implies P_1.$$

That is, we prove the implications in a cyclic order.

Thus we have successively reduced the number of implications needed to establish the equivalenceof N statements from N^2 to $N^2 - N$ to $2N$ to N. Is this the best we can do? Yes, it is:

PROPOSITION 6.18. *Let $N \geq 2$. In order to prove the equivalence of statements $P_1, \ldots, P_N$ via basic implications of the form $P_i \implies P_j$, we need at least N basic implications.*

PROOF. If we have fewer than N basic implications, then

$$S \coloneqq \{1 \leq i \leq n \mid P_i \text{ appears as a hypothesis of one of our basic implications}\}$$

has size less than N, which means there is at least one i such that P_i does not appear as a hypothesis in any of our basic implications. Suppose then that P_i is true and that all the other statements P_j are false. Then in every basic implication the hypothesis is false, hence each basic implication is true, but the statements $P_1, \ldots, P_N$ are not equivalent. □

One could try to push this further by asking whether there is any other way to establish the equivalence of N statements via N basic implications besides ordering them in a circle. In fact there is not. An argument for this begins as follows: given N basic implications of the form $P_i \implies P_j$ that suffice to establish the equivalence of P_1 through P_N, consider as in the proof of Proposition 6.18 the set S of indices i of statements P_i that appear a hypothesis of one of these basic implications. The same argument shows that $\#S = N$, so $S = [N]$. Similarly, let T be the set of indices j of statements P_j that appear as a conclusion of one of these basic implications. If there is $j \in [N] \setminus T$, then the statement P_j could be false, all

the other statements could be true, and all the basic implications would hold, so we must have

$$S = T = [N].$$

Thus we get a function $f : [N] \to [N]$ by mapping $i \in [N]$ to the unique $j \in [N]$ such that $P_i \implies P_j$ is one of our basic implications, and this function is surjective from a finite set to itself. We will pause the argument here and come back to it once we have discussed cycle types of bijective maps $f : [N] \to [N]$ in §9.1.

When we try to prove the equivalence of N statements using N basic implications arranged in a circle, we get to choose the ordering of the statements, so "arrange the basic implications in a circle" can be done in many different ways. Arranging the implications in *order* means writing down an irredundant list of length N from $[N]$, so there are $N!$ ways to do this. However, while

$$\ell_1 : 1, 2, 3;\ \ell_2 : 2, 3, 1;\ \ell_3 : 3, 1, 2$$

are three different lists, they yield the same set of implications: proving $P_1 \implies P_2 \implies P_3 \implies P_1$ is the same as proving $P_2 \implies P_3 \implies P_1 \implies P_2$ and the same as proving $P_3 \implies P_1 \implies P_2 \implies P_3$: in all cases we prove $P_1 \implies P_2$, $P_2 \implies P_3$ and $P_3 \implies P_1$. So we shouldn't distinguish among lists that can be obtrained from each other by "rotating" – i.e., repeatedly replacing the list

$$x_1, x_2, \dots, x_N$$

with the list

$$x_2, x_3, \dots, x_N, x_1.$$

Repeated rotation of an irredundant list of length N generates N such lists altogether, so the number of **cyclically inequivalent** lists is $\frac{N!}{N} = (N-1)!$. For $N = 3$ this gives us two choices: essentially we decide whether to prove $P_1 \implies P_2$ or to prove $P_2 \implies P_1$.

In practice all of this *can* matter, since just because N statements turn out to be equivalent does not mean that it is easy to argue *directly* that each one implies each other one. In fact as N grows, proving the equivalence of N statements often becomes quite cumbersome, because one has to try to figure out which basic implications are "doable" and decide how to string together enough of them to establish the equivalence. And the by the way, one may not always be able to make it work in N basic implications, so someimes one needs to prove more. When beginning a proof that $N \geq 4$ statements are equivalent, if you can do it in N basic implications, I suggest that you order the implications so that the proof follows the strategy

$$P_1 \implies P_2 \implies \dots \implies P_N \implies P_1.$$

If the basic implications are any more complicated than that, you should make very clear to the reader what the sequence of implications is going to be. Otherwise the reader will likely be distracted by wondering about this. Here is an example with $N = 6$ from my own writing: [**Cl-CA**, Thm. 17.21]. (Perhaps I could have done better about spelling out the pattern of implications in advance.)

7. Constructive vs. Nonconstructive Proofs

Having discussed the most common logical forms of statements in detail, let us now give a bit of consideration to the existentially quantified statement:

$$\exists\, x \in S,\ P(x). \tag{23}$$

The simplest way to prove (23) is to find a specific $x \in S$ for which $P(x)$ is true and prove it. Here is a nearly trivial example:

PROPOSITION 6.19. *There are prime numbers.*

PROOF. The number 2 is prime. □

This kind of existence proof is called **constructive**. Notice that a proof that there are infinitely many primes is harder to make constructive because it is harder to exhibit infinitely many things than to exhibit one thing, but nevertheless we tweaked the original proof by contradiction to make it more constructive by giving an algorithm that generates arbitrarily many primes.

What would it mean to prove a statement of the form (23) nonconstructively? Here is a classic example.

PROPOSITION 6.20. *There are positive, irrational real numbers x and y such that x^y is rational.*

PROOF. Case 1: Suppose that $\sqrt{2}^{\sqrt{2}} \in \mathbb{Q}$. Then since $\sqrt{2}$ is irrational, we may take $x = y = \sqrt{2}$.
Case 2: Suppose thae $\sqrt{2}^{\sqrt{2}} \notin \mathbb{Q}$. Then we may take $x = \sqrt{2}^{\sqrt{2}}$ and $y = \sqrt{2}$:

$$x^y = (\sqrt{2}^{\sqrt{2}})^{\sqrt{2}} = \sqrt{2}^{\sqrt{2}\sqrt{2}} = \sqrt{2}^2 = 2 \in \mathbb{Q}.$$ □

It is remarkable how nimbly the question of the irrationality $\sqrt{2}^{\sqrt{2}}$ is sidestepped: the proof works either way. But of course, unless we know which of Case 1 or Case 2 is true, we do not know explicit irrational x, y such that $x^y \in \mathbb{Q}$. But in fact the answer to this question is known, which makes the proof constructive. To explain, we need to introduce some terminology.

A real number x is **algebraic** if there are integers $a_0, \ldots, a_n$, not all 0, such that

$$a_n x^n + \ldots + a_1 x + a_0 = 0.$$

EXAMPLE 6.21.

a) *Every rational number is algebraic: if $x = \frac{a}{b}$, then $bx - a = 0$.*
b) *For all $N \in \mathbb{Z}^+$, the number $\sqrt{N}$ is algebraic: $x^2 - N = 0$.*

A real number is **transcendental** if it is not algebraic.

THEOREM 6.22.

a) *(Hermite, 1873) The number e is transcendental.*
b) *(Lindemann, 1882) The number π is transcendental.*

The following theorem answered the seventh of Hilbert's Problems, a list of 23 problems posed by David Hilbert (then the world's leading mathematician) in 1900.

THEOREM 6.23 (Gelfond-Schneider, 1934). *If a, b are positive algebraic numbers with $a \notin \{0, 1\}$ and $b \notin \mathbb{Q}$, then a^b is transcendental.*

Applying Gelfond-Schneider with $a = b = \sqrt{2}$, we find that $\sqrt{2}^{\sqrt{2}}$ is transcendental, hence certainly irrational. Knowing this, the proof of Proposition 6.20 shows that $x = \sqrt{2}^{\sqrt{2}}$ and $y = \sqrt{2}$ are two positive irrational numbers such that $x^y \in \mathbb{Q}$.

But in fact it is possible to give a much more elementary construction of irrational x, y such that $x^y \in \mathbb{Q}$.

PROPOSITION 6.24. *The real number number $\log_2(25)$ is irrational.*

PROOF. Put $y := \log_2(25)$. By definition then we have $2^y = 25$. Seeking a contradiction, we suppose $y = \frac{a}{b}$ with $a, b \in \mathbb{Z}^+$. Then

$$2^{a/b} = 25 = 5^2,$$

so

$$2^a = 5^{2b}.$$

But 2^a is even and 5^{2b} is odd, contradiction. □

Now take $x = \sqrt{2}$ and $y = \log_2(25)$, so x and y are positive irrational numbers and

$$x^y = \sqrt{2}^{\log_2(25)} = \sqrt{2}^{\log_2(5^2)} = \sqrt{2}^{2\log_2(5)} = (\sqrt{2}^2)^{\log_2(5)} = 2^{\log_2(5)} = 5 \in \mathbb{Q}.$$

In Exercise 6.16 you are asked to use the Fundamental Theorem of Arithmetic to show that for any coprime integers $a, b > 1$, the number $\log_a(b)$ is irrational.

Another example of a nonconstructive existence proof takes us back to the beginning of the chapter.

THEOREM 6.25 (Ball-Rivoal [**BR01**]). *There are infinitely many positive integers k such that*

$$\zeta(2k+1) := \sum_{n=1}^{\infty} \frac{1}{n^{2k+1}}$$

is irrational.

What is remarkable here is that while Apéry proved in 1979 that $\zeta(3)$ is irrational, the later work of Ball-Rivoal does not give any further *specific* values of k such that $\zeta(2k+1)$ is irrational, though it shows that infinitely many such k must exist.

8. Exercises

EXERCISE 6.1. *In each part, you will consider a statement of the form*

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

Your job is to decide (i) whether each statement is vacuously true and (ii) whether each statement is trivially true. Note that it is possible that both conditions hold and it is possible that neither of them hold.

a) *For all prime numbers $p \geq 3$, if $p + 1$ is prime, then $2p + 1$ is prime.*
b) *For all integers x, if $x^2 + x + 1$ is even, then $x^2 + x + 2$ is even.*
c) *For all $x, y \in \mathbb{Z}^+$, if $x + y$ is prime, then $\cos(2\pi xy) \in \mathbb{Z}$.*
d) *For all integers $x \geq 2$, if $2^x - 1$ is prime, then x is prime.*

EXERCISE 6.2.

a) *Show: if $a \in \mathbb{Z}$ is odd, then $8 \mid a^2 - 1$.*
b) *Show: for all $x \in \mathbb{Z}$, we have $8 \mid x(x-1)(x-2)(x-3)$.*
(Hint: show that exactly one of x, $x-1$, $x-2$, $x-3$ is divisible by 4. Then show that exactly one of the other three factors is divisible by 2.)

EXERCISE 6.3. *Let $x_1, \ldots, x_n \in \mathbb{R}$ and let $a_1, \ldots, a_n > 0$.*

a) *Show that*
$$a_1 x_1^2 + \ldots + a_n x_n^2 \geq 0.$$
b) *Show that $a_1 x_1^2 + \ldots + a_n x_n^2 = 0$ if and only if $x_1 = \cdots = x_n = 0$.*

EXERCISE 6.4. *Prove each of the following statements* by contradiction.

a) *If $a, b \in \mathbb{R}^{>0}$, then $a + b \geq 2\sqrt{ab}$.*
b) *For all $n \in \mathbb{Z}$, $4 \nmid (n^2 + 2)$.*
c) *For all $a, b \in \mathbb{Z}$, we have $a^2 + 4b + 5 \neq 0$.*

EXERCISE 6.5. *Go back to Exercise 6.4 and prove each of the statements* directly. *Then in each case comment on which proof you prefer.*

EXERCISE 6.6. *Show: there are no $x, y \in \mathbb{Z}^+$ such that $x^2 + x + 1 = y^2$.*
(Hint: what is the smallest square integer that is larger than x^2?)

EXERCISE 6.7.

a) *Let $N \in \mathbb{Z}+$. Suppose that the following statement holds:*
$$P(N): \ \forall x \in \mathbb{Z}, \ N \mid x^2 \implies N \mid x.$$
Adapt the proof of Theorem 6.8 to show that $\sqrt{N}$ is irrational.
b) *Show: $\sqrt{3}$ is irrational. (Hint: Combine part a) and Proposition 6.6.)*
c) *Show that $P(4)$ is false.*
d) *Show that $P(5)$, $P(6)$ and $P(7)$ are true, and deduce that $\sqrt{5}$, $\sqrt{6}$ and $\sqrt{7}$ are irrational.*

EXERCISE 6.8. *A positive integer N is* ***squarefree*** *if for no prime number p do we have $p^2 \mid N$.*

a) *Show: if $P(N)$ (cf. Exercise 6.7a)) holds, then N is squarefree.*
(Suggestion: Prove the contrapositive. If N is not squarefree, write $N = p^2 M$ for some $M \in \mathbb{Z}^+$. Show that for $x = pM$ we have $x^2 \mid N$ but $x \nmid N$.)
b) *Show: if N is squarefree, then $P(N)$ holds.*
(Suggestion: use the Fundamental Theorem of Arithmetic. In particular, show that N is squarefree if and only it is a product of distinct prime numbers.)
c) *Deduce: if $N \in \mathbb{Z}^+$ is squarefree, then $\sqrt{N}$ is irrational.*
d) *Show: For $N \in \mathbb{Z}$ we have that $\sqrt{N}$ rational if and only if $N = M^2$ for some $M \in \mathbb{Z}$.*

EXERCISE 6.9. *Let $x, y \in \mathbb{R}$.*

a) *Show: if $x, y \in \mathbb{Q}$ then $x + y, x - y, xy \in \mathbb{Q}$. If $y \neq 0$, show $\frac{x}{y} \in \mathbb{Q}$.*
b) *Suppose $x \in \mathbb{Q}$ and $y \in \mathbb{R} \setminus \mathbb{Q}$. Show: $x + y, x - y \in \mathbb{R} \setminus \mathbb{Q}$.*
c) *Suppose $x \in \mathbb{Q} \setminus \{0\}$ and $y \in \mathbb{R} \setminus \mathbb{Q}$. Show: $xy, \frac{y}{x}, \frac{x}{y} \in \mathbb{R} \setminus \mathbb{Q}$.*
d) *Find $x, y \in \mathbb{R} \setminus \mathbb{Q}$ such that $x + y \in \mathbb{R} \setminus \mathbb{Q}$.*
e) *Find $x, y \in \mathbb{R} \setminus \mathbb{Q}$ such that $x + y \in \mathbb{Q}$.*

f) *Find $x, y \in \mathbb{R} \setminus \mathbb{Q}$ such that $xy \in \mathbb{R} \setminus \mathbb{Q}$.*
g) *Find $x, y \in \mathbb{R} \setminus \mathbb{Q}$ such that $xy \in \mathbb{Q}$.*

EXERCISE 6.10. *Let $x, y \in \mathbb{R}$ be such that $x^3 = y^3$.*[5]
a) *Show that either $x = y$ or $x^2 + xy + y^2 = 0$. (Hint: factor $x^3 - y^3$.)*
b) *Show: if $y = 0$ and $x^2 + xy + y^2 = 0$, then $x = 0$.*
c) *Suppose $x^2 + xy + y^2 = 0$ and $y \neq 0$. Show that*

$$\left(\frac{x}{y}\right)^2 + \frac{x}{y} + 1 = 0.$$

d) *Let $X \in \mathbb{R}$. Show that*

$$X^2 + X + 1 = \left(X + \frac{1}{2}\right)^2 + \frac{3}{4}$$

and deduce that there are no real numbers a such that $X^2 + X + 1 = 0$.
e) *Show: $x = y$.*

EXERCISE 6.11. *Let $a, b, c \in \mathbb{R}$.*[6] *Suppose*

$$\Delta := b^2 - 4ac < 0.$$

Show that: for all $x, y \in \mathbb{R}$, if $ax^2 + bxy + cy^2 = 0$, then $x = y = 0$.
(Suggestion: as in the previous exercise, reduce to showing that there are no real numbers X such that $aX^2 + bX + c = 0$, and show this by completing the square.)

EXERCISE 6.12. *Let S be a finite nonempty set of size N, and let $P(x, y)$ be a symmetric statement with domain $(x, y) \in S \times S$ (recall that this means that for all $(x, y) \in S \times S$ we have $P(x, y)$ is logically equivalent to $P(y, x)$). We may list the elements of S as $s_1, \ldots, s_N$. Show: if $P(x, y)$ holds for all $x = s_i$ and $y = s_j$ with $i \leq j$ then it holds for all $x, y \in S$, and that the number of ordered pairs (s_i, s_j) with $1 \leq i \leq j \leq N$ is $\frac{N^2+N}{2}$.*

EXERCISE 6.13. *State and prove an analogue of the previous exercise for symmetric statements of the form $P(x, y, z)$ with domain $(x, y, z) \in S \times S \times S$, where S is a finite nonempty set with N elements.*
(Hint: the matter of it is to count $\{(i, j, k) \in \mathbb{Z}^3 \mid 1 \leq j \leq j \leq k \leq N\}$.)

EXERCISE 6.14. *Show that it is possible to have five people at a party such that for every trio of them some two of the three know each other but all three do not know each other.*

EXERCISE 6.15. *Determine, with proof, which of the following numbers are rational and which are irrational.*
a) $\log_2(10)$.
b) $\log_3(16)$.
c) $\log_4(8)$.

EXERCISE 6.16. *Let $a, b \geq 2$ be coprime integers. Show: $\log_a(b)$ is irrational.*
(You will want to use Euclid's Lemma or the Fundamental Theorem of Arithmetic.)

[5]Throughout this exercise, in place of $\mathbb{R}$ we could work in any ordered field.
[6]In place of $\mathbb{R}$ we could work in any ordered field.

CHAPTER 7

Induction

1. Inductive Subsets

We define an **inductive subset** $S \subseteq \mathbb{N}$ to be a set of non-negative integers satisfying both of the following:
(IS1) We have $0 \in S$.
(IS2) For all $n \in \mathbb{N}$, if $n \in S$, then $n+1 \in S$.

What can we say about an inductive subset $S \subseteq \mathbb{N}$? Well:
Step 1: By (IS1), we have $0 \in S$.
Step 2: By (IS2), since $0 \in S$, also $1 = 0+1 \in S$.
Step 3: By (IS2), since $1 \in S$, also $2 = 1+1 \in S$.
Step 4: By (IS2), since $1 \in S$, also $3 = 1+1 \in S$.

Evidently we can continue this argument, showing that $4 \in S$, then that $5 \in S$, and so forth. What is the final conclusion?

THEOREM 7.1 (Principle of Mathematical Induction for Subsets). *For a subset $S \subseteq \mathbb{N}$, the following are equivalent:*

(i) *The set S is an inductive subset of $\mathbb{N}$.*
(ii) *We have $S = \mathbb{N}$.*

PROOF. (i) $\implies$ (ii): Seeking a contradiction, suppose that $S \subsetneq \mathbb{N}$. Then $T := \mathbb{N} \setminus S$ is a nonempty subset of the well-ordered set $\mathbb{N}$, so it has a minimum, say a. By (IS1) we have $0 \in S$, so $0 \notin \mathbb{N} \setminus S = T$. Thus $a \geq 1$, so $a = b+1$ for some integer $b \geq 0$. Since a is the minimal element of T and b is a non-negative integer less than a, we must have $b \in S$, but then by (IS2) we have $b+1 = a \in S$. So $a \notin T$, a contradiction.
(ii) $\implies$ (i): Certainly $0 \in \mathbb{N}$, and since for all $n \in \mathbb{N}$ both n and $n+1$ lie in $\mathbb{N}$, the implication $n \in \mathbb{N} \implies n+1 \in \mathbb{N}$ is true. □

The proof method of (i) $\implies$ (ii) in Theorem 7.1 is sometimes called the method of **minimal counterexample**. If we have a collection of statements $\{P(x)\}_{x \in S}$ indexed by a well-ordered subset S of $\mathbb{R}$ (above we took $S = \mathbb{N}$), then if they are not all true, then the set $\{x \in S \mid P(x) \text{ is false}\}$ is nonempty, so has a least element, say x_0. This means that $P(x_0)$ is false but $P(x)$ is true for all $x \in S$ with $x < x_0$. So x_0 is a "minimal counterexample" to P, and often we can work from the truth of P for smaller values to the truth of P for x_0.

Conversely, if assume Theorem 7.1 we can *prove* the Well-Ordering Principle: seeking a contradiction, let $T \subseteq \mathbb{N}$ be a nonempty subset without a minimum. Let

$$S := \{x \in \mathbb{N} \mid \forall y \in T, x < y\}.$$

That is, S consists of the non-negative integers that are less than every element of T. We have $S \cap T = \varnothing$, since any element of both would be an element of T that is less than itself. It suffices to show that S is an inductive subset of $\mathbb{N}$, for then by Theorem 7.1 we have $S = \mathbb{N}$, so

$$\varnothing = S \cap T = \mathbb{N} \cap T = T.$$

First we claim that $0 \in S$. If not, then 0 is not less than every element of T, which means that $0 \in T$, so 0 is the minimum of T, contradiction. Next, suppose that $n \in S$ and $n + 1 \notin S$. This means that every element of T is greater than n but some element of T is not greater than $n + 1$. From this we deduce that $n + 1$ is the least element of T, contradiction.

This shows that Theorem 7.1 is not just a consequence of the Well-Ordering Principle – it is actually equivalent to it! In this text we have taken the well-orderedness of $\mathbb{N}$ as an *axiom*, i.e., something that we assume without proof. We have to assume *some* axioms in order to be able to make deductions.

Recall that in Euclidean geometry one studies points, lines, planes and so forth, but one does not start by saying what sort of object the Euclidean plane "really is".

(At least this is how Euclidean geometry has been approached for more than a hundred years. Euclid himself gave such "definitions" as: "A point is that which has position but not dimensions." "A line is breadth without depth." In the 19th century it was recognized that these are descriptions rather than definitions, in the same way that many dictionary definitions are actually descriptions:

"cat: A small carnivorous mammal domesticated since early times as a catcher of rats and mice and as a pet and existing in several distinctive breeds and varieties."

This helps you if you are already familiar with the animal but not the word, but if you have never seen a cat before this definition would certainly not allow you to determine with certainty whether any particular animal you encountered was a cat, and still less would it allow you to reason abstractly about the cat concept or "prove theorems about cats.")

Rather "point", "line", "plane" and so forth are taken as **undefined terms**. They are related by certain **axioms**, or abstract properties that they must satisfy.

In 1889, the Italian mathematician and proto-logician Gisueppe Peano came up with a similar (and, in fact, much simpler) system of axioms for the natural numbers. In slightly modernized form, this goes as follows:

The undefined terms are **zero**, **number** and **successor**.

There are five axioms that they must satisfy, the **Peano axioms**. The first four are:

(P1) Zero is a number.
(P2) Every number has a successor, which is also a number.
(P3) No two distinct numbers have the same successor.

(P4) Zero is not the successor of any number.

Using set-theoretic language we can clarify what is going on here as follows: the structures we are considering are triples $(X, 0, S)$, where X is a set, 0 is an element of X, and $S : X \to X$ is a function, subject to the above axioms.

From this we can deduce quite a bit. First, we have a number (i.e., an element of X) called $S(0)$. Is $0 = S(0)$? No, that is prohibited by (P4). We also have a number $S(S(0))$, which is not equal to 0 by (P4) and it is also not equal to $S(0)$, because then $S(0) = S(S(0))$ would be the successor of the distinct numbers 0 and $S(0)$, contradicting (P3). Continuing in this way, we can produce an infinite sequence of distinct elements of X:

$$0, S(0), S(S(0)), S(S(S(0))), \dots. \tag{24}$$

In particular X itself is infinite. The crux of the matter is this: is there any element of X that is *not* a member of the sequence (24), i.e., is not obtained by starting at 0 and applying the successor function finitely many times?

The axioms so far do not allow us to answer this question. For instance, suppose that the "numbers" consisted of the set $[0, \infty)$ of all non-negative real number. We define 0 to be the real number of that name, and we define the successor of x to be $x + 1$. This system satisfies (P1) through (P4) but has much more in it than just the natural numbers, so we must be missing an axiom! Indeed, the last axiom is:

(P5) If Y is a subset of the set X of numbers such that $0 \in Y$ and such that $x \in Y$ implies $S(x) \in Y$, then $Y = X$.

Notice that (P5) is a rephrasing of the assertion that the only inductive subset of $\mathbb{N}$ is $\mathbb{N}$ itself. Also the eaxmple we cooked up above fails (P5), since in $[0, \infty)$ the subset of natural numbers contains zero and contains the successor of each of its elements but is a proper subset of $[0, \infty)$.

Thus it was Peano's contribution to realize that mathematical induction is an axiom for the natural numbers in much the same way that the parallel postulate is an axiom for Euclidean geometry.

From the modern perspective, all mathematical structures are axiomatized in terms of *pure sets* (sets whose elements are again sets), so this axiomatization of Peano is intermediate to what we really want: define each natural number as a set (hint: $0 := \varnothing$ is a good start!) and then verify for the set of all natural numbers, the five axioms (P1) through (P5) hold. If we did that, we would have a complete proof of Theorem 7.1 hence also of the Well-Ordering Principle. But we will not do that here.

The history is interesting: this work of Peano is little more than one hundred years old, which in the scope of mathematical history is quite recent. Traces of what we now recognize as induction can be found from the mathematics of antiquity (including Euclid's Elements!) on forward. According to the (highly recommended!) Wikipedia article on mathematical induction, the first mathematician to formulate it explicitly was Blaise Pascal, in 1665. During the next hundred years various

equivalent versions were used by different mathematicians – notably the methods of infinite descent and minimal counterexample, which we shall discuss later – and the technique seems to have become commonplace by the end of the 18th century.

2. Principle of Mathematical Induction for Sentences

We have given Theorem 7.1 to provide a clean, solid foundation for induction. In truth, the following is the form of induction that we will almost always use.

THEOREM 7.2 (Principle of Mathematical Induction for Sentences). *Let $P(n)$ be an open sentence with domain $n \in \mathbb{N} = \{0, 1, 2, 3 \ldots\}$. Suppose that:*

(I1) *$P(0)$ is true, and*

(I2) *$\forall n \in \mathbb{N}$, $P(n) \implies P(n+1)$.*

Then: $P(n)$ is true for all $n \in \mathbb{N}$.

As we will now explain, our two Principles of Mathematical Induction – Theorem 7.1 and Theorem 7.2 – are equivalent. This is yet antoher instance of an equivalence between sets and logic.

Indeed, let us first show that the Principle of Mathematical Induction for Subsets implies the Principle of Mathematical Induction for Sentences: in other words, assuming Theorem 7.1 we will prove Theorem 7.2. Let $P(n)$ be an open sentence with domain $n \in \mathbb{N}$ such that $P(0)$ is true and for all $n \in \mathbb{N}$, $P(n) \implies P(n+1)$. As we have done in Chapter 4, consider the *truth locus* of P:

$$S \coloneqq \{n \in \mathbb{N} \mid P(n) \text{ is true}\}.$$

Our first assumption is that $P(0)$ is true, so $0 \in S$. Our second assumption is that for all $n \in \mathbb{N}$, $P(n) \implies P(n+1)$, so $\forall n \in \mathbb{N}$, $n \in S \implies n+1 \in S$. Thus S is an inductive subset of $\mathbb{N}$, so by Theorem 7.1 we have $S = \mathbb{N}$: done!

Now let us show that the Principle of Mathematical Induction for Sentences implies the Principle of Mathematical Induction for Subsets: in other words, assuming Theorem 7.2 we will prove Theorem 7.1. Let $S \in \mathbb{N}$ be a subset such that $0 \in S$ and $\forall n \in \mathbb{N}$, $n \in S \implies n+1 \in S$. We want to show that $S = \mathbb{N}$, and our task is to come up with an open sentence $P(n)$ with domain $n \in \mathbb{N}$ to help us show this. After a little thought the following presents itself: for $n \in \mathbb{N}$, we define $P(n)$ to be the assertion that n is an element of S. Then our first assumption tells us that $P(0)$ is true and our second assumption tells us that for all $n \in \mathbb{N}$, $P(n) \implies P(n+1)$. So by Theorem 7.2 we have that $P(n)$ is true for all n, which means that for all $n \in \mathbb{N}$ we have $n \in S$, so $S = \mathbb{N}$: done!

3. A Slight Generalization

In induction proofs we will not always have an open sentence $P(n)$ with domain $n \in \mathbb{N}$. In fact, more often than not our open sentence will have domain $n \in \mathbb{Z}^+$: i.e., we start at 1 rather than 0. It is no problem at all to work in this context, and indeed a bit more generally.

THEOREM 7.3 (Principle of Mathematical Induction for Sentences, v. II). *Let $N \in \mathbb{Z}$, and let $P(n)$ be an open sentence with domain $n \in \mathbb{Z}^{\geq N} = \{N, N+1, N+2 \ldots\}$. We suppose that:*

(I1) *$P(N)$ is true, and*
(I2) *$\forall n \in \mathbb{Z}^{\geq N}$, we have $P(n) \implies P(n+1)$.*
Then $P(n)$ is true for all integers $n \geq N$.

PROOF. There are several ways to proceed. We will simply "shift everthing back to 0.' Namely, put

$$S := \{n - N \mid n \in \mathbb{Z}^{\geq N} \text{ and } P(n) \text{ is true.}$$

Since n is an integer that is at least N, it follows that $n - N$ is an integer that is at least $N - N = 0$, which is to say that $n \in \mathbb{N}$. So S is a subset of $\mathbb{N}$. The key to the proof is the following simple observation:

$$\forall n \in \mathbb{N}, n \in S \iff P(n+N) \text{ is true.}$$

From this we can see that S is an inductive subset of $\mathbb{N}$: by (I1) $P(N)$ is true, so $N - N = 0 \in S$. And for all $n \in \mathbb{N}$, if $n \in S$, then $n + N \in \mathbb{Z}^{\geq N}$ and $P(n+N)$ is true, so by (I2) we have that $P(n+1+N)$ is true, so $n+1+N-N = n+1 \in S$. By Theorem 7.1 we have $S = \mathbb{N}$, meaning that for all $n \in N$, $P(n+N)$ is true, which means that $P(n)$ holds for all integers $n \geq N$. □

Here is another plausible approach. For $N \in \mathbb{Z}^+$, we define an inductive subset $S \subseteq \mathbb{Z}^{\geq N}$ to a set such that $N \in \S$ and $\forall n \in \mathbb{Z}^{\geq N}$, $n \in S \implies n+1 \in S$. By Exercise 5.3, $\mathbb{Z}^{\geq N}$ is also well-ordered, so arguing as in the proof of Theorem 7.1 we can show that the only inductive subset of $\mathbb{Z}^{\geq N}$ is $\mathbb{Z}^{\geq N}$ itself, and this leads to another proof of Theorem 7.3. You are asked to fill in the details in Exercise 7.1.

The moral here is that induction proofs work for statements parametrized by the integers starting at any fixed integer. Whether the starting integer is 0 or 1 or something else is not important, so long as the logic of induction is clearly understood so as to make a proper connection between the base case and the induction step. To undersatnd what that means, let us move on to actual proofs by induction!

4. The (Pedagogically) First Induction Proof

We will prove many things by induction in this text. But there is a traditional *first* result for students to see proved by induction, and we see no reason not to follow this tradition. So here we go:

PROPOSITION 7.4. *For all $n \in \mathbb{Z}^+$, $1 + \ldots + n = \frac{n(n+1)}{2}$.*

PROOF. We go by induction on n.
Base case ($n = 1$): Indeed $1 = \frac{1(1+1)}{2}$.
Induction step: Let $n \in \mathbb{Z}^+$ and suppose that $1 + \ldots + n = \frac{n(n+1)}{2}$. Then

$$1 + \ldots + n + n + 1 = (1 + \ldots + n) + n + 1 \stackrel{\text{IH}}{=} \frac{n(n+1)}{2} + n + 1$$

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

Here the letters "IH" signify that the induction hypothesis was used. □

Induction is such a powerful tool that once one learns how to use it one can prove many nontrivial facts with essentially no thought or ideas required, as is the case in the above proof. However thought and ideas are good things when you have them!

In many cases an inductive proof of a result is a sort of "first assault" which raises the challenge of a more insightful, noninductive proof. This is certainly the case for Proposition 7.4 above, which can be proved in many ways.

Here is one non-inductive proof: replacing n by $n-1$, it is equivalent to show:

$$\forall n \in \mathbb{Z}, n \geq 2 : \ 1 + \ldots + n - 1 = \frac{(n-1)n}{2}. \tag{25}$$

We recognize the quantity $\frac{(n-1)n}{2}$ on the right-hand side as the **binomial coefficient** $\binom{n}{2}$: it counts the number of 2-element subsets of an n element set. This raises the prospect of a **combinatorial proof**, i.e., to show that the number of 2-element subsets of an n element set is *also* equal to $1+2+\ldots+n-1$. This comes out immediately if we list the 2-element subsets of $\{1, 2, \ldots, n\}$ in a systematic way: we may write each such subset as $\{i, j\}$ with $1 \leq i \leq n-1$ and $i < j \leq n$. Then:

The subsets with least element 1 are $\{1,2\}, \{1,3\}, \ldots, \{1,n\}$, a total of $n-1$.
The subsets with least element 2 are $\{2,3\}, \{2,4\}, \ldots, \{2,n\}$, a total of $n-2$.
$\vdots$
The subset with least element $n-1$ is $\{n-1, n\}$, a total of 1.

Thus the number of 2-element subsets of $\{1, \ldots, n\}$ is on the one hand $\binom{n}{2}$ and on the other hand $(n-1) + (n-2) + \ldots + 1 = 1 + 2 + \ldots + n - 1$. This gives a combinatorial proof of Proposition 7.4.

For a very striking pictorial variation of the above argument, go to `http://mathoverflow.net/questions/8846/proofs-without-words` and scroll down to the first diagram.

5. The (Historically) First(?) Induction Proof

Recall Theorem 6.9: there are infinitely many primes. As we discussed, it may be expressed as a proof by contradiction but this aspect of the proof can be removed. How did Euclid himself express the proof? We quote from a complete online translation of *The Elements* made by David Joyce [**Jo**].

"*Prime numbers are more than any assigned multitude of prime numbers.*

Let A , B and C be the assigned prime numbers.

I say that there are more prime numbers than A, B, and C.

Take the least number DE measured by A, B and C. Add the unit DF to DE.

Then EF is either prime or not.

Firs, let it be prime. Then the prime numbers A, B, and EF have been found which are more than A, B, and C.

Next, let EF not be prime. Therefore it is measured by some prime number.

Let it be measured by the prime number G.

I say that G is not the same with any of the numbers A, B, and C.

If possible, let it be so. Now A, B, and C measure DE, therefore G also measures DE. But it also measures EF. Therefore G, being a number, measures the reaminder, the unit DF, which is absurd.

Therefore G is not the same with any one of the numbers A, B, and C. And by hypothesis it is prime. Therefore the prime numbers A, B, C, and G have been found which are more than the assigned multitude of A, B, and C.

Therefore, *prime numbers are more than any assigned multitude of prime numbers.*"

Our first comment is the relatively superficial one that Euclid interprets numbers as lengths of line segments and therefore expresses arithmetic statements via a light geometric code: for instance he says that one number x *can be measured by* another number y – here x and y are lengths of line segments, or what we would call positive real numbers – if one can build the segment x out of a positive whole number of copies of the segment y, or in other words if $\frac{x}{y} \in \mathbb{Z}^+$. By the least number measured by A , B and C he means the least common mulitple of A, B and C. Since A, B and C are distinct primes, this is indeed just $A \cdot B \cdot C$. At the end of the proof, the fact that if $n \mid x$ and $n \mid y$ then $n \mid x-y$ is understood geometrically.

Let us look deeper. The next thing that we notice is that Euclid does not explicitly deal with infinite sets: the statement of the result is that prime numbers are more than any assigned multitude, so evidently a "multitude" means either a finite set or the number of elements of a finite set. Thus rather than "There are infinitely many primes," a closer formulation to Euclid's is "For all $n \in \mathbb{Z}^+$, there are more than n prime numbers."

How, in broad terms, does Euclid prove that for any positive integer n, there are more than n prime numbers? Strictly speaking, he assumes that he has three primes, and shows that there is a fourth. This is not sufficient! Of course the argument is more general than this: as we have already seen, it shows that if you are given any n distinct prime numbers $p_1, \ldots, p_n$, any prime factor of the integer $p_1 \cdots p_n + 1 \geq 3$ gives you an new prime number p_{n+1}. Why does this suffice to show that there are, for any positive integer N, more than N primes? We didn't argue for this directly, but rather observed that the negation of "There are infinitely many primes" is "The set of all primes is finite," so – after observing that this set is nonempty! – the set of all primes is $\{p_1, \ldots, p_n\}$ for some $n \geq 1$, and this is impossible, because given any n primes, we must have another.

Here is a more direct explanation: for $n \in \mathbb{Z}^+$, let $P(n)$ be the statement that there are more than n primes. Euclid wants to prove

$$\forall n \in \mathbb{Z}^+, \ P(n). \tag{26}$$

He has proved

$$P(3) \implies P(4)$$

but as readers we are surely meant to understand that his argument establishes

$$\forall n \in \mathbb{Z}^+,\ P(n) \implies P(n+1).$$

Therefore he has very nearly proved (26)...by induction! To complete the induction proof he needs only to establish $P(1)$, i.e., that there is at least one prime. Okay, sure: again, 2 is a prime. Though I insist on the necessity of the base case for the validity of an inductive proof (cf. the Dubious Claim of §8), we can forgive a mathematician writing over 2200 years ago for neglecting it: in this case, the truth of $P(1)$ is certainly obvious. All in all, I think there is little doubt that Euclid intends his proof of Proposition IX.20 of *The Elements* to be by induction. This is of historical interest, because as mentioned earlier, the first explicit use of induction as a proof technique seems to have been by Pascal in 1665. Euclid's argument comes earlier – more than 1900 years earlier!

The commmon misconception that $p_1 \cdots p_n + 1$ must itself be prime cannot be pinned on Euclid: his argument is *not* a counterfactual one that forces that to be the case. Rather, he explicitly contemplates both the possibilities that it is prime and that it isn't prime.

We have discussed various mathematical misconceptions that can arise from the proof of Euclid's Proposition IX.20 on the infinitude of primes, but there remains the historical misconception that this is how Euclid himself proceeded. People really seem to like to phrase this argument as a proof by contradiction – including me! A further discussion of these related misconceptions occurs in [**HW09**]. They trace this back to the leading mathematician and highly influential mathematical writer G.H. Hardy[1]. Hardy includes Euclid's proof of the infinitude of primes in several of his texts, always cast as a proof by contradiction, accompanied by either explicit or implicit claims that this is what Euclid did. In Hardy's *A Mathematician's Apology* [**H**] – surely one of the most widely read of all texts about mathematics of recent years – Hardy presents as his two examples of beautiful proofs the irrationality of $\sqrt{2}$ and the infinitude of primes, both cast as proofs by contradiction. Moreover on [**H**, p. 68] he describes this proof techinque with unforgettable eloquence:

"The proof is by *reductio ad absurdum*, and *reductio ad absurdum*, which Euclid loved so much, is one of a mathematician's finest weapons. It is a far finer gambit than any chess gambit: a chess player may offer the sacrifce of a pawn or even a piece, but a mathematician ofers *the game*."

The current debriefing notwithstanding, the influence of Hardy's writing on the present text ought to be clear.

Remark 7.5. *Some scholars have suggested that what is essentially an argument by mathematical induction appears in the later middle Platonic dialogue* Parmenides*, lines 149a7-c3. But this argument is of mostly historical and philosophical interest.*[2] *The statement in question is, very roughly, that if n objects are*

[1]Godfrey Harold Hardy, 1877–1947.

[2]Since Plato lived from 427 BCE to 347 BCE and Euclid lived from circa 325 BC to circa 265 BCE, the *Parmenides* safely precedes the *Elements*.

placed adjacent to another in a linear fashion, the number of points of contact between them is $n-1$. (Maybe. To quote the lead in the wikipedia article on the Parmenides*: "It is widely considered to be one of the more, if not the most, challenging and enigmatic of Plato's dialogues.") There is not much mathematics here! Nevertheless, for a thorough discussion of induction in the* Parmenides *the reader may consult* [**Ac00**] *and the references cited therein.*

6. Closed Form Identities

The inductive proof of Proposition 7.4 is a prototype for a certain kind of induction proof (the easiest kind!) in which $P(n)$ is some algebraic identity: say $LHS(n) = RHS(n)$. In this case to make the induction proof work you need only (i) establish the base case and (ii) verify the equality of successive differences

$$LHS(n+1) - LHS(n) = RHS(n+1) - RHS(n).$$

We give two more familiar examples of this.

PROPOSITION 7.6. *For all $n \in \mathbb{Z}^+$, $1 + 3 + \ldots + (2n-1) = n^2$.*

PROOF. Let $P(n)$ be the statement "$1+3+\ldots+(2n-1) = n^2$". We will show that $P(n)$ holds for all $n \in \mathbb{Z}^+$ by induction on n.
Base case $n = 1$: indeed $1 = 1^2$.
Induction step: Let n be an arbitrary positive integer and assume $P(n)$:

$$1 + 3 + \ldots + (2n-1) = n^2. \tag{27}$$

Adding $2(n+1) - 1 = 2n + 1$ to both sides, we get

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

This is precisely $P(n+1)$, so the induction step is complete. □

PROPOSITION 7.7. *For all $n \in \mathbb{Z}^+$, $1^2 + 2^2 + \ldots + n^2 = \frac{n(n+1)(2n+1)}{6}$.*

PROOF. By induction on n.
Base case: $n = 1$. We have $1^2 = 1 = \frac{1(1+1)(2\cdot 1+1)}{6}$.
Induction step: Let $n \in \mathbb{Z}^+$ and suppose that $1^2 + \ldots + n^2 = \frac{n(n+1)(2n+1)}{6}$. Then

$$1 + \ldots + n^2 + (n+1)^2 \overset{\text{IH}}{=} \frac{n(n+1)(2n+1)}{6} + (n+1)^2 =$$

$$\frac{2n^3 + 3n^2 + n + 6 + 6n^2 + 12n + 1}{6} = \frac{2n^3 + 9n^2 + 13n + 7}{6}.$$

Expanding out $\frac{(n+1)((n+1)+1)(2(n+1)+1)}{6}$, we also get $\frac{2n^3+9n^2+13n+7}{6}$. □

Often a non-inductive proof, when available, offers more insight. Again returning to our archetypical example: $1 + \ldots + n$, it is time to tell the story of little Gauss. As a child of no more than 10 or so, Gauss and his classmates were asked to add up the numbers from 1 to 100. Most of the students did this by a laborious calculation and got incorrect answers in the end. Gauss reasoned essentially as follows: put

$$S_n = 1 + \ldots + (n-1) + n.$$

Of course the sum is unchanged if we we write the terms in descending order:

$$S_n = n + (n-1) + \ldots + 2 + 1.$$

Adding the two equations gives

$$2S_n = (n+1) + (n+1) + \ldots + (n+1) = n(n+1),$$

so

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

This is no doubt preferable to induction, so long as one is clever enough to see it.

Mathematical induction can be viewed as a particular incarnation of a much more general proof technique: try to solve your problem by reducing it to a previously solved problem. A more straightforward application of this philosophy allows us to deduce Proposition 7.6 from Proposition 7.4:

$$1+3+\ldots+(2n-1) = \sum_{i=1}^{n}(2i-1) = 2\sum_{i=1}^{n} i - \sum_{i=1}^{n} 1 = 2\left(\frac{n(n+1)}{2}\right) - n = n^2+n-n = n^2.$$

7. More on Power Sums

Suppose now we want to find a formula for $\sum_{i=1}^{n} i^3 = 1^3 + \ldots + n^3$.[3] A key point: we can't use induction yet because we don't know what the answer is![4]

So let's try to actually think about what's going on. We previously found a formula for $\sum_{i=1}^{n} i$ which was a quadratic polynomial in n, and then a formula for $\sum_{i=1}^{n} i^2$ which was a cubic polynomial in n. We might therefore guess that the desired formula for $\sum_{i=1}^{n} i^3$ is a fourth degree polynomial in n, say

$$a_4 n^4 + a_3 n^3 + a_2 n^2 + a_1 n + a_0.$$

If we think more seriously about Riemann sums, the fundamental theorem of calculus and the fact that $\frac{x^4}{4}$ is an antiderivative for x^3, this guess becomes more likely, and we can even guess that $a_4 = \frac{1}{4}$. Also by looking at the other examples we might guess that $a_0 = 0$. So we are looking for (presumably rational?) numbers a_1, a_2, a_3 such that

$$1^3 + \ldots + n^3 = \frac{1}{4}n^4 + a_3 n^3 + a_2 n^2 + a_1 n.$$

Now, inspired by the partial fractions technique in calculus, we can simply plug in a few values and solve for the coefficients. For instance, taking $n = 1, 2, 3$ we get

$$1^3 = 1 = \frac{1}{4} + a_3 + a_2 + a_1,$$

$$1^3 + 2^3 = 9 = 4 + 8a_3 + 4a_2 + 2a_1,$$

$$1^3 + 2^3 + 3^3 = 9 + 3^3 = 36 = \frac{81}{4} + 27a_3 + 9a_2 + 3a_1.$$

This gives us the linear system

$$a_1 + a_2 + a_3 = \frac{3}{4}$$

[3] Why might we want this? For instance, such sums arise in calculus as Riemann sums for the integral $\int_a^b x^3 dx$. Of course there is a better way to evaluate such integrals, via the Fundamental Theorem of Calculus. Perhaps it is safest to say that finding closed formulas for sums is an intrinsically interesting, and often quite challenging, endeavor.

[4] As we will see again and again, this is, like Kryptonite for Superman, induction's only weakness.

$$2a_1 + 4a_2 + 8a_3 = 5$$
$$3a_1 + 9a_2 + 27a_3 = \frac{63}{4}.$$

I will leave it to you to do the math here, in what way seems best to you.[5] The unique solution is $a_1 = 0,\ a_2 = \frac{1}{4},\ a_3 = \frac{1}{2}$, so that our conjectural identity is

$$1^3 + \ldots + n^3 = \frac{n^4}{4} + \frac{n^3}{2} + \frac{n^2}{4} = \frac{n^2}{4}(n^2 + 2n + 1) = \left(\frac{n(n+1)}{2}\right)^2. \tag{28}$$

In Exercise 7.4 you are asked to use induction to prove that (28) holds for all $n \in \mathbb{Z}^+$. In Exercise 7.5 you are asked to use a similar technique to discover and prove a closed form expression for $\sum_{i=1}^n i^4$.

The above method is a useful one for solving many types of problems: make a guess as to the general form the answer may take, plug that guess in and fine tune the constants accordingly. In this case the method has two limitations: first, it involves a rather large amount of calculation, and second we cannot find out whether our general guess is correct until after all the calculations have been made. In this case, there is a better way to derive formulas for the power sums

$$S_d(n) = 1^d + \ldots + n^d.$$

We begin with the sum

$$S = \sum_{i=1}^n \left((i+1)^{d+1} - i^{d+1}\right),$$

which we evaluate in two different ways. First, writing out the terms gives

$$S = 2^{d+1} - 1^{d+1} + 3^{d+1} - 2^{d+1} + \ldots + n^{d+1} - (n-1)^{d+1} + (n+1)^{d+1} - n^{d+1} = (n+1)^{d+1} - 1.$$

Second, by first expanding out the binomial $(i+1)^{d+1}$ we get

$$S = \sum_{i=1}^n \left((i+1)^{d+1} - i^{d+1}\right) = \sum_{i=1}^n \left(i^{d+1} + \binom{d+1}{1} i^d + \ldots + \binom{d+1}{d} i + 1 - i^{d-1}\right) =$$

$$\sum_{i=1}^n \left(\binom{d+1}{1} i^d + \ldots + \binom{d+1}{d} i\right) = \binom{d+1}{1} \sum_{i=1}^n i^d + \ldots + \binom{d+1}{d} \sum_{i=1}^n i + \sum_{i=1}^n 1 =$$

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

Equating our two expressions for S, we get

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

Solving this equation for $S_d(n)$ gives

$$S_d(n) = \frac{(n+1)^{d+1} - \left(\sum_{j=0}^{d-1} \binom{d+1}{j} S_j(n)\right) - 1}{(d+1)}. \tag{29}$$

[5] Yes, this is an allusion to *The Return of the King*.

This formula allows us to compute $S_d(n)$ recursively: that is, given exact formulas for $S_j(n)$ for all $0 \leq j < d$, we get an exact formula for $S_d(n)$. And getting the ball rolling is easy: $S_0(n) = 1^0 + \ldots + n^0 = 1 + \ldots 1 = n$.

EXAMPLE 7.8. *($d = 1$): Our formula gives*

$$1+\ldots+n = S_1(n) = (\frac{1}{2})((n+1)^2 - S_0(n) - 1) = (\frac{1}{2})(n^2+2n+1-n-1) = \frac{n(n+1)}{2}.$$

EXAMPLE 7.9. *($d = 2$): Our formula gives* $1^2 + \ldots + n^2 = S_2(n) =$

$$\frac{(n+1)^3 - S_0(n) - 3S_1(n) - 1}{3} = \frac{n^3 + 3n^2 + 3n + 1 - n - \frac{3}{2}n^2 - \frac{3}{2}n - 1}{3} =$$

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

Our formula (29) also has theoretical applications: with it in hand we can apply induction to a loftier goal, namely the proof of the following result.

THEOREM 7.10. *For all $d \in \mathbb{Z}^+$, there are $a_1, \ldots, a_d \in \mathbb{Q}$ such that for all $n \in \mathbb{Z}^+$ we have*

$$1^d + \ldots + n^d = \frac{n^{d+1}}{d+1} + a_d n^d + \ldots + a_1 n.$$

You are asked to prove Theorem 7.10 in Exercise 7.6.

There are *many* other approaches to evaluating the power sums $S_d(n)$. Let us briefly mention an interesting one taken in a recent paper of Dueñez, Hamakiotes and Miller [**DHM24**]. We begin with the finite geometric series: let $x \in \mathbb{R}$ and $n \in \mathbb{Z}^+$. Then

$$(1 + x + \ldots + x^{n-1} + x^n)(x - 1) = x^{n+1} - 1 :$$

indeed, when we expand out the left hand side we get

$$x + x^2 + \ldots + x^n + x^{n+1} - (1 + x + \ldots + x^n) = x^{n+1} - 1.$$

This is an instance of what is often called a "telescoping sum" (meaning that upon performing some simple arithmetic, most of the terms cancel out). Suppose now that $x \neq 1$, so we may divide by $x - 1$, getting:

$$\forall n \in \mathbb{N},\ \forall x \in \mathbb{R} \setminus \{1\},\ 1 + x + \ldots + x^n = \frac{x^{n+1} - 1}{x - 1}. \tag{30}$$

Now we view (30) as an identity of the form $f = g$, where $f, g : \mathbb{R} \setminus \{1\} \to \mathbb{R}$ are differentiable functions. It follows of course that $f'(x) = g'(x)$ for all $x \in \mathbb{R} \setminus \{1\}$; multiplying this by x, it follows that $xf'(x) = xg'(x)$ for all $x \in \mathbb{R} \setminus \{1\}$. If we write this out for $f(x) = 1 + x + \ldots + x^n$ and $g(x) = \frac{x^{n+1}-1}{x-1}$, we get

$$x + 2x^2 + \ldots + nx^n = x(1 + 2x + \ldots + nx^{n-1}) = xf'(x)$$

$$= xg'(x) = x\left(\frac{(n+1)x^n(x-1) - (x^{n+1} - 1)}{(x-1)^2}\right),$$

or, after some simplifications,

$$x + 2x^2 + \ldots + nx^n = \frac{nx^{n+2} - (n+1)x^{n+1} + x}{(x-1)^2}. \tag{31}$$

Okay – so what?!? Well, when we plug in $x = 1$ to the left hand side of (31), we get $S_1(n) = 1 + \ldots + n$. We cannot immediately plug in $x = 1$ to the right hand side, because there is an $x - 1$ in the denominator. However, because the left hand side is a polynomial, it is continuous at $x = 1$, so the limit of the right hand side as $x \to 1$ exists and equals the value of the left hand side at $x = 1$, namely $S_1(n)$. So if we can compute the limit on the right hand side, we will have computed $S_1(n)$. But the standard methods of calculus apply here: we can apply L'Hôpital's Rule twice. The first application of L'Hôpital gives

$$S_1(n) = \lim_{x \to 1} \frac{n(n+2)x^{n+1} - (n+1)^2 x^n + 1}{2(x-1)},$$

provided the latter limit exists. The latter limit is still of the form $\frac{0}{0}$; applying L'Hôpital again gives

$$S_1(n) = \lim_{x \to 1} \frac{n(n+1)(n+2)x^n - n(n+1)^2 x^{n-1}}{2}$$

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

This was certainly not the easiest or most elementary way to evaluate $S_1(n) = 1 + \ldots + n$, but (i) it's interesting; and (ii) it generalizes well: for $d \in \mathbb{Z}^+$, to evaluate the power sum $S_d(n) = 1^d + \ldots + n^d$ this way, "all" one needs to do is apply the operator $x\frac{d}{dx}$ to the basic identity (30) d times in succession: the left hand side will then be

$$1^d x + 2^d x^2 + \ldots + n^d x^n$$

and the right hand side will be of the form $\frac{P_{d,n}(x)}{(x-1)^{2^d}}$ where $P_{d,n}(x)$ is a polynomial of degree $n + 2^d$. Applying l'Hôpital 2^d times, we find that

$$S_d(n) = \frac{P_{d,n}^{(2^d)}(1)}{(2^d)!}, \tag{32}$$

where for a function f and $N \in \mathbb{Z}^+$, $f^{(N)}$ denotes its Nth derivative. The scare quotes around *all* are there to indicate that this procedure is in fact rather computationally intensive even for small values of d. For instance, taking $d = 2$ we get:

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

so that

$$P_{2,n}(x) = n^2x^{n+4} + (-3n^2 - 2n + 1)x^{n+3} + (3n^2 + 4n)x^{n+2} - (n+1)^2x^{n+1} - x^3 + x.$$

Then $P_{2.n}^{(4)}(x) =$

$$n(n+1)x^{n-3}(n^4(x-1)^3 + n^3(x-1)^2(9x+1) + n^2(26x^3 - 28x^2 - 2x + 3) + n(24x^3 - 7x^2 - 8x - 1) + 6x^2 - 2.$$

Evaluating at $x = 1$, we get

$$P_{2,n}^{(4)}(1) = 4n(2n^2 + 3n + 1),$$

so

$$S_2(n) = \frac{P_{2,n}^{(4)}(1)}{4!} = \frac{n(n+1)(2n+1)}{6}.$$

In Exercises 7.7, 7.8 and 7.9 you are asked to prove (32), confirm the calculations stated above for $d = 2$ and apply (32) to give another proof of Theorem 7.10.

8. Inequalities

PROPOSITION 7.11. *For all $n \in \mathbb{N}$, we have $2^n > n$.*

Proof analysis: For $n \in \mathbb{N}$, let $P(n)$ be the statement "$2^n > n$". We want to show that $P(n)$ holds for all natural numbers n by induction.
Base case: $n = 0$: $2^0 = 1 > 0$.
Induction step: let $n \in \mathbb{N}$ and asusme $P(n)$: $2^n > n$. Then

$$2^{n+1} = 2 \cdot 2^n > 2 \cdot n.$$

We would now like to say that $2n \geq n + 1$. But in fact this is true if and only if $n \geq 1$. Well, don't panic. We just need to restructure the argument a bit: we verify the statement separately for $n = 0$ and then use $n = 1$ as the base case of our induction argument. Here is a formal writeup:

PROOF. Since $2^0 = 1 > 0$ and $2^1 = 2 > 1$, it suffices to verify the statement for all natural numbers $n \geq 2$. We go by induction on n.
Base case: $n = 2$: $2^2 = 4 > 2$.

Induction step: Assume that for some integer $n \geq 2$ we have $2^n > n$. Then

$$2^{n+1} = 2 \cdot 2^n > 2 \cdot n > n + 1,$$

since $n > 1$. □

Exercise 7.10 reminds us that induction is not the *only* tool to prove inequalities.

PROPOSITION 7.12. *There exists $N_0 \in \mathbb{Z}^+$ such that for all $n \geq N_0$, $2^n \geq n^3$.*

Proof analysis: A little experimentation shows that there are several small values of n such that $2^n < n^3$: for instance $2^9 = 512 < 9^3 = 729$. On the other hand, it seems to be the case that we can take $N_0 = 10$: let's try.
Base case: $n = 10$: $2^{10} = 1024 > 1000 = 10^3$.

Induction step: Suppose that for some $n \geq 10$ we have $2^n \geq n^3$. Then

$$2^{n+1} = 2 \cdot 2^n \geq 2n^3.$$

Our task is then to show that $2n^3 \geq (n+1)^3$ for all $n \geq 10$. (By considering limits as $n \to \infty$, it is certainly the case that the left hand side exceeds the right hand side for all sufficiently large n. It's not guaranteed to work for $n \geq 10$; if not, we will replace 10 with some larger number.) Now,

$$2n^3 - (n+1)^3 = 2n^3 - n^3 - 3n^2 - 3n - 1 = n^3 - 3n^2 - 3n - 1 \geq 0$$
$$\iff n^3 - 3n^2 - 3n \geq 1.$$

Since everything in sight is a whole number, this is in turn equivalent to

$$n^3 - 3n^2 - 3n > 0.$$

Now $n^3 - 3n^2 - 3n = n(n^2 - 3n - 3)$, so this is equivalent to $n^2 - 3n - 3 \geq 0$. The roots of the polynomial $x^2 - 3x - 3$ are $x = \frac{3 \pm \sqrt{21}}{2}$, so $n^2 - 3n - 3 > 0$ if $n > 4 = \frac{3+\sqrt{25}}{2} > \frac{3+\sqrt{21}}{2}$. In particular, the desired inequality holds if $n \geq 10$, so by induction we have shown that $2^n \geq n^3$ for all $n \geq 10$.

We leave it to to the student to convert the above analysis into a formal proof.

REMARK 7.13. *We have $2^n \geq n^3$ for all natural numbers n* except $n = 2, 3, 4, 6, 7, 8, 9$. *It is interesting that the desired inequality is true for a little while (i.e., at $n = 0, 1$) then becomes false for a little while longer, and then becomes true for all $n \geq 10$. It follows from our analysis that if for any $N \geq 4$ we have $2^N \geq N^3$, then this equality remains true for all larger natural numbers n. Thus from the fact that $2^9 < 9^3$, we can in fact deduce that $2^n < n^3$ for all $4 \leq n \leq 8$.*

PROPOSITION 7.14. *For all $n \in \mathbb{Z}^+$, $1 + \frac{1}{4} + \ldots + \frac{1}{n^2} \leq 2 - \frac{1}{n}$.*

Proof analysis: By induction on n.
Base case ($n = 1$): $1 \leq 2 - \frac{1}{1}$.
Induction step: Let $n \in \mathbb{Z}^+$, and suppose that $1 + \frac{1}{4} + \ldots + \frac{1}{n^2} \leq 2 - \frac{1}{n}$. Then

$$1 + \frac{1}{4} + \ldots + \frac{1}{n^2} + \frac{1}{(n+1)^2} \leq 2 - \frac{1}{n} + \frac{1}{(n+1)^2}.$$

We want the left hand side to be less than $2 - \frac{1}{n+1}$, so it will suffice to establish the inequality

$$2 - \frac{1}{n} + \frac{1}{(n+1)^2} < 2 - \frac{1}{n+1}.$$

Equivalently, it suffices to show

$$\frac{1}{n+1} + \frac{1}{(n+1)^2} < \frac{1}{n}.$$

But we have

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

Everything is positive, so by clearing denominators, the desired inequality is equivalent to

$$n^2 + 2n = n(n+2) < (n+1)^2 = n^2 + 2n + 1,$$

which, at last, is a true inequality. Thus we have all the ingredients of an induction proof, but again we need to put things together in proper order, a task which we leave to the reader.

REMARK 7.15. *Taking limits as $n \to \infty$, it follows that $\sum_{n=1}^{\infty} \frac{1}{n^2} \leq 2$. In particular, this argument shows that the infinite series converges. The exact value of the sum is, in fact, $\frac{\pi^2}{6} \approx 1.64493$* [**Cl-HC**, Thm. 14.1].

9. Extending binary properties to n-ary properties

We begin by trying to prove a very unlikely sounding statement.

Dubious Claim: All horses have the same color.

Proposed proof: There are only finitely many horses in the world, so it will suffice to show that for all $n \in \mathbb{Z}^+$, $P(n)$ holds, where $P(n)$ is the statement that in any set of n horses, all of them have the same color.

Base case: In any set S of one horse, all of the horses in S have the same color!
Induction step: We suppose that for some $n \in \mathbb{Z}^+$, in any set of n horses, all horses

have the same color. Consider now a set of $n+1$ horses, which for specificity we label $H_1,\ H_2, \ldots, H_n,\ H_{n+1}$. Now we can split this into two sets of n horses:

$$S = \{H_1, \ldots,\ H_n\}$$

and

$$T = \{H_2, \ldots, H_n,\ H_{n+1}\}.$$

By induction, every horse in S has the same color as H_1: in particular H_n has the same color as H_1. Similarly, every horse in T has the same color as H_n: in particular H_{n+1} has the same color as H_n. But this means that $H_2, \ldots, H_n, H_{n+1}$ all have the same color as H_1. It follows by induction that for all $n \in \mathbb{Z}^+$, in any set of n horses, all have the same color.

Proof analysis: Naturally one suspects that there is a mistake somewhere, and there is. However it is subtle, and occurs in a perhaps unexpected place. In fact the argument is completely correct *except* that the induction step is not valid when $n = 1$: in this case $S = \{H_1\}$ and $T = \{H_2\}$ and these two sets are disjoint: they have no horses in common. We have been misled by the "dot dot dot" notation which suggests, erroneously, that S and T must have more than one element.

In fact, if only we could establish the argument for $n = 2$, then the proof goes through just fine. For instance, the result can be fixed as follows: if in a finite set of horses, any two have the same color, then they all have the same color.

There is a moral here: one should pay especially close attention to the smallest values of n to make sure that the argument has no gaps. On the other hand, there is a certain type of induction proof for which the $n = 2$ case is the most important (often it is also the base case, but not always), and the induction step is easy to show, but uses once again the $n = 2$ case. Here are some examples of this.

We begin with the following result, which we stated in §5.6:

THEOREM 7.16 (n-fold Euclid's Lemma). *Let p be a prime number, let $n \in \mathbb{Z}^+$, and let $a_1, \ldots, a_n \in \mathbb{Z}$. If $p \mid a_1 \cdots a_n$ then $p \mid a_i$ for some $1 \leq i \leq n$.*

PROOF. Let p be a prime number. We go by induction on n.
Base Case 1: $n = 1$. This is trivial: if $p \mid a_1$, then $p \mid a_1$.
Base Case 2: $n = 2$. Euclid's Lemma ((Theorem 5.19) says that for $a, b \in \mathbb{Z}$, if $p \mid ab$ then $p \mid a$ or $p \mid b$. This is precisely the $n = 2$ case of the present result, just with the integers called a_1, a_2 instead of a, b.
Inductive Step: Let $n \geq 2$, and suppose that for all prime numbers p and all integers $a_1, \ldots, a_n$, if $p \mid a_1 \cdots a_n$ then $p \mid a_i$ for some $1 \leq i \leq n$. Now let p be a prime number and let $a_1, \ldots, a_{n+1} \in \mathbb{Z}$ be such that $p \mid a_1 \cdots a_{n+1}$. Then

$$p \mid (a_1 \cdots a_n)(a_{n+1}),$$

so applying Euclid's Lemma with $a = a_1 \cdots a_n$ and $b = a_{n+1}$, we get $p \mid (a_1 \cdots a_n)$ or $p \mid a_{n+1}$. By induction, if $p \mid a_1 \cdots a_n$ then $p \mid a_i$ for some $1 \leq i \leq n$, so overall we get $p \mid a_i$ for some $1 \leq i \leq n+1$, completing the induction step. □

The characteristic features of this kind of induction proof are: (i) either we want to prove $P(n)$ for all $n \geq 1$ and the statement $P(1)$ is absolutely trivial or we want

to prove $P(n)$ for all $n \geq 2$; and (ii) all of the content resides in establishing $P(2)$. The inductive argument for $P(n) \implies P(n+1)$ for all $n \geq 2$ is very straightforward.

The following very special case of Theorem 7.16 is already of interest.

COROLLARY 7.17. *Let p be a prime, $n \in \mathbb{Z}^+$ and $a \in \mathbb{Z}$ be such that $p \mid a^n$. Then $p \mid a$.*

PROOF. Apply Theorem 7.16 with $a_1 = \ldots a_n = n$. □

In Exercise 7.13 you are asked to use Corollary 7.17 to show that for all $n \geq 2$ and all prime numbers p, we have $p^{1/n} \notin \mathbb{Q}$.

PROPOSITION 7.18. *Let $n \in \mathbb{Z}^{\geq 2}$, and let $f_1, \ldots, f_n : \mathbb{R} \to \mathbb{R}$ be differentiable functions. Then:*

$$(f_1 \cdots f_n)' = f_1' f_2 \cdots f_n + f_1 f_2' \cdots f_n + \ldots + f_1 \cdots f_{n-1} f_n'.$$

PROOF. We go by induction on n.
Base case ($n = 2$): The assertion is $(f_1 f_2)' = f_1' f_2 + f_1 f_2'$, which is the product rule from differential calculus [**Cl-HC**, Thm. 5.6].
Induction step: We assume the result is true for any n differentiable functions. If $f_1, \ldots, f_{n+1}$ are all differentiable, then

$$(f_1 \cdots f_n f_{n+1})' = ((f_1 \cdots f_n) f_{n+1})' \stackrel{*}{=} (f_1 \cdots f_n)' f_{n+1} + f_1 \cdots f_n f_{n+1}' =$$

$$(f_1' f_2 \cdots f_n) f_{n+1} \stackrel{**}{=} f_1 f_2' f_3 \cdots f_n f_{n+1} + \ldots + f_1 \cdots f_{n-1} f_n' f_{n+1} + f_1 \cdots f_n f_{n+1}'.$$

In the first starred equality we have applied the usual product rule and in the second starred equality we have applied the induction hypothesis. □

COROLLARY 7.19. *For all $n \in \mathbb{Z}^+$, if $f(x) = x^n$, then $f'(x) = nx^{n-1}$.*

You are asked to prove Corollary 7.19 in Exercise 7.14.

When teaching freshman calculus, it is very frustrating not to be able to prove the power rule by this simple inductive argument!

10. Miscellany

PROPOSITION 7.20. *Let $f(x) = e^{x^2}$. Then for all $n \in \mathbb{Z}^+$ there exists a polynomial $P_n(x)$, of degree n, such that*

$$\frac{d^n}{dx^n} f(x) = P_n(x) e^{x^2}.$$

PROOF. By induction on n.
Base case ($n = 1$):
$\frac{d}{dx} e^{x^2} = 2x e^{x^2} = P_1(x) e^{x^2}$, where $P_1(x) = 2x$, a degree one polynomial.

Inductive step: Assume that for some positive integer n there exists $P_n(x)$ of degree n such that $\frac{d^n}{dx^n} e^{x^2} = P_n(x) e^{x^2}$. So $\frac{d^{n+1}}{dx^{n+1}} e^{x^2} =$

$$\frac{d}{dx} \frac{d^n}{dx^n} e^{x^2} \stackrel{\text{IH}}{=} \frac{d}{dx} P_n(x) e^{x^2} = P_n'(x) e^{x^2} + 2x P_n(x) e^{x^2} = \left(P_n'(x) + 2x P_n(x)\right) e^{x^2}.$$

Now, since $P_n(x)$ has degree n, $P_n'(x)$ has degree $n-1$ and $2xP_n(x)$ has degree $n+1$. If f and g are two polynomials such that the degree of f is different from the degree of g, then $\deg(f+g) = \max(\deg(f), \deg(g))$. In particular, $P_{n+1}(x) := P_n'(x) + 2xP_n(x)$ has degree $n+1$, completing the proof of the induction step. □

PROPOSITION 7.21. *For all $n \in \mathbb{N}$, we have $\int_0^\infty x^n e^{-x} dx = n!$.*

PROOF. By induction on n.
Base case ($n = 0$): We have

$$\int_0^\infty e^{-x} = -e^{-x}\Big|_0^\infty = -e^{-\infty} - (-e^0) = -0 - (-1) = 1 = 0!$$

Induction step: let $n \in \mathbb{N}$ and assume $\int_0^\infty x^n e^{-x} dx = n!$. Now to make progress in evaluating $\int_0^\infty x^{n+1} e^{-x} dx$, we integrate by parts, taking $u = x^{n+1},\ dv = e^{-x} dx$. Then $du = (n+1)x^n dx,\ v = -e^{-x}$, and

$$\int_0^\infty -x^{n+1} e^{-x} dx = -x^{n+1} e^{-x}\Big|_0^\infty - \int_0^\infty (-e^{-x}(n+1)x^n) dx$$

$$= (0 - 0) + (n+1)\int_0^\infty x^n e^{-x} dx \stackrel{\text{IH}}{=} (n+1)n! = (n+1)!$$

To evaluate the improper integral at ∞ we used $\lim_{x\to\infty} \frac{(n+1)x^n}{e^x} = 0$, as established in Exercise 7.15. □

REMARK 7.22. *For any $x \in (0, \infty)$, one can define*

$$\Gamma(x) \coloneqq \int_0^\infty t^{x-1} e^{-t} dt.$$

(In fact this is an improper integral*: for $x \geq 1$ the function t^{x-1} is continuous at $x = 0$ so one only needs to check that $\lim_{A\to\infty} 0^A t^{x-1} e^{-t} dt$ exists, which is not so bad: e.g. one could do a comparison to $\int_0^\infty e^{-t/2}$. When $x < 1$ one also needs to show that $\int_0^1 t^{x-1} e^{-t}$ is convergent, which one can do by comparison to $\int_0^1 \frac{dx}{x^p}$.) Proposition 7.21 then shows:*

$$\forall n \in \mathbb{N},\ \Gamma(n+1) = n!$$

So the Gamma function gives a natural – maybe not at first, but it gets justified – continuous interpolation of the factorial function (with a small shift, which is annoying at first, but one gets used to it). It is useful in various parts of analysis and number theory. One justification for this is ***Stirling's Formula*** [**R**, p. 194]*:*

$$\lim_{x\to\infty} \frac{\Gamma(x+1)}{(x/e)^x \sqrt{2\pi x}} = 1. \tag{33}$$

This gives the asymptotic behavior of the factorial function:

$$\lim_{n\to\infty} \frac{n!}{(n/e)^n \sqrt{2\pi n}} = 1.$$

11. The Principle of Strong/Complete Induction

PROBLEM: A sequence is defined recursively by

$$\begin{cases} a_1 \coloneqq 1 \\ a_2 \coloneqq 2 \\ \forall n \geq 3,\ a_n \coloneqq 3a_{n-1} - 2a_{n-2} \end{cases}.$$

Find a general formula for a_n and prove it by induction.

ANALYSIS: Unless we know something better, we may as well examine the first few terms of the sequence and hope that a pattern jumps out at us. We have

$$a_3 = 3a_2 - 2a_1 = 3 \cdot 2 - 2 \cdot 1 = 4.$$
$$a_4 = 3a_3 - 2a_2 = 3 \cdot 4 - 2 \cdot 2 = 8.$$
$$a_5 = 3a_4 - 2a_3 = 3 \cdot 8 - 2 \cdot 4 = 16.$$
$$a_6 = 3a_5 - 2a_4 = 3 \cdot 16 - 2 \cdot 8 = 32.$$

The evident guess is therefore $a_n = 2^{n-1}$. Now a key point: it is not possible to prove this formula using the version of mathematical induction we currently have. Indeed, let's try: assume that $a_n = 2^{n-1}$. Then

$$a_{n+1} = 3a_n - 2a_{n-1}.$$

By the induction hypothesis we can replace a_n with 2^{n-1}, getting

$$a_{n+1} = 3 \cdot 2^{n-1} - 2a_{n-1};$$

now what?? A little reflection shows that *we want* $a_{n-1} = 2^{n-2}$. If for some reason it were logically permissible to make that substitution, we'd be in good shape:

$$a_{n+1} = 3 \cdot 2^{n-1} - 2 \cdot 2^{n-2} = 3 \cdot 2^{n-1} - 2^{n-1} = 2 \cdot 2^{n-1} = 2^n = 2^{(n+1)-1},$$

which is what we wanted to show. Evidently this goes a bit beyond the type of induction we have seen so far: in addition to assuming the truth of a statement $P(n)$ and using it to prove $P(n+1)$, we also want to assume the truth of $P(n-1)$.

There is a version of induction that allows this, and more. To ease the reader into it, we will start with the version for $\mathbb{Z}^+$ rather than the slightly more general version for $\mathbb{Z}^{\geq N}$ for some integer N.

THEOREM 7.23 (Principle of Strong/Complete Induction for Sentences). *Let $P(n)$ be an open sentence with domain $n \in \mathbb{Z}^+$. Suppose that:*

(SI1) *$P(1)$ is true, and*

(SI2) *For all $n \in \mathbb{Z}^+$, if all of $P(1), P(2), \ldots, P(n)$ are true, then $P(n+1)$ is true.*

Then $P(n)$ is true for all $n \in \mathbb{Z}^+$.

PROOF. We will give two proofs. The first deduces Theorem 7.23 from the Well-Ordering Principle, while the second deduces it from Theorem 7.3.

FIRST PROOF: Assume that (SI1) and (SI2) hold, and seeking a contradiction, suppose that it is not the case that $P(n)$ holds for all $n \in \mathbb{Z}^+$. Then $T \coloneqq \{n \in \mathbb{Z}^+ \mid P(n) \text{ is false}\}$ is nonempty, so by Well-Ordering has a minimum element A. We cannot have $A = 1$, since $P(1)$ is true by (SI1). Therefore $A \geq 2$ and $P(1), \ldots, P(A-1)$ are all true. But then, by (SI2) we must have that $P(A)$ is true, contradicting the fact that $A \in T$.

SECOND PROOF: Assume that (SI1) and (SI2). For $n \in \mathbb{Z}^+$, we put

$$Q(n) \coloneqq P(1) \wedge P(2) \wedge \ldots \wedge P(n).$$

That is, $Q(n)$ asserts that $P(k)$ is true for all k betweeen 1 and n. Then $Q(1) = P(1)$, which is true by (SI1). Moreover, if $Q(n)$ is true, then $P(1), \ldots, P(n)$ are true, so by (SI2) we have that $P(n+1)$ is true. Thus overalll we have that $P(1), \ldots, P(n+1)$ are true, so $Q(n+1)$ is true. By Theorem 7.3 we deduce that $Q(n)$ holds for all $n \in \mathbb{Z}^+$. Since $Q(n) \implies P(n)$, we have that $P(n)$ holds for all $n \in \mathbb{Z}^+$. □

In the exercises you are asked to prove a version of Theorem 7.23 for open sentences $P(n)$ with domain $n \in \mathbb{Z}^{\geq N}$ for some integer N and then show that this result implies Theorem **??**. The idea is that Strong/Complete Induction is "stronger" than Induction because if

$$P(n) \implies P(n+1),$$

then certainly

$$(P(1) \wedge P(2) \wedge \ldots \wedge P(n)) \implies P(n+1):$$

indeed

$$(P(1) \wedge \ldots \wedge P(n)) \implies P(n),$$

and implication is transitive.

Here is an example of a result that we previously proved using Well-Ordering that be just as easily proved by Strong/Complete Induction (but for which a proof by Induction would be somewhat awkward).

PROPOSITION 7.24. *Let $n > 1$ be an integer. Then there exist prime numbers $p_1, \ldots, p_k$ (for some $k \geq 1$) such that $n = p_1 \cdots p_k$.*

PROOF. We go by strong induction on n.
Base case: $n = 2$. Indeed 2 is prime, so we're good.
Induction step: Let $n > 2$ be any integer and assume that the statement is true for all integers $2 \leq k < n$. We wish to show that it is true for n.
Case 1: n is prime. As above, we're good.
Case 2: n is not prime. By definition, this means that there exist integers a, b, with $1 < a, b < n$, such that $n = ab$. But now our induction hypothesis applies to both a and b: we can write $a = p_1 \cdots p_k$ and $b = q_1 \cdots q_l$, where the p_i's and q_j's are all prime numbers. But then

$$n = ab = p_1 \cdots p_k q_1 \cdots q_l$$

is an expression of n as a product of prime numbers: done! □

12. The Fibonacci numbers

We now introduce a very famous sequence, the **Fibonacci numbers**:

$$F_1 = F_2 = 1, \ \forall n \geq 1, F_{n+2} = F_{n+1} + F_n.$$

Let us list some values:

$$F_3 = 2, \ F_4 = 3, \ F_5 = 5, \ F_6 = 8, \ F_7 = 13, \ F_8 = 21, \ F_9 = 34, \ F_{10} = 55,$$
$$F_{11} = 89, \ F_{12} = 144, \ F_{13} = 233, \ F_{14} = 377, \ F_{15} = 377,$$
$$F_{200} = 280571172992510140037611932413038677189525,$$
$$F_{201} = 453973694165307953197296969697410619233826.$$

This partial list suggests that F_n again grows exponentially in n. Indeed, if we compare ratios of successive values, it seems that the base of the exponential is somewhere between 1 and 2. Especially,

$$\frac{F_{201}}{F_{200}} = 1.618033988749894848204586834\ldots.$$

If you happen to be very familiar with numbers, you might just recognize this as the **golden ratio** $\varphi = \frac{1+\sqrt{5}}{2}$.

The truth of this and more comes out of the following remarkable result.

THEOREM 7.25 (Binet's Formula). *For all $n \in \mathbb{Z}^+$, the nth Fibonacci number is*

$$F_n = \frac{1}{\sqrt{5}}\left(\varphi^n - (1-\varphi)^n\right), \tag{34}$$

where $\varphi = \frac{1+\sqrt{5}}{2}$.

PROOF. We go by strong/complete induction on n. The following identities will be useful:

$$\varphi^2 = \varphi + 1,$$
$$(1-\varphi) = -\varphi^{-1},$$
$$1 - \varphi^{-1} = \varphi^{-2} = (-\varphi)^{-2},$$

Base Cases:

• $n = 1$: We have

$$\frac{1}{\sqrt{5}}\left(\varphi - (1-\varphi)\right) = \frac{1}{\sqrt{5}}(2\varphi - 1) = \frac{1}{\sqrt{5}} \cdot \sqrt{5} = 1 = F_1.$$

• $n = 2$: We have

$$\frac{1}{\sqrt{5}}\left(\varphi^2 - (1-\varphi)^2\right) = \frac{1}{\sqrt{5}}\left((\varphi+1) - (-\varphi^{-1})^2\right) = \frac{1}{\sqrt{5}}\left((\varphi+1) - \varphi^{-2}\right)$$
$$= \frac{1}{\sqrt{5}}\left(1 + \varphi - (1-\varphi^{-1})\right) = \frac{1}{\sqrt{5}}\left(1 + \varphi - (1+1-\varphi)\right) = \frac{1}{\sqrt{5}}(2\varphi - 1) = 1 = F_2.$$

Induction Step: Suppose that $n \geq 1$ and the formula holds for all positive integers that are less than $n+2$. Then, using the above identities, we compute

$$F_{n+2} = F_{n+1} + F_n = \frac{1}{\sqrt{5}}(\varphi^{n+1} + \varphi^n - (1-\varphi)^{n+1} - (1-\varphi)^n))$$
$$= \frac{1}{\sqrt{5}}(\varphi^n(\varphi+1) - (1-\varphi)^n(1-\varphi+1) =$$
$$\frac{1}{\sqrt{5}}(\varphi^n(\varphi^2) - (-\varphi)^{-n}((-\varphi)^{-1}+1)$$
$$= \frac{1}{\sqrt{5}}(\varphi^{n+2} - (-\varphi)^{-n}(-\varphi)^{-2}) = \frac{1}{\sqrt{5}}(\varphi^{n+2} - (-\varphi)^{-(n+2)})$$
$$= \frac{1}{\sqrt{5}}(\varphi^{n+2} - (1-\varphi)^{n+2}).$$ □

The Fibonacci numbers are a remarkably rich source of identities. We restrict ourselves to two examples, with some further identities given in the exericses.

PROPOSITION 7.26. *For all $n \in \mathbb{Z}^+$, we have*

$$F_1 + \ldots + F_n = F_{n+2} - 1. \tag{35}$$

PROOF. By induction on n.

Base Case: We have $F_1 = 1 = 2 - 1 = F_3 - 1$.

Induction Step: Let $n \in \mathbb{Z}^+$ and suppose that $F_1 + \ldots + F_n = F_{n+2} - 1$. Then

$$F_1 + \ldots + F_{n+1} = (F_1 + \ldots + F_n) + F_{n+1} = (F_{n+2} - 1) + F_{n+1}$$
$$= (F_{n+1} + F_{n+2}) - 1 = F_{n+3} - 1 = F_{(n+1)+2} - 1.$$ □

PROPOSITION 7.27 (Cassini Identity). *For all $n \in \mathbb{Z}^+$, we have*

$$F_{n+1}F_{n-1} - F_n^2 = (-1)^n. \tag{36}$$

PROOF. We establish (36) by induction on n.
Base Case: We have

$$F_{1+1}F_{1-1} - F_1^2 = F_2F_0 - F_1^2 = -1 = (-1)^n.$$

Induction Step: Let $n \in \mathbb{Z}^+$, and suppose that (36) holds for n. Our big idea is to add and subtract $F_{n+1}F_n$:

$$(-1)^n = F_{n+1}F_{n-1} - F_n^2 = F_{n+1}F_{n-1} + F_{n+1}F_n - F_n^2 - F_{n+1}F_n$$
$$= F_{n+1}(F_{n-1} + F_n) - F_n(F_n + F_{n+1}) = F_{n+1}^2 - F_{n+2}F_n.$$

Multiplying through by -1, we get

$$F_{n+2}F_n - F_{n+1}^2 = (-1)^{n+1},$$

which is (36) with $n+1$ in place of n, completing the induction step. □

The inductive proof of Cassini's Identity is straightforward but not especially interesting or enlightening. Those who are familiar with matrices and determinants may prefer to use induction to show:

$$\forall n \in \mathbb{Z}^+, \left[\begin{array}{cc} 1 & 1 \\ 1 & 0 \end{array}\right]^n = \left[\begin{array}{cc} F_{n+1} & F_n \\ F_n & F_{n-1} \end{array}\right], \tag{37}$$

from which Cassini's Identity follows upon equating the determinants of both sides. You are asked to do this in Exercise 7.25.

13. Solving Homogeneous Linear Recurrences

Our proof of Binet's Formula works as an application of Strong/Complete Induction. However, I want to call attention to the fact that our treatment, while mathematically correct, contained a significant shortcoming: where on earth did this formula come from?!? In this section we discuss some techniques for finding closed form expressions for certain recursively defined sequences. This continues an exploration begun in §10, since our motivating problem for Strong/Complete Induction concerned a sequence defined by $a_1 = 1$, $a_2 = 2$, and for all $n \geq 1$, $a_n = 3a_{n-1} - 2a_{n-2}$. By trial and error we guessed that $a_n = 2^{n-1}$, and this was easily confirmed using Strong/Complete Induction.

But this was very lucky (or worse: the example was constructed so as to be easy to solve). In general, it might not be so obvious what the answer is, and as above, this is induction's Kryptonite: it has no help to offer in guessing the answer.

Example: Suppose a sequence is defined by $x_0 = 2$, $x_n = 5x_{n-1} - 3$ for all $n \geq 1$.

Here the first few terms of the sequence are $x_1 = 7$, $x_2 = 32$, $x_3 = 157$, $x_4 = 782$, $x_5 = 3907$. What's the pattern? It's not so clear.

This is a case where a bit more generality makes things much clearer: it is often easier to detect a pattern involving algebraic expressions than a pattern involving

integers. So suppose that we have any three real numbers a, b, c, and we define a sequence recursively by $x_0 = c$, $x_n = ax_{n-1} + b$ for all $n \geq 1$. Now let's try again:

$$x_1 = ax_0 + b = ac + b.$$
$$x_2 = ax_1 + b = a(ac + b) + b = ca^2 + ba + b.$$
$$x_3 = ax_2 + b = a(ca^2 + ba + b) + b = ca^3 + ba^2 + ba + b.$$
$$x_4 = ax_3 + b = a(ca^3 + ba^2 + ba + b) + b = ca^4 + ba^3 + ba^2 + ba + b.$$

Aha: it seems that we have for all $n \geq 1$.

$$x_n = ca^n + ba^{n-1} + \ldots + ba + b.$$

Now we have something that induction can help us with: it is true for $n = 1$. Assuming it is true for n, we calculate

$$x_{n+1} = ax_n + b \stackrel{IH}{=} a(ca^n + ba^{n-1} + \ldots + ba + b) + b) = ca^{n+1} + ba^n + \cdots + ba^2 + ba + b,$$

which is what we wanted. So the desired expression is correct for all n. Indeed, we can simplify it:

$$x_n = ca^n + b\sum_{i=1}^{n} a_i = ca^n + b\left(\frac{a^{n+1} - 1}{a - 1}\right) = \frac{(ab + ac - c)a^n - b}{a - 1}.$$

In particular the sequence x_n grows exponentially in n.

Next we wish to find all sequences $\{x_n\}_{n=1}^{\infty}$ satisfying both of the following:

$$x_1 = A_1, x_2 = A_2, \tag{38}$$

$$\forall n \geq 1,\ x_{n+2} = bx_{n+1} + cx_n. \tag{39}$$

Step 1: In light of what we have seen above, it is reasonable to *guess* a nonzero exponential solution: suppose that there is some $r \in \mathbb{R} \setminus \{0\}$ such that $x_n = r^n$ for all $n \in \mathbb{Z}^+$. Let us see what this entails. Via (39) we would get

$$r^{n+2} = x_{n+2} = bx_{n+1} + cx_n = br^{n+1} + cr^n.$$

This last equation is equivalent to

$$r^n(r^2 - br - c) = 0,$$

which, since $r \neq 0$, holds if and only if $r^2 - br - c = 0$. So let us define the **characteristic polynomial**

$$P(x) = x^2 - bx - c.$$

What we found is that $x_n = r^n$ satisfies (39) if and only if r is a root of $P(x)$. Now we have several cases to consider depending upon the behavior of the roots of the characteristic polynomial. The easiest case is when P has distinct nonzero real roots $r_1 \neq r_2$. This is indeed the case that we are in for the Fibonacci numbers: since then $b = c = 1$, the characteristic polynomial is

$$x^2 - x - 1,$$

whose roots are $\frac{1 \pm \sqrt{5}}{2}$, i.e., the golden ratio $\varphi = \frac{1+\sqrt{5}}{2}$ and also $1 - \varphi = \frac{1-\sqrt{5}}{2}$. It is however pretty clear that neither φ^n nor $(1 - \varphi)^n$ gives the Fibonacci number F_n: in fact, neither of these are rational numbers! The point is that the Fibonacci numbers F_n satisfy the recurrence $F_{n+2} = F_{n+1} + F_n$ but also have the initial equations $F_1 = F_2 = 1$. So we need to somehow algebraically combine the two

exponential solutions we found to $x_{n+2} = x_{n+1} + x_n$ in order to satisfy the initial conditions $x_1 = x_2 = 1$.

Step 2: We observe that if $\{x_n\}_{n=1}^\infty$ and $\{y_n\}_{n=1}^\infty$ are two sequences, each of which satisfies the recurrence (39), then for any $\alpha, \beta \in \mathbb{R}$, the linear combination

$$z_n \coloneqq \alpha x_n + \beta y_n$$

satisfies the same recurrence: indeed

$$\begin{aligned} z_{n+2} &= \alpha x_{n+2} + \beta y_{n+2} = \alpha(ax_{n+1} + bx_n) + \beta(ay_{n+1} + by_n) \\ &= a(\alpha x_{n+1} + \beta y_{n+1}) + b(\alpha x_n + \beta y_n) = az_{n+1} + bz_n. \end{aligned}$$

So now we have a clear strategy in the case of distinct nonzero real roots r_1, r_2 : show that we can choose $\alpha, \beta \in \mathbb{R}$ such that

$$z_n = \alpha r_1^n + \beta r_2^n$$

satisfies (38), namely:

$$\begin{aligned} z_1 &= \alpha r_1 + \beta r_2 = A_1, \\ z_2 &= \alpha r_1^2 + \beta r_2^2 = A_2. \end{aligned}$$

Multiplying the first equation by r_1 and subtracting it from the second, we get

$$\beta r_2(r_2 - r_1) = A_2 - r_1 A_1,$$

so

$$\beta = \frac{A_2 - r_1 A_1}{r_2(r_2 - r_1)}.$$

Substituting this back into the first equation, we solve for α, getting

$$\alpha = \frac{r_2 A_1 - A_2}{r_1(r_2 - r_1)}.$$

Thus in the case that $P(x)$ has distinct nonzero real roots, the unique solution to (38) and (39) is

$$x_n = \frac{r_2 A_1 - A_2}{r_1(r_2 - r_1)} r_1^n + \frac{A_2 - r_1 A_1}{r_2(r_2 - r_1)} r_2^n. \tag{40}$$

In the case of Fibonacci numbers, we have $r_1 = \varphi$, $r_2 = 1 - \varphi$, $A_1 = A_2 = 1$, so

$$\begin{aligned} F_n &= \left(\frac{(1-\varphi) - 1}{\varphi(1 - \varphi - \varphi)}\right)\varphi^n + \left(\frac{1 - \varphi}{(1-\varphi)(1 - \varphi - \varphi)}\right)(1-\varphi)^n \\ &= \left(\frac{-1}{1 - 2\varphi}\right)\varphi^n + \left(\frac{1}{1 - 2\varphi}\right)(1-\varphi)^n. \end{aligned}$$

Since

$$\frac{1}{1 - 2\varphi} = \frac{1}{1 - (1 + \sqrt{5})} = \frac{-1}{\sqrt{5}},$$

we finally recover **Binet's Formula**:

$$F_n = \frac{1}{\sqrt{5}}\varphi^n - \frac{1}{\sqrt{5}}(1-\varphi)^n.$$

Now we consider some other possibilities for the characteristic polynomial $P(x) = x^2 - bx - c$. Above did not allow 0 as a root. Well, what happens if 0 is a root? This occurs if and only if $c = 0$, in which case the other root is b. If $c = 0$ and $b \neq 0$ then we get $x_{n+2} = bx_{n+1}$ for all $n \geq 1$. This is a much simpler recursion: it is just telling us that starting at $n = 2$, the ratio of any two consecutive terms

is b. So since $x_2 = A_2$, we have $x_n = A_2 b^{n-2}$ for all $n \geq 2$. This is also a special case of the one-term linear recursion we studied at the beginning of this section. If $b = c = 0$, things get more degenerate still: we just have $x_n = 0$ for all $n \geq 3$.

Next we consider the case in which the characteristic polynomial $P(x)$ has non-real roots: thus

$$P(x) = (x - r_1)(x - r_2)$$

with $r_1 = a + bi$, $r_2 = a - bi$ with $a \in \mathbb{R}$, $b \in \mathbb{R} \setminus \{0\}$. At lesat pending some comfort with complex numbers (which we do not assume; indeed this is the only part of the text in which we make any meaningful use of them), one sees that the entire discussion could have taken place over $\mathbb{C}$ instead of $\mathbb{R}$ without any change whatsoever. That is, $b, c, A_1, A_2 \in \mathbb{C}$, if $P(x) = x^2 - bx - c$ has distinct roots in $\mathbb{C}$, there is a unique complex sequence $\{x_n\}_{n=1}^{\infty}$ satisfying (38) and (39), still given by the same formula

$$x_n = \frac{r_2 A_1 - A_2}{r_1(r_2 - r_1)} r_1^n + \frac{A_2 - r_1 A_1}{r_2(r_2 - r_1)} r_2^n. \tag{41}$$

There is something a bit strange about this formula, which we will draw out in the following example.

Example 7.28. *We consider the recurrence*

$$x_1 = 1,\ x_2 = 7,\ \forall n \geq 1,\ x_{n+2} = -x_n.$$

In this case just a little a little thought shows that the sequence of odd-numbered terms alternates between 1 *and* -1*, while the sequence of even-numbered terms alternates between* 7 *and* -7*, so overall the sequence is*

$$1, 7, -1, -7, 1, 7, -1, -7, \ldots$$

Let us nevertheless solve this using the method we just introduced. Since $b = 0$ *and* $c = -1$*, the characteristic polynomial is*

$$P(x) = x^2 + 1 = (x + i)(x - i),$$

so the roots are $r_1 = i$ *and* $r_2 = -i$*. We have* $A_1 = 1$ *and* $A_2 = 7$*, so (41) gives*

$$x_n = \left(\frac{-i - 7}{i(-i - i)}\right) i^n + \left(\frac{7 - i}{-i(-i - i)}\right) (-i)^n,$$

so, simplifying, we get

$$x_n = \frac{1}{2}\left((-7 - i)i^n + (-7 + i)(-i)^n\right). \tag{42}$$

Just as in Binet's Formula, where at first it looks unlikely that the right hand side is an integer, let alone the n*th Fibonacci number, in (42) it is not immediately clear that the right hand side is even a real number. However, every complex number* $z = a + bi$ *has a complex conjugate* $\overline{z} = a - bi$*. Since a complex number is real if and only if its imaginary part (the coefficient of* i*) is* 0*, we find that*

$$z = a + bi \in \mathbb{R} \iff b = 0 \iff a + bi = a - bi \iff z = \overline{z}.$$

Therefore for any complex number z*, the sum* $z + \overline{z}$ *is real, since complex conjugation switches the two and thus preserves the sum.*

It still requires some case analysis to check directly that (42) gives 1*, then* 7*, then* -1*, then* -7*, and so forth: we leave this to the reader.*

Do we have to prove in general that the right hand side of (41) is a real number? No, our discussion shows that there is a unique sequence of complex numbers x_n satisfying any two term homogeneous linear recursion $x_{n+2} = bx_{n+1} + c$ with $b, c \in \mathbb{C}$ and subject to the initial conditions $x_1 = A_1 \in \mathbb{C}$, $x_2 = A_2 \in \mathbb{C}$, and that the formula for x_n is given by (41). Moreover, if $b, c, A_1, A_2 \in \mathbb{R}$ then it must be the case that every x_n lies in $\mathbb{R}$: this follows immediately by Strong/Complete Induction. So if $, b, c, A_1, A_2 \in \mathbb{R}$ and r_1 and r_2 are the complex roots of $P(x) = x^2 - bx - c$, then – hey presto – it must be the case that $\frac{r_2 A_1 - A_2}{r_1(r_2 - r_1)} r_1^n + \frac{A_2 - r_1 A_1}{r_2(r_2 - r_1)} r_2^n \in \mathbb{R}$ for all positive integers n. Nevertheless you might feel comforted to prove this directly by a complex conjugation argument: we ask the interested reader to take this up in Exercise 7.30.

Finally, we consider the case in which the characteristic polynomial is of the form

$$x^2 - bx - c = P(x) = (x - r)^2$$

for some nonzero real number r: i.e., we have a repeated nonzero root. Since $(x - r)^2 = x^2 - 2rx + r^2$ we must have

$$b = 2r, \ c = -r^2,$$

so (39) becomes

$$x_{n+2} = 2rx_{n+1} - r^2 x_n.$$

In this case the issue is that whereas above we found two "fundamental" solutions r_1^n and r_2^n, since we have only one root we now only have one fundamental solution r^n. Our task is to find another fundamental solution that is not simply a scalar multiple of the first one. Lacking expository inspiration, I will just write down a second fundamental solution, namely

$$y_n = nr^n.$$

Indeed, we have

$$2ry_{n+1} - r^2 y_n = 2r(n+1)r^{n+1} - nr^2 r^n = r^{n+2}(2n + 2 - n) = (n+2)r^{n+2} = y_{n+2}.$$

Again we claim that every solution to (39) in this case is of the form

$$z_n = \alpha r^n + \beta n r^n$$

for $\alpha, \beta \in \mathbb{R}$: again this amounts to the simple linear algebraic fact that for real numbers A_1, A_2, there are unique α, β such that

$$A_1 = z_1 = \alpha r + \beta r$$

$$A_2 = z_2 = \alpha r^2 + \beta(2r^2).$$

Multiplying the first equation by r and subtracting it from the second, we get

$$\alpha = \frac{2rA_1 - A_2}{r^2}, \ \beta = \frac{A_2 - rA_1}{r^2}.$$

EXAMPLE 7.29. *Consider the sequence defined by*

$$x_1 = 1, \ x_2 = 2, \ \forall n \geq 1, \ x_{n+2} = 2x_{n+1} - x_n.$$

The characteristic polynomial is $P(x) = x^2 - 2x + 1 = (x-1)^2$*, so we have a repeated root of* $r = 1$*. Therefore the closed form is*

$$\forall n \in \mathbb{Z}^+, \ x_n = \left(\frac{2 \cdot 1 \cdot 1 - 2}{1^2}\right) 1^n + \left(\frac{2 - 1 \cdot 1}{1^2}\right) n \cdot 1^n = n.$$

This discussion extends in a rather straightforward manner to all homogeneous linear recursions, i.e., in which for some $d \geq 1$ we define $x_{n+d} = a_{d-1}x_{n+d-1} + \ldots + a_0 x_{n+1}$ for some $a_1, \ldots, a_d \in \mathbb{R}$. There is again a characteristic polynomial $P(x) = x^d - a_{d-1}x^{d-1} - \ldots - a_0$, and one builds d linearly independent solutions using its complex roots. There are three points worth mentioning:

(i) We need to know that every degree d polynomial $P(x)$ with complex coefficients has d complex roots, counted with multiplicity (in other words, there are complex numbers $r_1, \ldots, r_d$, not necessarily distinct, such that $P(x) = (x - r_1) \cdots (x - r_d)$). Whereas it is essentially built into the definition of complex numbers that every real number has a complex square root and thus every quadratic polynomial with real coefficients has complex roots, the fact that every complex polynomial has a complex root is a major result, the **Fundamental theorem of Algebra** (see e.g. [**Cl-HC**, Thm. 15.13] or [**Cl-FT**, Thm. 14.31]).

(ii) If $P(x)$ has d distinct roots, $r_1, \ldots, r_d$, then the fundamental solutions are indeed just $r_1^n, \ldots, r_d^n$. If a root occurs with mutiplicity m – i.e., $x - r$ occurs exactly m times as a factor of $P(x)$ – then we need to find m fundamental solutions involving that root. Above we encountered this in the case $m = 2$ and found that r^n and nr^n were solutions. It turns out that if the root occurs with multiplicity m that $r^n, nr^n, n^2r^n, \ldots, n^{m-1}r^n$ are all solutions.

(iii) Finally, one must show that linear combinations of the d fundamental solutions give all solutions and give expressions for the coefficients of the linear combination in terms of the initial conditions $x_1 = A_1, \ldots, x_d = A_d$. Whereas when $d = 2$ we just had a system of two linear equations in two unknowns to solve so could do it directly, now we have a system of d linear equations in d unknowns which we want to have a unique solution. That is, a certain $d \times d$ matrix must be shown to be invertible. This is best left to a linear algebra course.

The considerations of this section will be eerily familiar to those who have studied homogeneous linear differential equations. For a more systematic exposition on "discrete analogues" of calculus concepts (with applications to the determination of power sums as in §3), see [**Cl-DC**].

14. Upward-Downward Induction

PROPOSITION 7.30. *(Upward-Downward Induction) Let $P(x)$ be a sentence with domain the positive integers. Suppose that:*
(i) For all $n \in \mathbb{Z}^+$, $P(n+1)$ is true $\implies$ $P(n)$ is true, and
(ii) For every $n \in \mathbb{Z}^+$, there exists $N > n$ such that $P(N)$ is true.
Then $P(n)$ is true for all positive integers n.

PROOF. Let S be the set of positive integers n such that $P(n)$ is false. Seeking a contradiction we suppose that S is nonempty. Then by Well-Ordering S has a least element n_0. By condition (ii) there exists $N > n_0$ such that $P(N)$ is true.

Now an inductive argument using condition (i) shows that $P(N)$ is true for all positive integers less than N. To be formal about it, for any negative integer let $P(n)$ be any true statement (e.g. $1 = 1$). Then, for $n \in \mathbb{N}$, define $Q(n) = P(N - n)$.

Then $Q(0) = P(N)$ holds, and for all $n \in \mathbb{N}$, if $Q(n) = P(N-n)$ holds, then by (ii) $P(N-(n+1)) = Q(n+1)$ holds, so by induction $Q(n)$ holds for all n, which means that $P(n)$ holds for all $n < N$.

In particular $P(n_0)$ is true, contradiction. □

It is not every day that one proves a result by Upward-Downward Induction. But there are a few nice applications of it, including the following argument of Cauchy.

THEOREM 7.31. *(Arithmetic-Geometric Mean Inequality) Let $n \in \mathbb{Z}^+$ and let $a_1, \ldots, a_n$ be positive real numbers. Then:*

$$(a_1 \cdots a_n)^{\frac{1}{n}} \leq \frac{a_1 + \ldots + a_n}{n}. \tag{43}$$

Equality holds in (43) iff $a_1 = \ldots = a_n$.

PROOF. Step 0: We will prove the result by Upward-Downward Induction on n. For $n \in \mathbb{Z}^+$ let $P(n)$ be the statement of the theorem. Then we will show:
• $P(1)$ and $P(2)$ hold.
• For all $n \in \mathbb{Z}^+$, if $P(n)$ holds, then $P(2n)$ holds.
• For all $n > 1$, if $P(n)$ holds then $P(n-1)$ holds.
By Proposition 7.30 this suffices to prove the result.
Step 1 (Base Cases): $P(1)$ is simply the assertion that $a_1 = a_1$, which is indeed true. Now let a_1, a_2 be any two positive numbers. Then

$$\left(\frac{a_1 + a_2}{2}\right)^2 - a_1 a_2 = \frac{a_1^2 + 2a_1a_2 + a_2^2}{4} - \frac{4a_1a_2}{4} = \frac{(a_1 - a_2)^2}{4} \geq 0,$$

with equality iff $a_1 = a_2$. This proves $P(2)$.
Step 2 (Doubling Step): Suppose that for some $n \in \mathbb{Z}^+$ $P(n)$ holds, and let $a_1, \ldots, a_{2n}$ be any positive numbers. Applying $P(n)$ to the n positive numbers $a_1, \ldots, a_n$ and then to the n positive numbers $a_{n+1}, \ldots, a_{2n}$ we get

$$a_1 + \ldots + a_n \geq n\,(a_1 \cdots a_n)^{\frac{1}{n}}$$

and

$$a_{n+1} \ldots + a_{2n} \geq n\,(a_{n+1} \cdots a_{2n})^{\frac{1}{n}}.$$

Adding these inequalities together gives

$$a_1 + \ldots + a_{2n} \geq n\left((a_1 \cdots a_n)^{\frac{1}{n}} + (a_{n+1} \cdots a_{2n}\right)^{\frac{1}{n}}.$$

Now apply P(2) with $\alpha = (a_1 \cdots a_n)^{\frac{1}{n}}$ and $\beta = (a_{n+1} \cdots a_{2n})^{\frac{1}{n}}$ to get

$$n(a_1 \cdots a_n)^{\frac{1}{n}} + n(a_{n+1} \cdots a_{2n})^{\frac{1}{n}} = n(\alpha + \beta) \geq 2n(\sqrt{\alpha\beta})$$

$$= 2n\,(a_1 \cdots a_{2n})^{\frac{1}{2n}},$$

so

$$\frac{a_1 + \ldots + a_{2n}}{2n} \geq (a_1 \cdots a_{2n})^{\frac{1}{2n}}.$$

Also equality holds iff $a_1 = \ldots = a_n$, $a_{n+1} = \ldots = a_{2n}$ and $\alpha = \beta$ iff $a_1 = \ldots = a_{2n}$.
Step 3 (Downward Step): Let $n > 1$ and suppose $P(n)$ holds. Let $a_1, \ldots, a_{n-1}$ be

any positive numbers, and put $s = a_1 + \ldots + a_{n-1}$, $a_n = \frac{s}{n-1}$. Applying the result with $a_1, \ldots, a_n$ we get

$$a_1 + \ldots + a_n = s + \frac{s}{n-1} = \left(\frac{n}{n-1}\right) s \geq n \left(\frac{a_1 \cdots a_{n-1} s}{n-1}\right)^{\frac{1}{n}},$$

so

$$s^{\frac{n-1}{n}} \geq (n-1)^{\frac{n-1}{n}} (a_1 \cdots a_{n-1})^{\frac{1}{n}}$$

and thus

$$a_1 + \ldots + a_{n-1} = s \geq (n-1)(a_1 \cdots a_{n-1})^{\frac{1}{n-1}}.$$

We have equality iff $a_1 = \ldots = a_n$ iff $a_1 = \ldots = a_{n-1}$. □

15. The Fundamental Theorem of Arithmetic Revisited

15.1. Euclid's Lemma and the Fundamental Theorem of Arithmetic.

We recall the following two results, both stated and prove in Chapter 5.

THEOREM 7.32 (Euclid's Lemma). *Let p be a prime number and $a, b \in \mathbb{Z}$. Suppose that $p \mid ab$. Then $p \mid a$ or $p \mid b$.*

THEOREM 7.33. *The factorization of any integer $n > 1$ into primes is unique, up to the order of the factors. Explicitly, suppose that*

$$n = p_1 \cdots p_k = q_1 \cdots q_l,$$

are two factorizations of n into primes, with $p_1 \leq \ldots \leq p_k$ and $q_1 \leq \ldots \leq q_l$. Then $k = l$ and $p_i = q_i$ for all $1 \leq i \leq k$.

Theorem 7.33 is part b) of Theorem 5.28. Part a) of Theorem 5.28 establishes the *existence* of prime factorizations for integers $n > 1$, which was first proof using the Well-Ordering Principle and then again by strong induction (Proposition 7.24). The proof of theorem 5.28b) that we gave above was a quick consequence of Euclid's Lemma (more precisely, of its $n-$fold generalization, Theorem 7.16, which as we saw, follows from Euclid's Lemma via a particularly simple inductive argument). However, our proof of Euclid's Lemma was based on a rather lengthy discussion of division, the Euclidean algorithm, writing the greatest common divisor as a linear combination, and so forth: taking all the intermediate results together, we spent more time proving it than perhaps any other result in this text.

In this section we give a different approach that showcases the power of induction. First is a fundamental observation: given the easier result on the *existence* of prime factorizations of integers $n > 1$ (which, again, we have proved twice: once using Well-Ordering and once using Strong Induction), Theorem 5.19 (Euclid's Lemma = **EL**) and Theorem 7.33) (the uniqueness part of the Fundamental Theorem of Arithmetic = **FTA**) are *equivalent results*: not just in the logical sense (they are both true after all) but in the stronger sense that each can be easily deduced from the other. We already saw that EL $\implies$ Theorem 7.16 $\implies$ FTA. Conversely:

FTA implies EL: Assume that every integer greater than one factors *uniquely* into a product of primes, and let p be a prime, and let $a, b \in \mathbb{Z}$ be such that $p \mid ab$. If $a = 0$ then $p \mid a$, while if $b = 0$ then $p \mid b$, so we may assume then $a, b \in \mathbb{Z} \setminus \{0\}$. Then using Proposition 5.3 we may assume that $a, b \in \mathbb{Z}^+$. If $a = 1$ then the

hypothesis is $p \mid 1 \cdot b = b$, so $p \mid b$; similarly, if $b = 1$ we get $p \mid a$. So we may assume that $a, b \in \mathbb{Z}^{\geq 2}$ and therefore have unique prime factorizations

$$a = p_1 \cdots p_r, \ b = q_1 \cdots q_s;$$

our assumption that p divides ab means $ab = kp$ for some $k \in \mathbb{Z}^+$ and thus

$$ab = p_1 \cdots p_r q_1 \cdots q_s = kp.$$

The right hand side of this equation shows that p must appear in the prime factorization of ab. Since the prime factorization is unique, we must have at least one p_i or at least one q_j equal to p. In the first case p divides a; in the second case p divides b.

We will now show that *each of* Euclid's Lemma and Theorem 7.33 can be proven directly by induction, without any number-theoretic preliminaries beyond the basic properties of divisibility introduced in §5.2.

15.2. Rogers' Inductive Proof of Euclid's Lemma.

Here is a proof of Euclid's Lemma using the Well-Ordering Principle, following K. Rogers [**Ro63**].

As we saw earlier in the course, one can prove Euclid's Lemma for any particular prime p by consideration of cases. In particular we have already seen that Euclid's Lemma holds for all a and b when $p = 2$, and so forth. So suppose for a contradiction that there exists at least one prime such that Euclid's Lemma does not hold for that prime, and among all such primes, by WOP we consider the least one, say p. What this means that there exist $a, b \in \mathbb{Z}^+$ such that $p \mid ab$ but $p \nmid a$ and $p \nmid b$. Again we apply WOP to choose the least positive integer a such that there exists at least one positive integer b with $p \mid ab$ and $p \nmid a$, $p \nmid b$.

Now consider the following equation:

$$ab = (a - p)b + pb,$$

which shows that $p \mid ab \iff p \mid (a - p)b$. There are three cases:

Case 1: $a - p$ is a positive integer. Then, since $0 < a - p < a$ and a was by assumption the *least* positive integer such that Euclid's Lemma fails for the prime p, we must have that $p \mid a - p$ or $p \mid b$. By assumption $p \nmid b$, so we must have $p \mid a - p$, but then $p \mid (a - p) + p = a$, contradiction!
Case 2: We have $a = p$. But then $p \mid a$, contradiction.
Case 3: We have $a < p$. On the other hand, certainly $a > 1$ – if $p \mid 1 \cdot b$, then indeed $p \mid b$! – so that a is divisible by at least one prime (a consequence of Proposition 7.24) q, and $q \mid a < p$, so $q < p$. Therefore q is a prime which is smaller than the least prime for which Euclid's Lemma fails, so Euclid's Lemma holds for q. Since $p \mid ab$, we may write $pk = ab$ for some $k \in \mathbb{Z}^+$, and now $q \mid a \implies q \mid ab = pk$, so by Euclid's Lemma for q, $q \mid p$ or $q \mid k$. The first case is impossible since p is prime and $1 < q < p$, so we must have $q \mid k$. Therefore

$$p\left(\frac{k}{q}\right) = \left(\frac{a}{q}\right) b,$$

so $p \mid \frac{a}{q}b$. But $1 < \frac{a}{q} < a$ and a is the *least* positive integer for which Euclid's Lemma fails for p and a, so it must be that $p \mid \frac{a}{q}$ (so in particular $p \mid a$) or $p \mid b$. Contradiction. Therefore Euclid's Lemma holds for all primes p.

15.3. The Lindemann-Zermelo Inductive Proof of FTA.

Here is a proof of FTA using the Well-Ordering Principle, following Lindemann [**Li33**] and Zermelo [**Ze34**].

Let us say that a prime factorization $n = p_1 \cdots p_r$ is in **standard form** if $p_1 \leq \ldots p_r$. Every prime factorization becomes a standard form prime factorization upon listing the prime factors in weakly increasing order (i.e., from the smallest to the greatest, with repetitions allowed). We claim that the standard form factorization of a positive integer is unique. Assume not; then the subset of $\mathbb{Z}^{\geq 2}$ of integers that have at least two different standard form factorizations is nonempty, so it has a least element, say n, where:

$$n = p_1 \cdots p_r = q_1 \cdots q_s. \tag{44}$$

Here the p_i's and q_j's are prime numbers, not necessarily distinct from each other. However,we must have $p_1 \neq q_j$ for any j. Indeed, if we had such an equality, then after relabelling the q_j's we could assume $p_1 = q_1$ and then divide through by $p_1 = q_1$ to get a smaller positive integer $\frac{n}{p_1}$. By the assumed minimality of n, the prime factorization of $\frac{n}{p_1}$ must be unique: i.e., $r - 1 = s - 1$ and $p_i = q_i$ for all $2 \leq i \leq r$. But then multiplying back by $p_1 = q_1$ we see that we didn't have two different factorizations after all. (In fact this shows that for all i, j, $p_i \neq q_j$.)

In particular $p_1 \neq q_1$. Without loss of generality, assume $p_1 < q_1$. Then, if we subtract $p_1 q_2 \cdots q_s$ from both sides of (44), we get

$$m := n - p_1 q_2 \cdots q_s = p_1(p_2 \cdots p_r - q_2 \cdots q_s) = (q_1 - p_1)(q_2 \cdots q_s). \tag{45}$$

Evidently $0 < m < n$, so by minimality of n, the prime factorization of m must be unique. However, (45) gives two different factorizations of m, and we can use these to get a contradiction. Specifically, $m = p_1(p_2 \cdots p_r - q_2 \cdots q_s)$ shows that $p_1 \mid m$. Therefore, when we factor $m = (q_1 - p_1)(q_2 \cdots q_s)$ into primes, at least one of the prime factors must be p_1. But $q_2, \ldots, q_j$ are already primes which are different from p_1, so the only way we could get a p_1 factor is if $p_1 \mid (q_1 - p_1)$. But this implies $p_1 \mid q_1$, and since q_1 is also prime this implies $p_1 = q_1$. Contradiction!

16. Exercises

Exercise 7.1. *Let $N \in \mathbb{Z}$.*

a) *Define a subset $S \subseteq \mathbb{Z}^{\geq N}$ to be inductive if $N \in S$ and for all $n \in \mathbb{Z}^{\geq N}$, if $n \in S$ then also $n + 1 \in S$. Use the fact that $\mathbb{Z}^{\geq N}$ is well-ordered to prove that the only inductive subset of $\mathbb{Z}^{\geq N}$ is $\mathbb{Z}^{\geq N}$ itself.*

b) *Use part a) to give another proof of Theorem 7.3.*

Exercise 7.2. *Show: for all $n \in \mathbb{Z}^+$, we have*

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

Exercise 7.3.

a) *An infinite sequence* $\{a_n\}_{n=1}^{\infty}$ *of real numbers is* ***arithmetic*** *if there is* $d \in \mathbb{R}$ *such that for all* $n \in \mathbb{Z}^+$ *we have* $a_{n+1} - a_n = d$. *For such a sequence, show:*

$$\forall n \in \mathbb{Z}^+,\ a_n = a_1 + (n-1)d.$$

b) *An infinite sequence* $\{a_n\}_{n=0}^{\infty}$ *of nonzero real numbers is* ***geometric*** *if there is* $r \in \mathbb{R} \setminus \{0\}$ *such that for all* $n \in \mathbb{N}$ *we have* $\frac{a_{n+1}}{a_n} = r$. *For such a sequence, show:*

$$\forall n \in \mathbb{N},\ a_n = a_0 r^n.$$

EXERCISE 7.4. *Show by induction that for all* $n \in \mathbb{Z}^+$, *we have*

$$1^3 + \ldots + n^3 = \frac{n^4}{4} + \frac{n^3}{2} + \frac{n^2}{4} = \frac{n^2}{4}(n^2 + 2n + 1) = \left(\frac{n(n+1)}{2}\right)^2.$$

EXERCISE 7.5. *We presented a technique for guessing closed form identites for power sums and used the technique to derive a closed form expression for* $\sum_{i=1}^n i^3$. *Use a similar technique to guess a formula for* $\sum_{i=1}^n i^4$ *and then use induction to prove that your formula is correct.*

EXERCISE 7.6. *Prove Theorem 7.10.*

EXERCISE 7.7. *Let* $d, n \in \mathbb{Z}^+$.

a) *Show:*

$$(x\frac{d}{dx})^d(1 + x + \ldots + x^n) = 1^d x + \ldots + n^d x^n.$$

b) *Let* D *be an integer that is greater than* d, *let* $P(x)$ *be a polynomial of degree* D, *and let*

$$f(x) := \frac{P(x)}{(x-1)^d}.$$

Show there is a polynomial $Q(x)$ *of degree* $D + d$ *such that*

$$f'(x) = \frac{Q(x)}{(x-1)^{2d}}.$$

c) *Show: there is a polynomial* $P_{d,n}(x)$ *of degree* $n + 2^d$ *such that*

$$(x\frac{d}{dx})^d\left(\frac{x^{n+1} - 1}{x - 1}\right) = \frac{P_{d,n}(x)}{(x-1)^{2^d}}.$$

d) *Applying l'Hôpital's Rule* 2^d *times , show:*

$$S_d(n) := 1^d + \ldots + n^d = \frac{P_{d,n}^{2^d}(1)}{(2^d)!}.$$

(Make sure to check that l'Hôpital's Rule actually applies.)

EXERCISE 7.8. *We retain the notation of Exercise 7.7*

a) *In §7.7 there are formulas stated for* $P_{2,n}(x)$, *its fourth derivative, and its fourth derivative evaluated at* 1. *Confirm these calculations and thereby complete this proof that* $S_2(n) = 1^2 + \ldots + n^2 = \frac{n(n+1)(2n+1)}{6}$.

b) *Try computing* $P_{3,n}(x)$ *and its eighth derivative by hand. If you find this calculation too onerous, get a mathematical software package to do it for you, and thereby show that* $S_3(n) = 1^3 + \ldots + n^3 = \left(\frac{n(n+1)}{2}\right)^2$.

EXERCISE 7.9. *We retain the notation of Exercise 7.7. Use the formula* $S_d(n) = \frac{P_{d,n}^{2^d}(1)}{(2^d)!}$ *– stated in the text as equation (32) and proved in Exercise* **??** *to give another proof of Theorem 7.10*

EXERCISE 7.10. *Use calculus to show:*

$$\forall x \in \mathbb{R},\ 2^x > x.$$

(Suggestion: it should be helpful to graph the two functions. Of course, merely drawing a picture will not be a sufficient proof.)

EXERCISE 7.11. a) *Show:*[6] *for all* $n \in \mathbb{Z}^+$ *we have*

$$\sum_{k=1}^{2^n} \frac{1}{k} \geq 1 + \frac{n}{2}.$$

b) *Deduce:* $\sum_{k=1}^{\infty} \frac{1}{k} = \infty$.

EXERCISE 7.12.

a) *Prove* ***Bernoulli's Inequality:*** *For all* $x \geq -1$ *and for all* $n \in \mathbb{Z}^+$,

$$(1+x)^n \geq 1 + nx.$$

b) *Show: for all* $n \in \mathbb{Z}^+$, *we have* $\left(1 - \frac{1}{n+1}\right)^n \geq \frac{1}{n+1}$.

EXERCISE 7.13. *Use Corollary 7.17 to show that if* p *is a prime number and* $n \geq 2$, *then* $p^{\frac{1}{n}}$ *is irrational.*

EXERCISE 7.14. *Prove Corollary 7.19.*

EXERCISE 7.15. *Show: for all* $n \in \mathbb{Z}^+$, $\lim_{x \to \infty} \frac{x^n}{e^x} = 0$.
(Suggestion: use induction and L'Hôpital's Rule.[7]*)*

EXERCISE 7.16. *Use induction to prove that every finite nonempty subset* $S \subseteq \mathbb{R}$ *has a minimum element.*
(Suggestion: induct on the size of S*.)*

EXERCISE 7.17. *Proposition 3.5s says: for* $N \geq 2$ *and finite sets* $A_1, \ldots, A_N$ *we have*

$$\#(\prod_{i=1}^{N} A_i) = \prod_{i=1}^{N} \#A_i.$$

We proved this for $N = 2$. *Use induction to deduce the general case from this.*

EXERCISE 7.18. *Let* $N \in \mathbb{Z}$. *Prove the following mild generalization of the Principle of Strong/Complete Induction for Sentences: let* $P(n)$ *be an open sentence with domain* $n \in \mathbb{Z}^{\geq N}$. *Suppose that:*

($\mathrm{SI}_N 1$) $P(N)$ *is true, and*
($\mathrm{SI}_N 2$) *For all* $n \in \mathbb{Z}^{\geq N}$, *if all of* $P(N), P(N+1), \ldots, P(n)$ *are true, then* $P(n+1)$ *is true.*

---

[6] One need not use induction for this, but perhaps it makes the basic idea simpler / clearer.

[7] Math professors tend to roll their eyes at what we view as unnecessary applications of L'Hôpital's Rule. This problem will cause some eyes to roll: there are certainly other ways to establish this. But it makes for a nice induction problem!

Show: $P(n)$ is true for all $n \in \mathbb{Z}^{\geq N}$. (Suggestion: adapt either or both of the proofs of Theorem 7.23.)

EXERCISE 7.19. *Let $N \in \mathbb{Z}$. A subset $S \subseteq \mathbb{Z}^{\geq N}$ is* ***completely inductive*** *if for all $M \in \mathbb{Z}^{\geq N}$, if $\{n \in \mathbb{Z} \mid N \leq n < M\} \subseteq S$, then $M \in S$.*

a) *Show: if S is completely inductive, then $N \in S$.*
b) *Show: a subset $S \subseteq \mathbb{Z}^{\geq N}$ is completely inductive if and only if $S = \mathbb{Z}^{\geq N}$. (Suggestion: apply the Well-Ordering Principle to $T := \mathbb{Z}^{\geq N} \setminus S$.) Explain why this result can be viewed as a* ***Principle of Strong/Complete Induction for Subsets of*** *$\mathbb{Z}^{\geq N}$.*
c) *Show (easily!): an inductive subset $S \subseteq \mathbb{Z}^{\geq N}$ is completely inductive.*
e) *Show: the Principle of Strong/Complete Induction for Subsets of $\mathbb{Z}^{\geq N}$ implies the result of Exercise 7.18.*
f) *Show: Exercise 7.18 implies the Principle of Strong/Complete Induction for Subsets of $\mathbb{Z}^{\geq N}$.*

EXERCISE 7.20. *Recall that we extended the Fibonacci numbers F_n to negative integer indices as well. Find all $n \in \mathbb{Z}$ such that $F_n < 0$.*

EXERCISE 7.21. *Let F be any number system satisfying the field axioms. Show: for all $n \in \mathbb{Z}^+$ and all $x_1, \ldots, x_n \in F$, the following are equivalent:*

(i) *We have $x_1 \cdots x_n = 0$.*
(ii) *We have $x_i = 0$ for some $1 \leq i \leq n$.*

EXERCISE 7.22. *In this exercise, F_n denotes the nth Fibonacci number. Show: for all $n \in \mathbb{Z}^+$, $F_n < 2^n$.*

EXERCISE 7.23. *Suppose you have an enormous supply of red 1 inch by 1 inch square tiles and blue 2 inch by 1 inch rectangular tiles. For $n \in \mathbb{Z}^+$, you are seeking to tile an n inch by 1 inch long rectangular strip by some cominbation of red and blue tiles. Show that the number $T(n)$ of different ways to do this is F_{n+1}, where F_n is the nth Fibonacci number.*
(Suggestion: Show that $T_1 = 1$, $T_2 = 2$ and for all $n \geq 3$ we have $T_n = T_{n-1} + T_{n-2}$.

EXERCISE 7.24. *Show: for all $m, n \in \mathbb{N}$ we have*

$$F_{m+n+1} = F_{m+1}F_{n+1} + F_mF_n.$$

(Suggestion: for each $m \in \mathbb{N}$, prove the result for all $n \in \mathbb{N}$ by induction on n.)

EXERCISE 7.25.

a) *Use induction to prove (37).*
b) *Deduce Cassini's Identity.*

EXERCISE 7.26. *As we explained, writing $F_{n-1} = F_{n+1} - F_n$ allows us to define negatively indexed Fibonacci numbers: e.g.*

$$F_{-1} = F_1 - F_0 = 1,$$
$$F_{-2} = F_0 - F_{-1} = -1,$$
$$F_{-3} = F_{-1} - F - 2 = 2.$$

Show: for all $n \in \mathbb{Z}$ we have $F_{-n} = (-1)^{n+1}F_n$.

EXERCISE 7.27. *In this exercise, F_n denotes the nth Fibonacci number.*

a) *Show: for all $n \in \mathbb{Z}^+$, we have $F_1 + F_3 + \ldots + F_{2n-1} = F_{2n}$.*

b) *Show: for all $n \in \mathbb{Z}^+$, we have $F_2 + F_4 + \ldots + F_{2n} = F_{2n+1} - 1$.*

EXERCISE 7.28 (Lekkerkerker-Zeckendorf Theorem). *Two non-negative Fibonacci numbers F_{n_1} and F_{n_2} are* ***consecutive*** *if there is no Fibonacci number F_{n_3} with $F_{n_1} < F_{n_3} < F_{n_2}$. ($F_1 = 1$ and $F_3 = 2$ are consecutive Fibonacci numbers even though their indices 1 and 3 are not consecutive positive integers.)*

a) *Let $a \in \mathbb{Z}^+$. Show that a is a sum of distinct Fibonacci numbers.*
(Suggestion: use the ***greedy algorithm****. That is, let F_n be the largest Fibonacci number that is less than or equal to a. Show that $a_1 := a - F_n < F_n$, and proceed by induction.)*

b) *Show that a can be written as a sum of distinct Fibonacci numbers, no two of which are consecutive.*
(Suggestion: with notation as in part a), show that $a_1 := a - F_n < F_{n-1}$.)

c) *Show that the expression of a positive integer as a sum of distinct Fibonacci numbers, no two of which are consecutive, is unique.*
(Suggestion: let F_n be the largest Fibonacci number less than or equal to a. Use the previous exercise to show that for any sum $S = F_{i_1} + \ldots + F_{i_k}$ of distinct nonconsecutive Fibonacci numbers, each of which is less than F_n, we have $S < F_n \leq a$. Thus in any representation of a as a sum of distinct nonconsecutive Fibonacci numbers, F_n must appear. Apply Strong Induction.)

EXERCISE 7.29. *Consider the sequence $\{x_n\}_{n=1}^{\infty}$ defined by:*
- *$x_1 = 1$, $x_2 = 2$, and*
- *For all $n \geq 3$, we have $x_n = 2x_{n-1} - x_{n-2}$.*

Show: for all $n \in \mathbb{Z}^+$, we have $x_n = n$.

EXERCISE 7.30. *Let $A_1, A_2 \in \mathbb{R}$, and let $r_1 = a + bi$, $r_2 = a - bi$ be a complex conjugate pair of complex numbers. Show:*

$$\forall n \in \mathbb{Z}^+, \ \frac{r_2 A_1 - A_2}{r_1(r_2 - r_1)} r_1^n + \frac{A_2 - r_1 A_1}{r_2(r_2 - r_1)} r_2^n \in \mathbb{R}.$$

EXERCISE 7.31. *Let $n \in \mathbb{Z}^+$. A popular game called the* ***Towers of Hanoi*** *is played with three tall wooden pegs drilled into a flat surface and n thin wooden disks with a hole drilled through the center, so that the disks can be slid down along any of the pegs to lie flat on the surface. Moreover we suppose the disks are labeled D_1 through D_n and that their radii are strictly increasing. The game begins with all n disks resting on the first peg, with the smallest disk D_1 at the top, followed by the next largest disk D_2, and so forth, with the largest disk D_n at the bottom.*

The object of the game is to transfer all disks from the first tower to one of the other two towers, by a sequence of legal moves*. A legal move consists of selecting any peg that has at least one disk, removing the topmost disk from that peg, and sliding it down one of the other two pegs,* subject to the following condition: *if the peg we are moving our disk to contains any disks already, then our disk must be smaller than the topmost disk on that peg. In other words, what we are* not *allowed to do is move a disk and place it on top of a smaller disk.*

a) *Convince yourself that after any finite sequence of legal moves, each of the three pegs will either be empty or consist of disks that increase in size from top to bottom.*

b) *Show: it is possible to win the game in $2^n - 1$ legal moves.*

c) *Show: it is not possible to win the game in fewer than $2^n - 1$ legal moves.*
d) *Show: there are precisely two sequences of $2^n - 1$ legal moves that win the game: one in which all of the disks end up on the second peg, and one in which all of the disks end up on the third peg. (In other words, if you want to win as fast as possible, your first move doesn't matter, but it determines all your other moves.)*
e) *In case that was too easy...what happens if there are four (or more!) pegs?*

EXERCISE 7.32. *Let $n \in \mathbb{Z}^+$, and let $f, g : \mathbb{R} \to \mathbb{R}$ be n times differentiable functions. Show:*

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

(Recall that for $k \in \mathbb{N}$, we denote by $f^{(k)}$ the kth derivative of f, and by convention we have $f^{(0)} = f$.)

EXERCISE 7.33. *Prove **Lagrange's Identity**: for all $n \in \mathbb{Z}^+$ and for all $(x_1, y_1), \ldots (x_n, y_n) \in \mathbb{R}^2$, we have*

$$\left(\sum_{i=1}^{n} a_i^2\right)\left(\sum_{i=1}^{n} b_i^2\right) - \left(\sum_{i=1}^{n} a_i b_i\right)^2 = \sum_{1 \le i < j \le n} (a_i b_j - a_j b_i)^2 = \frac{1}{2} \sum_{1 \le i,j \le n} (a_i b_j - a_j b_i)^2.$$

CHAPTER 8

Relations and Functions

1. Relations

1.1. The idea of a relation. Let X and Y be two sets. We would like to formalize the idea of a **relation** between X and Y. Intuitively speaking, this is a well-defined "property" R such that given any $x \in X$ and $y \in Y$, either x bears the property R to y, or it doesn't (and not both!). Some important examples:

EXAMPLE 8.1. *Let X be a set of objects and let Y be a set of sets. Then "membership" is a relation R from X to Y: i.e., we have xRy if $x \in y$.*

EXAMPLE 8.2. *Let S be a set, and let $X = Y = 2^S$, the power set of S (recall that this is the set of all subsets of S. Then containment, $A \subseteq B$ is a relation between X and Y. (Proper containment, $A \subsetneq B$, is also a relation.)*

EXAMPLE 8.3. *Let $X = Y$. Then equality is a relation from X to Y: we say xRy iff $x = y$. Also inequality is a relation between X and Y: we say xRy iff $x \neq y$.*

EXAMPLE 8.4. *Let $X = Y = \mathbb{R}$. Then $\leq, <, \geq, >$ are relations between $\mathbb{R}$ and $\mathbb{R}$.*

EXAMPLE 8.5. *For any sets X and Y we have the **full relation** R_F: every element of X relates to every element of Y.*

EXAMPLE 8.6. *Let $X = Y = \mathbb{Z}$. Then divisibility is a relation between $\mathbb{Z}$ and $\mathbb{Z}$: we say xRy if $x \mid y$.*

EXAMPLE 8.7. *Let $X = Y = \mathbb{Z}$. Then "having the same parity" is a relation between $\mathbb{Z}$ and $\mathbb{Z}$.*

In many of the above examples we have $X = Y$. This will often (but certainly not always!) be the case, and when it is we may speak of relations **on X**.

1.2. The formal definition of a relation.

We still have not given a formal definition of a relation between sets X and Y. In fact the above way of thinking about relations is easily formalized, as was suggested by Adam Osborne[1]: namely, we can think of a relation R as a function from $X \times Y$ to the two-element set $\{T, F\}$ (for "true" and "false"). In other words, for $(x, y) \in X \times Y$, we say that xRy if and only if $f((x, y)) = T$.

This is a great way of thinking about relations. But it has one foundational drawback: it makes the definition of a relation depend on that of a function, whereas the

[1]Adam Osborne was a student in my 2009 Math 3200 course at UGA.

standard practice is the reverse: we want to define a function as a kind of relation.

The correspondence between logic and set theory leads us instead to define:

A **relation** R between two sets X and Y is a subset of the Cartesian product $X \times Y$.

(Thus we have replaced the basic logical dichotomy "TRUE/FALSE" with the basic set-theoretic dichotomy "is a member of/is not a member of".) This definition has some geometric appeal: we are essentially identifying a relation R with its *graph* in the sense of precalculus mathematics.

We take advantage of the definition to adjust the terminology: rather than speaking (slightly awkwardly) of relations "from X to Y" we will now speak of relations **on** $\mathbf{X} \times \mathbf{Y}$. When $X = Y$ we may (sometimes) speak of relations **on X**.

EXAMPLE 8.8. *Any curve*[2] *in* $\mathbb{R}^2$ *defines a relation on* $\mathbb{R} \times \mathbb{R}$. *E.g. the unit circle*

$$x^2 + y^2 = 1$$

is a relation in the plane: it is just a set of ordered pairs.

1.3. Basic terminology and further examples. Let X and Y be sets. We consider the set of all relations on $X \times Y$ and denote it by $\mathcal{R}(X, Y)$. According to our formal definition we have

$$\mathcal{R}(X, Y) = 2^{X \times Y},$$

i.e., the set of all subsets of the Cartesian product $X \times Y$.

EXAMPLE 8.9. *Relations on Empty Sets:*

a) *Suppose* $X = \varnothing$. *Then* $X \times Y = \varnothing$ *and* $\mathcal{R}(X \times Y) = 2^{\varnothing} = \{\varnothing\}$. *That is: if* X *is empty, then the set of ordered pairs* (x, y) *for* $x \in X$ *and* $y \in Y$ *is empty, so there is only one relation: the empty relation.*

b) *Suppose* $Y = \varnothing$. *Again* $X \times Y = \varnothing$ *and the discussion is the same as above.*

EXAMPLE 8.10. *Relations on a One Point Set:*

a) *Suppose* $X = \{\bullet\}$ *consists of a single element. Then* $X \times Y = \{(\bullet, y) \mid y \in Y\}$; *in other words,* $X \times Y$ *is essentially just* Y *itself, since the first coordinate is always the same. Thus a relation* R *on* $X \times Y$ *corresponds to a subset of* Y: *formally, the set of all* $y \in Y$ *such that* $\bullet R y$.

b) *Suppose* $Y = \{\bullet\}$ *consists of a single element. The discussion is analogous to that of part a), and relations on* $X \times Y$ *correspond to subsets of* X.

EXAMPLE 8.11. *Counting Relations:*

Suppose X *and* Y *are finite nonempty sets, with* $\#X = m$ *and* $\#Y = n$. *Then* $\mathcal{R}(X, Y) = 2^{X \times Y}$ *is finite, of cardinality*

$$\#2^{X \times Y} = 2^{\#X \times Y} = 2^{\#X \cdot \#Y} = 2^{mn}.$$

The function 2^{mn} *grows rapidly with both* m *and* n, *and the upshot is that if* X *and* Y *are even moderately large finite sets, the set of all relations on* $X \times Y$ *is very large. For instance if* $X = \{a, b\}$ *and* $Y = \{1, 2\}$ *then there are* $2^{2 \cdot 2} = 16$ *relations on* $X \times Y$. *In Exercise 8.1 you are asked to write them all down. However, if*

[2]This is true whatever definition of "curve" one chooses to take.

$X = \{a, b, c\}$ and $Y = \{1, 2, 3\}$ then there are $2^{3 \cdot 3} = 512$ relations on $X \times Y$, and writing them all down is not as easy as the Jackson Five would have us believe.

In Exercise 8.2 you are asked to show that if X and Y are nonempty sets, at least one of which is infinite, then $\mathcal{R}(X, Y)$ is infinite.

Given two relations R_1 and R_2 between X and Y, it makes sense to say that $R_1 \subseteq R_2$: this means that R_1 is "stricter" than R_2 or that R_2 is "more permissive" than R_1. This is a very natural idea: for instance, if X is the set of people in the world, R_1 is the brotherhood relation – i.e., $(x, y) \in R_1$ iff x and y are brothers – and R_2 is the sibling relation – i.e., $(x, y) \in R_2$ iff x and y are siblings – then $R_1 \subsetneq R_2$: if x and y are brothers then they are also siblings, but not conversely.

Among all elements of $\mathcal{R}(X, Y)$, there is one relation $R_\varnothing$ that is the strictest of them all, namely $R_\varnothing = \emptyset$:[3] that is, for *no* $(x, y) \in X \times Y$ do we have $(x, y) \in R_\varnothing$. Indeed $R_\varnothing \subseteq R$ for any $R \in \mathcal{R}(X, Y)$. We call this the **empty relation**. At the other extreme, there is a relation which is the most permissive, namely $R_{X \times Y} = X \times Y$ itself: that is, for *all* $(x, y) \in X \times Y$ we have $(x, y) \in R_{X \times Y}$. This is (still: cf. Example 8.5) called the **full relation** on $X \times Y$. And indeed $R \subseteq R_{X \times Y}$ for any $R \in \mathcal{R}(X, Y)$.

A relation $R \subseteq X \times Y$ is a **function** if for each $x \in X$ there is exactly one $y \in Y$ such that $(x, y) \in R$. We will study functions specifically in the next section.

EXAMPLE 8.12. *Let $X = Y$. The equality relation $\{(x, x) \mid x \in X\}$ can be thought of geometrically as the diagonal of $X \times X$. We often denote it as Δ or Δ_X.*

The **domain** $D(R)$ of a relation $R \subseteq X \times Y$ is the set of $x \in X$ such that there exists $y \in Y$ with $(x, y) \in R$. In other words, it is the set of all elements in x which relate to at least one element of Y.

REMARK 8.13. *To be honest, I am not thrilled with this definition, as it treats X differently from Y, for no apparent reason. Probably the correct thing to do would be to define both a **left domain***

$$D_l(R) \coloneqq \{x \in X \mid \exists\ (x, y) \in R\}$$

*and also a **right domain***

$$D_r(R) \coloneqq \{y \in Y \mid \exists\ (x, y) \in R\}.$$

But in truth, we will not do much with the domain of a relation except in the special case of functions, so we do not really need such refined notation.

EXAMPLE 8.14. *If $R \subseteq X \times Y$ is a function, then its domain is X.*

EXAMPLE 8.15. *The relation $\{(x, y) \in \mathbb{R}^2 \mid x^2 + y^2 = 1\}$ has domain $[-1, 1]$.*

Given a relation $R \subseteq X \times Y$, we can define the **inverse relation** $R^{-1} \subseteq Y \times X$ by interchanging the order of the coordinates. Formally, we put

$$R^{-1} = \{(y, x) \in Y \times X \mid (x, y) \in R\}.$$

[3]The notation here is just to emphasize that we are viewing $\varnothing$ as a relation on $X \times Y$.

Geometrically, this corresponds to reflecting across the line $y = x$. If we did define left and right domains, then taking the inverse would switch them:

$$D_l(R^{-1}) = D_r(R) \text{ and } D_r(R^{-1}) = D_l(R).$$

EXAMPLE 8.16. *Consider the relation $R \subseteq \mathbb{R} \times \mathbb{R}$ given by*

$$R = \{(x, x^2) \mid x \in \mathbb{R}\}.$$

This relation is a function, whose graph is an upward-opening parabola: it can also be described by the equation $y = x^2$. The inverse relation R^{-1} is $\{(x^2, x) \mid x \in \mathbb{R}\}$, which corresponds to the equation $x = y^2$ and geometrically is a parabola opening right-ward. The domain of R is $\mathbb{R}$ (cf. Example 8.14), while the domain of R^{-1} is $[0, \infty)$. Thus R^{-1} is not *a function.*

EXAMPLE 8.17. *Consider the relation*

$$R = \{(x, x^3) \mid x \in \mathbb{R}\}.$$

This relation is a function, more commonly described by the equation $y = x^3$. ITs domain is $\mathbb{R}$. Consider the inverse relation

$$R^{-1} = \{(x^3, x) \mid x \in \mathbb{R}\},$$

which is described by the equation $x = y^3$. Since every real number has a unique real cube root (cf. Exercise 6.10), this is equivalent to $y = x^{\frac{1}{3}}$. This time R^{-1} is again a function, and its domain is $\mathbb{R}$.

When we study functions later in detail, one of our main goals will be to understand the difference between Examples 8.16 and 8.17.

1.4. Properties of relations.

Let X be a set. We now consider various properties that a relation R on X – i.e., a subset $R \subseteq X \times X$ – may or may not possess.

Reflexivity: For all $x \in X$, $(x, x) \in R$.

In other words, each element of X bears relation R to itself. Another way to say this is that the relation R contains the equality relation on X, as above contains the "diagonal"

$$\Delta = \Delta_X \coloneqq \{(x, x) \mid x \in X\}.$$

Anti-reflexivity: For all $x \in X$, $(x, x) \notin R$.

In other words, a relation is anti-reflexive if and only if it is disjoint from the diagonal Δ.

No relation on X is both reflexive and anti-reflexive (except in the trivial case $X = \varnothing$, when both properties hold vacuously). However, when X has more than one element, a relation need not be either reflexive nor anti-reflexive: it may contain some, but not all, points of Δ.

Symmetry: For all $x, y \in X$, if $(x, y) \in R$, then $(y, x) \in R$.

This has a geometric interpretation in terms of symmetry across the diagonal Δ.

For instance, the relation associated to the function $y = \frac{1}{x}$ is symmetric since interchanging x and y changes nothing, whereas the relation associated to the function $y = x^2$ is not. (Looking ahead a bit, a function $y = f(x)$ is symmetric if and only if it coincides with its own inverse function.)

EXAMPLE 8.18. *Let V be a set. A **(simple, loopless, undirected) graph** is given by a relation E on V that is anti-reflexive and symmetric. For $x, y \in V$, we say that x and y are **adjacent** if $(x, y) \in E$. The assumed properties of anti-reflexivity and symmetry amount to: x is never adjacent to itself, and if x is adjacent to y, then y is adjacent to x.*

Anti-Symmetry: for all $x, y \in X$, if $(x, y) \in R$ and $(y, x) \in R$, then $x = y$.

Strong Anti-Symmetry: for all $x, y \in X$, if $(x, y) \in R$, then $(y, x) \notin R$.

In Exercise 8.8 you are asked to show that a relation is strongly anti-symmetric if and only if it is anti-reflexive and anti-symmetric.

Transitivity: for all $x, y, z \in X$, if $(x, y) \in R$ and $(y, z) \in R$, then $(x, z) \in R$.

EXAMPLE 8.19. *Let X be the set of all human beings, alive and dead.*

Let P the parenthood relation on X: $(x, y) \in P$ if and only if x is a parent of y. The parenthood relation is anti-reflexive and strongly anti-symmetric. It is neither, reflexive, symmetric nor transitive.

Let A be the ancestry relation on X: $(x, y) \in A$ if and only if x is an ancestor of y. The ancestry relation is anti-reflexive, strongly anti-symmetric, and transitive. It is neither reflexive nor symmetric.

Now we make two further defintions of relations that possess certain combinations of these basic properties. The first is the most important definition in this section.

An **equivalence relation** on a set X is a relation on X which is reflexive, symmetric and transitive.

A **partial ordering** on a set X is a relation on X which is reflexive, anti-symmetric and transitive.

We often denote equivalence relations by a tilde – $x \sim y$ – and read $x \sim y$ as "x is equivalent to y". For instance, the relation "having the same parity" on $\mathbb{Z}$ is an equivalence relation, and $x \sim y$ means that x and y are both even or both odd. Thus it serves to group the elements of $\mathbb{Z}$ into subsets which share some common property. In this case, all the even numbers are being grouped together and all the odd numbers are being grouped together. We will see shortly that this is a characteristic property of equivalence relations: every equivalence relation on a set X determines a **partition** on X and conversely, given any partition on X we can define an equivalence relation.

The concept of a partial ordering should be regarded as a "generalized less than or equal to" relation. Perhaps the best example is the containment relation $\subseteq$ on the power set $\mathcal{P}(S)$ of a set S. This is a very natural way of regarding one set as "bigger" or "smaller" than another set. Thus the insight here is that containment

satisfies many of the formal properties of the more familiar $\leq$ on numbers. However there is one property of $\leq$ on numbers that does not generalize to $\subseteq$ (and hence not to an arbitrary partial ordering): namely, given any two real numbers x, y we must have either $x \leq y$ or $y \leq x$. However for sets this does not need to be the case (unless S has at most one element). For instance, in the power set of the positive integers, we have $A = \{1\}$ and $B = \{2\}$, so neither is it true that $A \subseteq B$ or that $B \subseteq A$. This is a much stronger property of a relation:

Totality: For all $x, y \in X$, either $(x, y) \in R$ or $(y, x) \in R$.

A **total ordering** (or **linear ordering**) on a set X is a partial ordering that also has the property of totality.

EXAMPLE 8.20. *The relation $\leq$ on $\mathbb{R}$ is a total ordering.*

There is an entire branch of mathematics – **order theory** – devoted to the study of partial orderings.[4] In my opinion order theory gets short shrift in the standard mathematics curriculum (especially at the advanced undergraduate and graduate levels): most students learn only a few isolated results which they apply frequently but with little context or insight. In this text we also do not do fulljustice to order theory, although in §9.4 we discuss three classic results in this area and in §12.4 we discuss Zorn's Lemma.

2. Equivalence Relations

Let X be a set, and let $\sim$ be an equivalence relation on X.

For $x \in X$, we define the **equivalence class of x** as

$$\mathfrak{c}(x) = \{y \in X \mid y \sim x\}.$$

For example, if $\sim$ is the relation "having the same parity" on $\mathbb{Z}$, then

$$\mathfrak{c}(2) = \{\ldots, -4, -2, 0, 2, 4, \ldots\},$$

i.e., the set of all even integers. Similarly

$$\mathfrak{c}(1) = \{\ldots - 3, -1, 1, 3, \ldots\}$$

is the set of all odd integers. But an equivalence class in general has many "representatives". For instance, sticking with the equivalence relation of having the same parity, the equivalence class $\mathfrak{c}(4)$ is the set of all integers having the same parity as 4, so is again the set of all even integers:

$$\mathfrak{c}(4) = \mathfrak{c}(2).$$

More generally, we have

$$\forall \text{ even integers } n,\ \mathfrak{c}(n) = \mathfrak{c}(0)$$

and

$$\forall \text{ odd integers } n,\ \mathfrak{c}(n) = \mathfrak{c}(1).$$

Thus we we have partitioned the integers into two subsets: the even integers and the odd integers.

[4]There is even a journal called **Order**, which published in particular the following: [**Cl15**].

We claim that given any equivalence relation $\sim$ on a set X, the set $\{\mathfrak{c}(x) \mid x \in X\}$ forms a partition of X. Before we proceed to demonstrate this, observe that we are now strongly using our convention that there is no "multiplicity" associated to membership in a set: e.g. the sets $\{4, 2+2, 1^1+3^0+2^1\}$ and $\{4\}$ are equal. The above representation $\{\mathfrak{c}(x) \mid x \in X\}$ is highly redundant: for instance in the above example we are writing down the set of even integers and the set of odd integers infinitely many times, but it only "counts once" in order to build the set of subsets which gives the partition.

With this disposed of, the verification that $\mathcal{P} = \{\mathfrak{c}(x) \mid x \in X\}$ gives a partition of X comes down to recalling the definition of a partition and then following our noses. There are three properties to verify:

(i) That every element of $\mathcal{P}$ is nonempty. Indeed, the element $\mathfrak{c}(x)$ is nonempty because it contains x! This is by reflexivity: $x \sim x$, so $x \in \{y \in X \mid y \sim x\}$.
(ii) That the union of all the elements of $\mathcal{P}$ is all of X. But again, the union is indexed by the elements x of X, and we just saw that $x \in \mathfrak{c}(x)$, so every x in X is indeed in at least one element of $\mathcal{P}$.
(iii) Finally, we must show that if $\mathfrak{c}(x) \cap \mathfrak{c}(y) \neq \varnothing$, then $\mathfrak{c}(x) = \mathfrak{c}(y)$: i.e., any two elements of $\mathcal{P}$ which have a common element must be the same element. So suppose that there exists $z \in \mathfrak{c}(x) \cap \mathfrak{c}(y)$. Writing this out, we have $z \sim x$ and $z \sim y$. By symmetry, we have $y \sim z$; from this and $z \sim x$, we deduce by transitivity that $y \sim x$, i.e., $y \in \mathfrak{c}(x)$. We claim that it follows from this that $\mathfrak{c}(y) \subseteq \mathfrak{c}(y)$. To see this, take any $w \in \mathfrak{c}(y)$, so that $w \sim y$. Since $w \sim x$, we conclude $w \sim x$, so $w \in \mathfrak{c}(x)$. Rerunning the above argument with the roles of x and y interchanged we get also that $\mathfrak{c}(y) \subseteq \mathfrak{c}(x)$, so $\mathfrak{c}(x) = \mathfrak{c}(y)$. This completes the verification.

Note that the key fact underlying the proof was that any two equivalence classes $[x]$ and $[y]$ are either disjoint or coincident. Note also that we did indeed use all three properties of an equivalence relation.

Now we wish to go in the other direction. Suppose X is a set and $\mathcal{P} = \{U_i\}_{i \in I}$ is a partition of X(here I is just an index set). We can define an equivalence relation $\sim$ on X as follows: we say that $x \sim y$ if there exists $i \in I$ such that $x, y \in U_i$. In other words, we are decreeing x and y to be equivalent exactly when they lie in the same "piece" of the partition. Let us verify that this is an equivalence relation. First, let $x \in X$. Then, since $\mathcal{P}$ is a partition, there exists some $i \in I$ such that $x \in U_i$, and then x and x are both in U_i, so $x \sim x$. Next, suppose that $x \sim y$: this means that there exists $i \in I$ such that x and y are both in U_i; but then sure enough y and x are both in U_i ("and" is commutative!), so $y \sim x$. Similarly, if we have x, y, z such that $x \sim y$ and $y \sim z$, then there exists i such that x and y are both in U_i and a possibly different index j such that y and z are both in U_j. Since $y \in U_i \cap U_j$, we must have $U_i = U_j$ so that x and z are both in $U_i = U_j$ and $x \sim z$.

Moreover, the processes of passing from an equivalence relation to a partition and from a partition to an equivalence relation are mutually inverse: if we start with an equivalence relation R, form the associated partition $\mathcal{P}(R)$, and then form the associated equivalence relation $\sim (\mathcal{P}(R))$, then we get the equivalence relation R

that we started with, and similarly in the other direction.

When we study functions in detail, we will give a third take on equivalence relations involving fibers of surjective functions.

2.1. Congruence modulo N. Let $N \in \mathbb{Z}^+$. We will define a relation on the integers called **congruence modulo** N: for integers x and y, we have that x is equivalent to y if $N \mid x - y$. We denote the equivalence by

$$x \equiv y \pmod{N}.$$

We will first verify that congruence modulo N is an equivalence relation and then give another way to think about it in terms of division with remainder. There is *much* more to say about this relation; some of the many things that one could and/or should say will be treated later on, in §9.1.

That congruence modulo N is an equivalence relation is quite straightforward:
• Reflexivity: For $x \in \mathbb{Z}$, we have $N \mid 0$, so $N \mid (x - x)$, so $x \equiv x \pmod{N}$.
• Symmetry: For $x, y \in \mathbb{Z}$, if $x \equiv y \pmod{N}$, then $N \mid x - y$, so by Proposition 5.3 we have $N \mid -(x - y)$. Thus $N \mid y - x$, so $y \equiv x \pmod{N}$.
• Transitivity: For $x, y, z \in \mathbb{Z}$, if $x \equiv y \pmod{N}$ and $y \equiv z \pmod{N}$, then $N \mid (x-y)$ and $N \mid (y-z)$, so by Proposition 5.4 we have $N \mid ((x - y) + (y - z)) = x-z$, and thus $x \equiv z \pmod{N}$.

Suppose $N = 1$. Since every integer is divisible by 1, we have $x \equiv y \pmod{1}$ for all $x, y \in \mathbb{Z}$. This the relation of congruence modulo 1 is the *trivial relation* on $\mathbb{Z}$ in which everything relates to everything else.

Suppose $N = 2$. The relation of congruence modulo 2 is more interesting but still familiar: indeed we claim that integers x and y are congruent modulo 2 if and only if they have the same parity. Indeed, suppose first that $x \equiv y \pmod{2}$. If x is even, then $2 \mid (x - y)$ and $2 \mid x$, so $2 \mid x - (x - y) = y$, so y is also even. If y is even, then $2 \mid (x - y)$ and $2 \mid y$, so $2 \mid (x - y) + y = x$. Thus x is even if and only if y is even, so x and y have the same parity. Conversely, suppose that x and y have the same parity. If x and y are both even, then for some $k, l \in \mathbb{Z}$ we have $x = 2k$, $y = 2l$, so

$$x - y = 2k - 2l = 2(k - l)$$

is even. On the other hand, if x and y are both odd, then for some $k, l \in \mathbb{Z}$ we have $x = 2k + 1$, $y = 2l + 1$, so

$$x - y = (2k + 1) - (2l + 1) = 2(k - 1)$$

is even.

Another way to express that integers x and y have the same parity is that upon division by 2 they leave the same remainder. Thus integers x and y are congruent modulo 2 if and only if they leave the same remainder upon division by 2.

This observation generalizes to congruence modulo N for all $N \in \mathbb{Z}^+$:

PROPOSITION 8.21. *Let $N \in \mathbb{Z}^+$. For integers $x, y \in \mathbb{Z}$, the following are equivalent:*

(i) *We have $x \equiv y \pmod{N}$.*

(ii) *The integers x and y leave the same remainder upon division by N. More precisely: there are $q_x, q_y \in \mathbb{Z}$ and $0 \leq r < N$ such that*

$$x = q_x N + r \text{ and } y = q_y N + r.$$

PROOF. (i) $\implies$ (ii): Suppose that $x \equiv y \pmod{N}$, and write $x = q_x N + r_x$ and $y = q_y N + r_y$ with $0 \leq r_x, r_y < N$. Then

$$N \mid x - y = (q_x N + r_x) - (q_y N + r_y) = (q_x - q_y)N + (r_x - r_y).$$

Since $N \mid (q_x - q_y)N$, it follows that $N \mid (r_x - r_y)$. Since $0 \leq r_x, r_y < N$, we have $|r_x - r_y| < N$, so being a multiple of N, we must have $r_x - r_y = 0$, i.e., $r_x = r_y$.
(ii) $\implies$ (i): If there is $0 \leq r < N$ such that $x = q_x N + r$ and $y = q_y N + r$, then

$$x - y = (q_x N + r) - (q_y N + r) = (q_x - q_y)N,$$

so $N \mid x - y$. □

Thus two integers are congruent modulo N precisely when they leave the same remainder upon division by N. Thus the congruence classes are precisely

$$\mathfrak{c}(0) = \{\ldots, -3N, -2N, -N, 0, N, 2N, 3N, \ldots\},$$
$$\mathfrak{c}(1) = \{\ldots, -3N+1, -2N+1, -N+1, 1, N+1, 2N+1, 3N+1, \ldots\},$$
$$\mathfrak{c}(2) = \{\ldots, -3N+2, -2N+2, -N+2, 2, N+2, 2N+2, 3N+2, \ldots\},$$
$$\vdots$$
$$\mathfrak{c}(N-1) = -\{\ldots, -3N+(N-1), -2N+(N-1), -N+(N-1), N-1, N+(N-1), 2N+(N-1), \ldots\}.$$

Since $-N + (N-1) = -1$, the last class is equal to

$$\mathfrak{c}(-1) = \{\ldots, -2N-1, -N-1, -1, N-1, 2N-1, 3N-1, \ldots\}.$$

3. Composition of Relations

Suppose we have a relation $R \subseteq X \times Y$ and a relation $S \subseteq Y \times Z$. We can define a **composite relation** $S \circ R \subseteq X \times Z$ in a way which will generalize compositions of functions. Compared to composition of functions, composition of relations is much less well-known, although as with many abstract concepts, once it is pointed out to you, you begin to see it "in nature'. This section is certainly optional reading.

The definition is simply this:

$$S \circ R = \{(x, z) \in X \times Z \mid \exists y \in Y \text{such that } (x, y) \in R \text{ and } (y, z) \in S\}.$$

In other words, we say that x in the first set X relates to z in the third set Z if there exists at least one intermediate element y in the second set such that x relates to y and y relates to z.

Exercise 8.12 asks you to show that composition of relations is associative and has identity elements.

We can always compose relations on a single set X. As a special case, given a relation R, we can compose it with itself: say

$$R^{(2)} = R \circ R = \{(x, z) \in X \times X \mid \exists y \in X \text{ such that } xRy \text{ and } yRz\}.$$

More generally, for any $n \geq 2$, we put

$$R^{(n)} \coloneqq R \circ R \circ \cdots \circ R,$$

$(n-1)$ compositions in all, which is equal to

$$\{(x_1, x_{n+1}) \in X \times X \mid \exists x_2, \ldots, x_n \in X \text{ with } (x_1, x_2), (x_2, x_3), \ldots, (x_n, x_{n+1}) \in R\}.$$

We also put $R^{(0)} \coloneqq \Delta_X$, the diagonal (or identity) relation.

PROPOSITION 8.22.

a) *For a relation R on X, the following are equivalent:*
   (i) *The relation R is transitive.*
   (ii) *For all $n \in \mathbb{Z}^+$, we have $R^{(n)} \subseteq R$.*
   (iii) *We have $R^{(2)} \subseteq R$.*

b) *If R is reflexive, then for all integers $0 \leq m \leq n$ we have $R^{(m)} \subseteq R^{(n)}$.*

PROOF. a) (i) $\implies$ (ii): Suppose R is transitive, let $n \in \mathbb{Z}^+$ and let $(x, y) \in R^{(n)}$. Then there are $a_1, \ldots, a_{n-1} \in X$ such that $(x, a_1), (a_1, a_2), \ldots, (a_{n-1}, y) \in R$. Since R is transitive, we get (by induction, strictly speaking) that $(x, y) \in R$.
(ii) $\implies$ (iii): If $R^{(n)} \subseteq R$ for all $n \geq 2$, then taking $n = 2$ we get $R^{(2)} \subseteq R$.
(iii) $\iff$ (i): Since $R^{(2)}$ is the set of $(x, z) \in X \times X$ for which there is $y \in X$ with $(x, y), (y, z) \in R$, we have $R^{(2)} \subseteq R$ if and only if for all $x, y, z \in X$, we have $(x, y), (y, z) \in R$ implies $(x, z) \in R$. The latter is precisely the transitive property.
b) Suppose R is reflexive. If $(x, y) \in R^{(m)}$, there are $a_1, \ldots, a_{m-1} \in X$ such that

$$(x, a_1), (a_1, a_2), \ldots, (a_{m-1}, y) \in R,$$

and since R is reflexive we have

$$(x, a_1), (a_1, a_2), \ldots, (a_{m-1}, y), (y, y), \ldots (y, y) \in R,$$

where we include $n - m$ instances of (y, y). Thus $(x, y) \in R^{(n)}$. □

Exercise 8.14 explores the failure of the converse of Proposition 8.22b).

4. Some Relational Closures

When a relation R on a set X lacks a desired certain property – like being an equivalence relation – do we just give up? We don't have to: often we can define a new relation in terms of R that does possess that property. In general this could be done in many ways, but let us suppose that we are looking for a relation $\tilde{R}$ *containing* R that has that property: that is, the new relation $\tilde{R}$ is obtained from R by adding further ordered pairs $(x, y) \in X \times X$.

• We begin with the property of reflexivity. A relation on X is reflexive if and only if it contains the diagonal $\Delta = \{(x, x) \mid x \in X\}$, so if R is any relation on X then the relation

$$R_r \coloneqq R \cup \Delta$$

obtained from R by adjoining the diagonal is reflexive, and moreover any reflexive relation $S \supseteq R$ must also contain R_r. This gives a sense in which R_r is the *minimal* reflexive relation containing R. Let us call R_r the **reflexive closure** of R.

• Consider now the property of symmetry. A relation R on X is symmetric if and only if it contains its inverse R^{-1}. If R is any relation on X then the relation

$$R_s \coloneqq R \cup R^{-1}$$

is symmetric, but this time there is something to check: if $(x, y) \in R_s$ then either $(x, y) \in R$ or $(y, x) \in R$, so (y, x) lies in either R^{-1} or in R, hence $(y, x) \in R_s$. Moreover any symmetric relation $S \supseteq R$ must also contain R_s. This gives a sense in which R_s is the *minimal* symmetric relation containing R. Let us call R_s the **symmetric closure** of R.

• And now consider the property of transitivity, which is more interesting. By Proposition 8.22, a relation R on a set X is transitive if and only if $R \supseteq R^{\circ 2}$, so if $S \supseteq R$ is any transitive relation then $S \supseteq R \cup R^{(2)}$. By comparison with the last two cases one might guess that $R \cup R^{(2)}$ is transitive. But this need not be the case:

EXAMPLE 8.23. *Let R be the relation on $\mathbb{Z}$ given by $(x, y) \in R$ if and only if $|x - y| \leq 1$. In other words, for all $n \in \mathbb{Z}$, n relates to $n - 1$, to n, to $n + 1$ and to no other integers. This relation is reflexive and symmetric but not transitive, since e.g. $(0, 1) \in R$ and $(1, 2) \in R$ but $(0, 2) \notin R$. By Proposition 8.22b) we have $R \subseteq R^{(2)}$, so*

$$R \cup R^{(2)} = R^{(2)} = \{(x, y) \in \mathbb{Z} \times \mathbb{Z} \mid |x - y| \leq 2\}.$$

The relation $R^{(2)}$ is still not transitive: it contains $(0, 2)$ and $(2, 3)$ but not $(0, 3)$.

We can continue on in this manner: for all $n \in \mathbb{N}$ we have

$$R^{(n)} \coloneqq \{(x, y) \in \mathbb{Z} \mid |x - y| \leq n\},$$

so we have

$$\Delta = R^{(0)} \subsetneq R \subsetneq R^{(2)} \subsetneq R^{(3)} \subsetneq \ldots \subsetneq R^{(n)} \subsetneq \ldots.$$

By Proposition 8.22a), any transitive relation S that contains R must also contain $R^{(n)}$ for all n and thus

$$S \supseteq \bigcup_{n \geq 0} R^{(n)} = \bigcup_{n \geq 0} \{(x, y) \in \mathbb{Z} \times Z \mid |x - y| \leq n\} = \mathbb{Z} \times \mathbb{Z}.$$

Thus the only transitive relation containing R is the trivial relation $\mathbb{Z} \times \mathbb{Z}$.

PROPOSITION 8.24. *Let R be a relation on a set X.*

a) *The relation $R_t \coloneqq \bigcup_{n \geq 1} R^{(n)}$ is a transitive relation containing R.*
b) *If $S \supseteq R$ is any transitive relation, then $S \supseteq R_t$.*

*The relation R_t is called the **transitive closure** of R.*

PROOF. a) It is clear that $R_t \supseteq R^{(1)} = R$, so it suffices to show that R_t is transitive. Suppose $x, y, z \in X$ are such that $(x, y), (y, z) \in R_t$. Then there are $m, n \in \mathbb{Z}^+$ such that $(x, y) \in R^{(m)}$ and $(y, z) \in R^{(n)}$, so

$$(x, z) \in R^{(m)} \circ R^{(n)} = R^{(m+n)} \subseteq R_t.$$

b) By Proposition 8.22a), if S is transitive, then for all $n \in \mathbb{Z}^+$ we have $S \supseteq S^{(n)} \supseteq R^{(n)}$, so $S \supseteq \bigcup_{n=1}^{\infty} R^{(n)} = R_t$. □

EXAMPLE 8.25. *Let X be the set of all human beings, alive and dead. Let R be the relation of parenthood on X: that is $(x, y) \in R$ if and only if x is the parent of y. Then the transitive closure R_t is the ancestry relation: $(x, y) \in R_t$ if and only if x is an ancestor of y.*

LEMMA 8.26. *Let R and S be relations on a set X.*

a) *We have:*
   (i) *$R \subseteq R_r$, with equality if and only if R is reflexive.*
   (ii) *$R \subseteq R_s$, with equality if and only if R is symmetric.*
   (ii) *$R \subseteq R_t$, with equality if and only if R is transitive.*

b) *We have $(R_r)_r = R_r$, $(R_s)_s = R_s$ and $(R_t)_t = R_t$.*

c) *Show: if $R \subseteq S$, then $R_r \subseteq S_r$, $R_s \subseteq S_s$ and $R_t \subseteq T_t$.*

PROOF. a) This essentially summarizes what we have already shown. We know that R_r contains R, is reflexive and that if $S \supseteq R$ is a reflexive relation, then $S \supseteq R_r$, so if R is reflexive we may take $S = R$ to get $R \supseteq R_r$ and thus $R_r = R$, and conversely, certainly R is reflexive if R is equal to the reflexive relation R_f. This shows part (i), and exactly the same arguments with R_s and R_t in place of R_r show parts (ii) and (iii).
b) By part a) applied to the relation R_r, we know that $(R_r)_r$ contains R_r, with equality if and only if R_r is reflexive. But R_r *is* reflexive, so $(R_r)_r = R_r$. Exactly the same argument shows that $(R_s)_s = R_s$ and that $(R_t)_t = R_t$.
c) Suppose $R \subseteq S$. Then

$$R_r = R \cup \Delta \subseteq S \cup \Delta = R_s.$$

Moreover, since $R \subseteq S$ we have $R^{-1} \subseteq S^{-1}$: indeed if $(x, y) \in R^{-1}$ then $(y, x) \in R \subseteq S$, so $(x, y) \in S^{-1}$. So

$$R_s = R \cup R^{-1} \subseteq S \cup S^{-1} = S_s.$$

Since $R \subseteq S$, for all $n \geq 1$ we have $R^{(n)} \subseteq S^{(n)}$: if $(x, y) \in R^{(n)}$, there are $a_1, \ldots, a_{n-1} \in X$ such that $(x, a_1), (a_1, a_2), \ldots, (a_{n-1}, y) \in R \subseteq S$, and thus $(x, y) \in S^{(n)}$. So

$$R_t = \bigcup_{n \geq 1} R^{(n)} \subseteq \bigcup_{n \geq 1} S^{(n)} = S_t. \qquad \square$$

For any relation R, we write R_{rs} for $(R_r)_s$. Thus R_{rs} is obtained from R by adjoining the diagonal and then adjoining the reflections across the diagonal of all the elements of R. Since $R_{\mathrm{rs}} \supseteq R_r \supseteq \Delta$, it is reflexive. It is certainly also symmetric. If $S \supseteq R$ is any relation that is both reflexive and symmetric, then by Lemma 8.26 we have

$$S = S_r \supseteq R_r,$$

so

$$S = S_{\mathrm{rs}} \supset (R_r)_s = R_{\mathrm{rs}}.$$

Thus R_{rs} is the minimal relation containing R that is reflexive and symmetric. We call it the **reflexive-symmetric closure** of X.

For a relation R, we now put

$$R_\sim \coloneqq (R_{\mathrm{rs}})_t = ((R_r)_s)_t.$$

THEOREM 8.27. *Let R, S be relations on a set X. Then:*

a) *The relation $R_\sim$ is an equivalence relation containing R.*

b) *For any equivalence relation S containing R, we have $S \supseteq R_\sim$.*
c) *We have $R = R_\sim$ if and only if R is an equivalence relation.*
d) *We have $(R_\sim)_\sim = R_\sim$.*
e) *If $R \subseteq S$, then $R_\sim \subseteq S_\sim$.*

PROOF. a) Since $R \subseteq R_r \subseteq (R_r)_s \subseteq ((R_r)_s)_t$, the relation $R^\sim$ contains R.

Since $\Delta \subseteq R_{\text{rs}} \subseteq (R_{\text{rs}})_t = R_\sim$, so $R_\sim$ is reflexive. Let $(x, y) \in R_\sim$. Then there is $n \in \mathbb{Z}^+$ and $a_1, \ldots, a_{n-1} \in X$ such that $(x, a_1), (a_1, a_2), \ldots, (a_n, y) \in R_{\text{rs}}$. Since R_{rs} is symmetric, also $(y, a_n), (a_n, a_{n-1}), \ldots, (a_2, a_1), (a_1, x) \in R_{\text{rs}}$, so $(y, x) \in (R_{\text{rs}})_t = R_\sim$. Thus $R_\sim$ is an equivalence relation.
b) If $S \supseteq R$ is any equivalence relation, then S is reflexive, so

$$S = S_r \supseteq R_r$$

and thus, since S is symmetric,

$$S = S_s \supseteq (R_r)_s = R_{\text{rs}}$$

and then finally, since S is transitive

$$S = S_t \supseteq (R_{\text{rs}})_t = R_\sim.$$

c) If R is an equivalence relation, then taking $S = R$ in part b) gives $R \supseteq R_\sim$, and since we always have $R \subseteq R_\sim$, we conclude that $R = R_\sim$. Conversely, if $R = R_\sim$ then since $R_\sim$ is an equivalence relation, so is R.
d) By parts a) and c) we know that $(R_\sim)_\sim$ contains $R_\sim$, with equality if and only if $R_\sim$ is an equivalence relation...which it is.
e) If $R \subseteq S$, then $R_r \subseteq S_r$, so $(R_r)_s \subseteq (S_r)_s$ and then finally

$$R_\sim = ((R_r)_s)_t \subseteq ((S_r)_s)_t = S_\sim. \qquad \square$$

It would be reasonable to call $R_\sim$ the **equivalence closure of** R. However it is more common to call $R_\sim$ the **equivalence relation generated by R**.

It should be clear by now that very similar arguments have been made several times over, and one must also suspect that there are more general principles at work here. I will now attempt a debriefing.[5]

First of all, here is a very common situation in mathematics: we have a certain "structure" A – A is in particular a set. We also have a notion of a subset $R \subseteq A$ being a "substructure" of A: depending upon R, this may or may not be the case.

In our example, A is the trivial relation $X \times X$ on X, so a subset $R \subseteq A$ is precisely a relation on X. What do we mean by a substructure? Well, each property of relation that we have considered gives a (different) such notion: the trivial relation $X \times X$ is reflexive, so we can say that a relation R is a substructure if it is reflexive. Similarly for symmetric, transitive, or equivalence relation.

Now if a subset $R \subseteq A$ is *not* a substructure, then we may ask for the **substructure generated by S**: this should be a substructure $\tilde{R}$ of A such that $\tilde{R}$ contains R and for any substructure T of A that contains X, also $T \supseteq \tilde{R}$.

There is a very general sufficient condition for the "substructure generated by S" to exist: namely, if for any family $\{R_i\}_{i \in I}$ of substructures of A, the intersection

[5]The debriefing is itself somewhat lengthy and complicated. But I think it will be of significant interest to some readers, at least.

$\bigcap_{i \in I} R_i$ is again a substructure of A, then the substructure of A generated by R is $\tilde{R} := \bigcap T$, where T ranges over all substructures of A that contain R. Indeed, as the intersection of a family of sets each containing R, certainly $\tilde{R}$ contains R, and it is a substructure because of our assumption that the intersection over any family of substructures is again a substructure. Finally, if T is any substructure of A that contains R, then T is one of the things we intersected over to get $\tilde{R}$, so $\tilde{R} \subseteq T$.

It it is straightforward to show that any intersection of reflexive relations is a reflexive relation, and similarly with "reflexive" replaced by any of: symmetric, reflexive-symmetric, transitive, equivalence relation. Thus in a much quicker way we can show that for any relation R on a set X, there is an equivalence relation $\tilde{R}$ containing R and with the property that for any equivalence relation $S \supseteq R$ we have $S \supseteq \tilde{R}$. By Theorem 8.27b) the relation $R_\sim$ has these properties, and since these properties of $\tilde{R}$ and $R_\sim$ force

$$\tilde{R} \supseteq R_\sim \text{ and } R_\sim \supseteq \tilde{R},$$

we must have $\tilde{R} = R_\sim$. In other words, the equivalence closure $R_\sim$ that we built up over several pages is just the intersection of all equivalence relations on X containing R. So what was the point of all the work we did in constructing $R_\sim$?

Here is the point: if we only want to ensure that the substructure $\tilde{R}$ generated by a subset R of the structure A exists, the "intersection method" is the ideal solution. It works essentially verbatim in the following contexts, among many others:

- To show that the subgroup generated by a subset S of a group G exists.
- To show that the subspace generated by a subset S of a vector space V exists.
- To show that the ideal generated by a subset S of a commutative ring R exists.

In each case, just as soon as your learn the definition of subgroup / subspace / ideal, you can show that the intersection over any family of subgroups / subspaces / ideals is another subgroup / subspace / ideal.

The catch however is if we want to know something about $\tilde{R}$ other than that it is the unique minimal substructure of A that contains R...which, of course, we usually do. Intersecting over an enormous family of sets usually tells us nothing concrete. So it is extremely useful to supplement this **top-down** description of $\tilde{R}$ with a **bottom-up** description of $\tilde{R}$, namely with some sort of *procedure* that tells us which elements of A to adjoint to R to get $\tilde{R}$.

- In the case of a subset S of a group G, it is much more useful to know that the subgroup of G generated by S consists of all finite products $g_1 \cdots g_n$ where each g_i is either an element $s \in S$ or the inverse s^{-1} of an element of S.
- In the case of a subset S of a vector space V over a field F, it is much more useful to know that the subspace of V generated by S consists of all finite F-linear combinations $a_1 s_1 + \ldots + a_n s_n$ for $a_i \in F$ and $s_i \in S$.
- In the case of a subset S of a commutative ring R, it is much more useful to know that the ideal of R generated by S consists of all finite R-linear combinations $a_1 r_1 + \ldots + a_n r_n$ for $a_i \in R$ and $s_i \in S$.

Which brings us back to our description of the equivalence closure of a relation R as $R_\sim$ rather than the (equal, but differently described) relation $\tilde{R}$. To form $R_\sim$ from R we first adjoin the diagonal Δ, then we adjoin the reflections through the diagonal (y, x) of the elements (x, y), then finally we form all finite lists $x_1, \ldots, x_n$ of elements of X such that eac h x_i relates to the next element x_{i+1} (using the relation we have at this point, which is R_{rs}) and adjoin the pairs (x_1, x_n). This third step is certainly more complicated than the first two, but it is a lot better than "intersect over some crazy family of sets" and in fact the purpose of Example 8.23 is to convince you that our description of the equivalence closure of a general relation cannot be any simpler than this.

We also took the time to develop results like Lemma 8.26 that were not necessary for our concrete description of $R_\sim$. There was a different purpose for this. Go back to the idea of starting with all subsets R of a structure A and trying to understand $\tilde{R}$, the substructure generated by R. This defines a map (function: coming up soon!)

$$T : 2^A \to 2^A, \ R \mapsto \tilde{R}$$

from the family of subsets of A to itself. In all five of the examples considered above – namely reflexive, symmetric, transitive, reflexive-symmetric and equivalence closures – in which we had $A = X \times X$, this map T satisfied the following properties:

(MC1) For all $R \subseteq A$, we have $R \subseteq \tilde{R}$.
(MC2) For all $R_1 \subseteq R_2 \subseteq A$, we have $\widetilde{R_1} \subseteq \widetilde{R_2}$.
(MC3) For all $R \subseteq A$, we have $\tilde{\tilde{R}} = \tilde{R}$.

The property (MC1) is simple to understand: it just says that $\tilde{R}$ should be obtained from R by adjoining further elements.[6] Property (MC2) says that passage from R to $\tilde{R}$ perserves the property of one subset being contained in another. This is a nice propety: e.g. thatit holds for the subspace generated by a subset of $\mathbb{R}^N$ means that if $S_1 \subseteq S_2$ are subsets of $\mathbb{R}^N$, then the span of S_1 is contained in the span of S_2. Property (MC3) says our procedure of forming $\tilde{R}$ from R stabilizes after the first step. It can be thought of as saying that if applying T gives you a property that you want, then if you already have that property then nothing further is done.

An operator $T : 2^A \to 2^A$ satisfying (MC1), (MC2) and (MC3) is called a **Moore closure operator**, and given such an operator, for any $R \subseteq A$ we can think of $T(R)$ as "the closure of R with resepct to a certain property."

Once you know to look for them, Moore closure operators are everywhere. See for yourself!

[6]We didn't justify why we wanted this, and we won't now except to point out that things worked out rather nicely. Lots of other things are possible, and one cannot explore everything at once!

5. Functions

Let X and Y be sets. A **function** $f : X \to Y$ is a special kind of relation between X and Y. Namely, it is a relation $R \subseteq X \times Y$ satisfying the following condition: for all $x \in X$ there exists exactly one $y \in Y$ such that $(x, y) \in R$. Because element of y attached to a given element x of X is unique, we may denote it by $f(x)$.

Geometrically, a function is a relation which passes the **vertical line test**: every vertical line $x = c$ intersects the graph of the function in exactly one point. In particular, the domain of any function is all of X.

EXAMPLE 8.28. *The equality relation $\{(x, x) \mid x \in X\}$ on X is a function: $f(x) = x$ for all x. We call this the **identity function** and denote it by 1_X.*

EXAMPLE 8.29. *a) Let Y be a set. Then $\varnothing \times Y = \varnothing$, so there is a unique relation on $\varnothing \times Y$. This relation is – vacuously – a function.*
b) Let X be a set. Then $X \times \varnothing = \varnothing$, so there is a unique relation on $X \times \varnothing$, with domain $\varnothing$. If $X = \varnothing$, then we get the empty function $f : \varnothing \to \varnothing$. If $X \neq \varnothing$ then the domain is not all of X so we do not get a function.

If $f : X \to Y$ is a function, the second set Y is called the **codomain** of f. Note the asymmetry in the definition of a function: although every element x of the domain X is required to be associated to a unique element y of Y, the same is not required of elements y of the codomain: there may be multiple elements x in X such that $f(x) = y$, or there may be none at all.

The **image** of $f : X \to Y$ is $\{y \in Y$ such that $y = f(x)$ for some $x \in X.\}$[7]

In calculus one discusses functions with domain some subset of $\mathbb{R}$ and codomain $\mathbb{R}$. Moreover in calculus a function is usually (but not always...) given by some relatively simple algebraic/analytic expression, and the convention is that the domain is the largest subset of $\mathbb{R}$ on which the given expression makes sense.

EXAMPLE 8.30.

a) *The function $y = 3x$ is a function from $\mathbb{R}$ to $\mathbb{R}$. Its range is all of $\mathbb{R}$.*
b) *The function $y = x^2$ is a function from $\mathbb{R}$ to $\mathbb{R}$. Its range is $[0, \infty)$.*
c) *The function $y = x^3$ is a function from $\mathbb{R}$ to $\mathbb{R}$. Its range is all of $\mathbb{R}$.*
d) *The function $y = \sqrt{x}$ is a function from $[0, \infty)$ to $\mathbb{R}$. Its range is $[0, \infty)$.*
e) *The arctangent $y = \arctan x$ is a function from $\mathbb{R}$ to $\mathbb{R}$. Its range is $(\frac{-\pi}{2}, \frac{\pi}{2})$.*

Let X and Y be sets. We denote the set of all functions $f : X \to Y$ by Y^X. This notation gets some justification from Proposition 8.60 later on.

5.1. Injective functions.

From the perspective of our course, the most important material on functions are the concepts *injectivity, surjectivity and bijectivity* and the relation of these properties with the existence of inverse functions.

[7]Some people call this the **range**, but also some people call the set Y (what we called the codomain) the range, so the term is ambiguous and perhaps best avoided.

A function $f : X \to Y$ is **injective** if every element y of the codomain is associated to at most one element $x \in X$. That is, f is injective if for all $x_1, x_2 \in X$, $f(x_1) = f(x_2)$ implies $x_1 = x_2$.

Let us reflect a bit on the property of injectivity. One way to think about it is via a horizontal line test: a function is injective if and only if each horizontal line $y = c$ intersects the graph of f in **at most** one point. Another way to think about an injective function is as a function which entails no loss of information. That is, for an injective function, if your friend tells you $x \in X$ and you tell me $f(x) \in Y$, then I can, in principle, figure out what x is because it is uniquely determined.

Consider for instance the two functions $f(x) = x^2$ and $f(x) = x^3$. The first function $f(x) = x^2$ is not injective: if y is any positive real number then there are two x-values such that $f(x) = y$, $x = \sqrt{y}$ and $x = -\sqrt{y}$. Or, in other words, if $f(x) = x^2$ and I tell you that $f(x) = 1$, then you are in doubt as to what x is: it could be either $+1$ or -1. On the other hand, $f(x) = x^3$ is injective, so if I tell you that $f(x) = x^3 = 1$, then we can conclude that $x = 1$.

5.2. Surjective functions. A function $f : X \to Y$ if its image $f(X)$ is equal to the codomain Y. More plainly, for all $y \in Y$, there is $x \in X$ such that $f(x) = y$.

In many ways surjectivity is the "dual property" to injectivity. For instance, it can also be verified by a horizontal line test: a function f is surjective if and only if each horizontal line $y = c$ intersects the graph of f in **at least one point**.

EXAMPLE 8.31. *Let m and b be real numbers. Is $f(x) = mx + b$ surjective? indent Answer: It is surjective if and only if $m \neq 0$. First, if $m = 0$, then $f(x) = b$ is a constant function: it maps all of $\mathbb{R}$ to the single point b and therefore is at the opposite extreme from being surjective. Conversely, if $m \neq 0$, write $y = mx + b$ and solve for x: $x = \frac{y-b}{m}$. Note that this argument also shows that if $m \neq 0$, f is injective: given an arbitary y, we have solved for a unique value of x.*

5.3. Using Calculus to Study Injectivity and Surjectivity. A familiar and important class of functions are those with domain and codomain the real numbers, i.e.,

$$f : \mathbb{R} \to \mathbb{R}.$$

If f is moreover continuous and/or differentiable, then the methods of calculus may be usefully brought to bear to help study the injectivity and surjectivity of f. In this text we do not develop the theory of continuity or differentiability, but we will now make some references to it. The relevant definitions and results are drawn from Chapters 4 through 6 of [**Cl-HC**].

A function $f : \mathbb{R} \to \mathbb{R}$ is **strictly increasing** if for all $x_1, x_2 \in \mathbb{R}$, $x_1 < x_2 \implies f(x_1) < f(x_2)$. Similarly, f is **strictly decreasing** if for all $x_1, x_2 \in \mathbb{R}$, $x_1 < x_2 \implies f(x_1) > f(x_2)$.

PROPOSITION 8.32. *If $f : \mathbb{R} \to \mathbb{R}$ is either strictly increasing or strictly decreasing, then f is injective.*

PROOF. Suppose that $f : \mathbb{R} \to \mathbb{R}$ is strictly increasing, and let $x_1 \neq x_2$ be real numbers. By totality, we have either $x_1 < x_2$ or $x_2 < x_1$. If $x_1 < x_2$, then $f(x_1) < f(x_2)$, so $f(x_1) \neq f(x_2)$. If $x_2 < x_1$, then $f(x_2) < f(x_1)$, so $f(x_1) \neq f(x_2)$.

It follows that f is injective. The case where f is strictly decreasing is very similar and left to the reader as Exercise 8.21. □

The converse of Proposition 8.32 is false:

EXAMPLE 8.33. *Let $f : \mathbb{R} \to \mathbb{R}$ be given by*

$$x \in \mathbb{R} \mapsto \begin{cases} \frac{1}{x} & x \neq 0 \\ 0 & x = 0 \end{cases}.$$

This function is injective: 0 is not the reciprocal of any real number, so $f(x) = 0$ implies $x = 0$, while every nonzero real number x is the reciprocal of a unique real number $\frac{1}{x}$. (Moreover f is surjective.)

• *Since $1 < 2$ but $f(1) = 1 > f(2) = \frac{1}{2}$, the function f is not strictly increasing.*
• *Since $-2 < -1$ but $f(-2) = \frac{-1}{2} > -1 = f(-1)$, the function f is not strictly decreasing.*

Any devotee of calculus would, upon seeing Example 8.33, notice that the function f is not continuous. Indeed, it turns out that all such examples have this property.

THEOREM 8.34. *If $f : \mathbb{R} \to \mathbb{R}$ is injective and continuous, then f is either strictly increasing or strictly decreasing.*

PROOF. See [**Cl-HC**, Thm. 3.53]. (The proof uses the Intermediate Value Theorem and also a slightly tricky case analysis. Perhaps because of the latter, the result is not as standard and well-known as it perhaps should be.) □

Thus for continuous functions $f : \mathbb{R} \to \mathbb{R}$, determining whether f is injective is the same as determining whether it is either strictly increasing or strictly decreasing. As one learns in calculus, we can say a lot about the latter question as long as f is not merely continuous but also differentiable.

THEOREM 8.35. *Let $f : \mathbb{R} \to \mathbb{R}$ be a differentiable function.*

a) *If $f'(x) > 0$ for all $x \in \mathbb{R}$, then f is strictly increasing.*
b) *If $f'(x) < 0$ for all $x \in \mathbb{R}$, then f is strictly decreasing.*

PROOF. We will use the Mean Value Theorem [**Cl-HC**, Thm. 5.16].
a) By contrapositive: if f is not strictly increasing, there are $a < b$ in $\mathbb{R}$ with $f(a) \geq f(b)$. Applying the Mean Value Theorem to f on the interval $[a, b]$, there is $a < c < b$ such that

$$f'(c) = \frac{f(b) - f(a)}{b - a} \leq 0.$$

b) This is very similar to part a) and left to the reader as Exercise 8.22. □

EXAMPLE 8.36.

a) *Let $f : \mathbb{R} \to \mathbb{R}$ by $f(x) = \arctan x$. We claim f is injective. Indeed, it is differentiable and its derivative is $f'(x) = \frac{1}{1+x^2} > 0$ for all $x \in \mathbb{R}$. Therefore f is strictly increasing, hence injective.*
b) *Let $f : \mathbb{R} \to \mathbb{R}$ by $f(x) = -x^3 - x$. We claim f is injective. Indeed, it is differentiable and its derivative is $f'(x) = -3x^2 - 1 = -(3x^2 + 1) < 0$ for all $x \in \mathbb{R}$. Therefore f is strictly decreasing, hence injective.*

Let $f : \mathbb{R} \to \mathbb{R}$ be given by $f(x) = x^3$. We showed in Example 6.17 that f is injective (without using that word). Because f is, like all polynomial functions, continuous, it follows from Theorem 8.34 that it must be either strictly increasing or strictly decreasing. Because $0 < 1$ and $f(0) = 0 < 1 = f(1)$, it must be strictly increasing. But this is quite a roundabout argument: why are we not applying Theorem 8.35 directly?

The reason is because Theorem 8.35 doesn't apply: we have $f'(x) = 3x^2$, which is always non-negative but is 0 at $x = 0$. This is a bit of a downer: to prove even that $f(x) = x^5$ is injective "by hand" as we did for $x \mapsto x^3$ is not so easy. We would like to be able to use calculus to show, among other things, that for any positive integer n, the function $f(x) = x^n$ is strictly increasing, even though $f'(x) = nx^{n-1}$ is positive for all nonzero $x \in \mathbb{R}$ but 0 at $x = 0$. For this we need to sharpen Theorem 8.35. This justifies the following result (whose statement is more complicated than that of Theorem 8.35, but this now seems unavoidable).

THEOREM 8.37. *Let $f : \mathbb{R} \to \mathbb{R}$ be a differentiable function.*

a) *Suppose that $f'(x) \geq 0$ for all x and that there is no $a < b$ such that $f'(x) = 0$ for all $x \in (a, b)$. Then f is strictly increasing (hence injective).*
b) *Suppose that $f'(x) \leq 0$ for all x and that there is no $a < b$ such that $f'(x) = 0$ for all $x \in (a, b)$. Then f is strictly decreasing (hence injective).*

PROOF. We prove part a); the proof of part b) is identical. Again we go by contrapositive: suppose that f is not strictly increasing, so that there exists $a < b$ such that $f(a) \leq f(b)$. If $f(a) < f(b)$, then applying the Mean Value Theorem, we get a c in between a and b such that $f'(c) < 0$, contradiction. So we may assume that $f(a) = f(b)$. Exactly the same Mean Value Theorem argument shows that if $f'(x) \geq 0$ for all x, then $x_1 \leq x_2 \implies f(x_1) \leq f(x_2)$. But such a function that also satisfies $f(a) = f(b)$ must be constant on the entire interval $[a, b]$, hence $f'(x) = 0$ for all x in (a, b), contradicting the hypothesis. □

EXAMPLE 8.38. *We will show that for any $n \in \mathbb{Z}^+$, the function $f : \mathbb{R} \to \mathbb{R}$ given by $x \mapsto x^{2n+1}$ is injective. Indeed we have $f'(x) = (2n + 1)x^{2n}$, which is non-negative for all $x \in \mathbb{R}$ and is 0 only at $x = 0$. So Theorem 8.35a) applies to show that f is strictly increasing, hence injective.*

In Exercise 8.23, for each odd $n \geq 3$ you are asked to exhibit a degree n polynomial $P : \mathbb{R} \to \mathbb{R}$ that is *not* injective.

We can also use calculus to give a criterion for a continuous function $f : \mathbb{R} \to \mathbb{R}$ *not* to be injective. To prepare for the proof, we recall that for $f : \mathbb{R} \to \mathbb{R}$, $\lim_{x\to\infty} f(x) = \infty$ means: for all real numbers M, there is a real number X such that for all $x \geq X$ we have $f(x) \geq M$. Similarly, $\lim_{x\to\infty} f(x) = -\infty$ means: for all real numbers m, there is a real number X such that for all real numbers $x \geq X$ we have $f(x) \leq m$; moreover, $\lim_{x\to-\infty} f(x) = \infty$ means: for all real numbers M, there is a real number X such that for all real numbers $x \leq X$ we have $f(x) \geq M$; and finally $\lim_{x\to-\infty} f(x) = -\infty$ means: for all real numbers m, there is a real number X such that for all real numbers $x \geq X$ we have $f(x) \leq m$.

THEOREM 8.39. *Let $f : \mathbb{R} \to \mathbb{R}$ be a continuous function. Suppose that either*

(i) $\lim_{x\to\infty} f(x) = \lim_{x\to-\infty} = \infty$, *or*
(ii) $\lim_{x\to\infty} f(x) = \lim_{x\to-\infty} = -\infty$.

Then f is not *injective.*

PROOF. We will show that condition (i) implies that f is not injective. The argument that condition (ii) implies that f is not injective is left as Exercise 8.27.

Since $\lim_{x\to\infty} f(x) = \infty$, for every real number $M > f(1)$ there is $x > 1$ such that $f(x) > M$. By the Intermediate Value Theorem, it follows that f takes all values in $[f(1), M]$ on the interval $(1, \infty)$, and since every real number $x \geq f(1)$ lies in some interval $[f(1), M)]$ – namely, $x \in [f(1), x]$ – it follows that

$$f([1, \infty)) \supseteq [f(1), \infty).$$

Using $\lim_{x\to\infty} f(x) = \infty$, a very similar argument shows that

$$f((-\infty, -1]) \supseteq [f(-1), \infty).$$

Let $S \coloneqq \max(f(1), f(-1))$. Then

$$[S, \infty) \subseteq [f(1), \infty) \cap [f(-1), \infty),$$

so every element $y \in [S, \infty)$ is both of the form $f(x_1)$ for some $x_1 \geq 1$ and of the form $f(x_2)$ for some $x_2 \leq -1$. Thus f is not injective. □

More precisely, assuming hypothesis (i) of Theorem 8.39 there is $m \in \mathbb{R}$ such that $f(\mathbb{R}) = [m, \infty)$, and assuming hypotheses (ii) of Theorem 8.39 there is $M \in \mathbb{R}$ such that $f(\mathbb{R}) = (-\infty, M)$. You are asked to show this in Exercise 8.25.

COROLLARY 8.40. *Let $f : \mathbb{R} \to \mathbb{R}$ be a polynomial function of even degree: that is, there is a non-negative even integer n and $a_0, \ldots, a_n \in \mathbb{R}$ with $a_n \neq 0$ such that for all $x \in \mathbb{R}$, $f(x) = a_n x^n + \ldots + a_1 x + a_0$. Then f is* not *injective.*

PROOF. If $n = 0$, then f is constant, which is certainly not injective. Now assume that n is positive. We will use the following calculus fact: since n is even, if $a_n > 0$ then $\lim_{x\to\infty} f(x) = \lim_{x\to-\infty} f(x) = \infty$, so Theorem 8.39 implies that f is not injective. Similarly, if $a_n < 0$, then $\lim_{x\to\infty} f(x) = \lim_{x\to-\infty} f(x) = -\infty$, so once again Theorem 8.39 implies that f is not injective. □

We now move on to surjectivity of functions $f : \mathbb{R} \to \mathbb{R}$.

THEOREM 8.41. *Let $f : \mathbb{R} \to \mathbb{R}$ be a continuous function.*

a) *Suppose that one of the following holds:*
   (i) $\lim_{x\to\infty} f(x) = \infty$ *and* $\lim_{x\to-\infty} f(x) = -\infty$, *or*
   (ii) $\lim_{x\to\infty} f(x) = -\infty$ *and* $\lim_{x\to-\infty} f(x) = \infty$.
   Then f is surjective.

b) *Suppose that one of the following holds:*
   (i) $\lim_{x\to\infty} f(x) = \lim_{x\to-\infty} f(x) = \infty$, *or*
   (ii) $\lim_{x\to\infty} f(x) = \lim_{x\to-\infty} f(x) = -\infty$.
   Then f is not *surjective.*

PROOF. a) Either condition (i) or condition (ii) implies that for any real number y, there is $x_1 \in \mathbb{R}$ such that $f(x_1) < y$ and $x_2 \in \mathbb{R}$ such that $f(x_2) > y$. Since f is continuous, the Intermediate Value Theorem implies to show that there is x_3 in between x_1 and x_2 such that $f(x_3) = y$. Thus f is surjective.

b) We will show that condition (i) implies that f is not surjective; the (similar) proof that condition (ii) implies that f is not surjective is left as Exercise 8.27.

Since $\lim_{x\to\infty} f(x) = \infty$ there is $b \in \mathbb{R}$ such that $f(x) \geq 0$ for all $x \geq b$. Since $\lim_{x\to-\infty} f(x) = \infty$, there is $a \leq b$ such that $f(x) \geq 0$ for all $x \leq a$. Thus f could

take negative values only on the closed bounded interval $[a, b]$. But by the Extreme Value Theorem [**Cl-HC**, Thm. 5.13] the continuous function f restriced to the closed bounded interval $[a, b]$ takes a minimum value m. Putting $\underline{m} := \min(m, 0)$, we find that $f(\mathbb{R}) \subseteq [\underline{m}, \infty)$, so f is not surjective. □

COROLLARY 8.42. *Let $n \in \mathbb{N}$, let $a_0, \ldots, a_n \in \mathbb{R}$ with $a_n \neq 0$, and let*

$$P : \mathbb{R} \to \mathbb{R},\ x \mapsto a_n x^n + \ldots + a_1 x + a_0$$

be the associated degree n polynomial function.

a) *If n is even, then P is not surjective.*
b) *If n is odd, then P is surjective.*

PROOF. a) If $n = 0$, then P is constant, hence certainly not surjective. Now suppose n is positive and even. Then:
If $a_n > 0$, then condition (i) of Theorem 8.41b) holds, so P is not surjective.
If $a_n < 0$, then condition (ii) of Theorem 8.41b) holds, so P is not surjective.
b) Suppose n is odd. Then:
If $a_n > 0$, then condition (i) of Theorem 8.41a) holds, so P is surjective.
If $a_n < 0$, then condition (ii) of Theorem 8.41b) holds, so P is surjective. □

5.4. Bijective functions.

A function $f : X \to Y$ is **bijective** if it is both injective and surjective.

A function is bijective iff for every $y \in Y$, there exists a unique $x \in X$ such that $f(x) = y$.

EXAMPLE 8.43. *Let $n \in \mathbb{Z}^+$. We claim that the function $f : \mathbb{R} \to \mathbb{R}$ by $x \mapsto x^n$ is bijective if and only if n is odd.*

First suppose that n is even. Then $f(-x) = f(x)$ for all $x \in \mathbb{R}$, so f fails to be injective. For good measure, by Theorem 8.37 the function f fails to be surjective. So this gives two reasons why f is not surjective.

Now suppose that n is odd. By Example 8.38 the function f is injective. By Theorem 8.37 the function f is surjective. So f is bijective.

The following result is easy but of the highest level of importance.

THEOREM 8.44. *For a function $f : X \to Y$, the following are equivalent:*

(i) *The function f is bijective.*
(ii) *The inverse relation $f^{-1} : Y \to X = \{(f(x), x) \mid x \in X\}$ is itself a function.*

PROOF. The inverse relation f^{-1} is a function if and only if for each $y \in Y$, there is a unique $x \in X$ such that $y = f(x)$. The existence of such an x for each $y \in Y$ means that f is surjective, and the assertion that for no $y \in Y$ are there distinct $x_1 \neq x_2$ in X such that $f(x_1) = y = f(x_2)$ means that f is injective. It follows that f^{-1} is a function if and only if f is bijective. □

5.5. Direct and Inverse Images. Let X and Y be sets, and let $f : X \to Y$ be a function.

For $A \subseteq X$, we define

$$f(A) := \{f(x) \mid x \in A\} \subseteq Y.$$

We call $f(A)$ the **direct image** of A under f.

For $B \subseteq Y$, we define

$$f^{-1}(B) \coloneqq \{x \in X \mid f(x) \in B\}.$$

We call $f^{-1}(B)$ the **inverse image** (or **preimage**) of B under f. For $y \in Y$, we write $f^{-1}(y)$ for

$$f^{-1}(\{y\}) \coloneqq \{x \in X \mid f(x) = y\}$$

and call it **the fiber of f over y**.

Here is an example showing some first uses of this terminology.

EXAMPLE 8.45. *Let $f : X \to Y$ be a function.*

a) *The function f is surjective if and only if $f(X) = Y$ if and only if, for all $y \in Y$ the fiber $f^{-1}(y)$ is nonempty.*
b) *The function f is injective if and only if for all $y \in Y$, the fiber $f^{-1}(y)$ has at most one element.*
c) *The function f is bijective if and only if for all $y \in Y$, the fiber $f^{-1}(y)$ has exactly one element.*

PROPOSITION 8.46. *Let $f : X \to Y$ be a function. Let $A_1, A_2 \subseteq X$ and $B_1, B_2 \subseteq Y$.*

a) *We have $f(A_1 \cup A_2) = f(A_1) \cup f(A_2)$.*
b) *We have $f^{-1}(B_1 \cup B_2) = f^{-1}(B_1) \cup f^{-1}(B_2)$.*
c) *We have $f(A_1 \cap A_2) \subseteq f(A_1) \cap f(A_2)$. Equality holds if f is injective.*
d) *We have $f^{-1}(B_1 \cap B_2) = f^{-1}(B_1) \cap f^{-1}(B_2)$.*

PROOF. a) Let $y \in f(A_1 \cup A_2)$. Then there is $x \in A_1 \cup A_2$ such that $f(x) = y$. We have either $x \in A_1$ or $x \in A_2$. If $x \in A_1$, then $y = f(x) \in f(A_1)$, while if $x \in A_2$, then $y = f(x) \in f(A_2)$, so either way we have $y \in f(A_1) \cup f(A_2)$.

Now let $y \in f(A_1) \cup f(A_2)$, so $y \in f(A_1)$ or $y \in f(A_2)$. If $y \in f(A_1)$, then there is $x \in A_1$ with $f(x) = y$. We then also have $x \in A_1 \cup A_2$, so $y \in f(A_1 \cup A_2)$. Similarly, if $y \in f(A_2)$, then there is $x \in A_2$ with $f(x) = y$, and since $A_2 \subseteq A_1 \cup A_2$ we have $x \in A_1 \cup A_2$ and thus $x \in f(A_1 \cup A_2)$.

b) Let $x \in f^{-1}(B_1 \cup B_2)$. Then $f(x) \in B_1 \cup B_2$, so $f(x) \in B_1$ or $f(x) \in B_2$. If $f(x) \in B_1$ then $x \in f^{-1}(B_1)$, while if $f(x) \in B_2$ then $x \in f^{-1}(B_2)$, so either way we have $x \in f^{-1}(B_1) \cup f^{-1}(B_2)$.

Let $x \in f^{-1}(B_1) \cup f^{-1}(B_2)$, so $x \in f^{-1}(B_1)$ or $x \in f^{-1}(B_2)$. If $x \in f^{-1}(B_1)$, then $f(x) \in B_1 \subseteq B_1 \cup B_2$, so $x \in f^{-1}(B_1 \cup B_2)$. Similarly, if $x \in f^{-1}(B_2)$, then $f(x) \in B_2 \subseteq B_1 \cup B_2$, so $x \in f^{-1}(B_1 \cup B_2)$.

c) Let $y \in f(A_1 \cap A_2)$. Then there is $x \in A_1 \cap A_2$ such that $f(x) = y$. In particular we have $x \in A_1$, so $y = f(x) \in f(A_1)$ and also $x \in A_2$, so $y = f(x) \in f(A_2)$. Thus $y \in f(A_1) \cap f(A_2)$.

Now suppose f is injective, and let $y \in f(A_1) \cap f(A_2)$. Since $y \in f(A_1)$ there is $x_1 \in A_1$ with $f(x_1) = y$. Since $y \in f(A_2)$ there is $x_2 \in A_2$ with $f(x_2) = y$. Thus

$$f(x_1) = y = f(x_2),$$

and since f is injective we conclude $x_1 = x_2 = x$, say, and $x \in A_1 \cap A_2$, so $y = f(x) \in f(A_1 \cap A_2)$.

d) Let $x \in f^{-1}(B_1 \cap B_2)$, so $f(x) \in B_1 \cap B_2$. Thus $f(x) \in B_1$, so $x \in f^{-1}(B_1)$, and

also $f(x) \in B_2$, so $x \in f^{-1}(B_2)$. Thus $x \in f^{-1}(B_1) \cap f^{-1}(B_2)$.

Let $x \in f^{-1}(B_1) \cap f^{-1}(B_2)$, so $f(x) \in B_1$ and $f(x) \in B_2$. Thus $f(x) \in B_1 \cap B_2$, so $x \in f^{-1}(B_1 \cap B_2)$. □

In general we *do not* have $f(A_1) \cap f(A_2) \subseteq f(A_1 \cap A_2)$. For instance, let $f : \mathbb{R} \to \mathbb{R}$ by $f(x) = x^2$, and let $A_1 = [-2, -1]$ and $A_2 = [1, 2]$. Then $A_1 \cap A_2 = \varnothing$ and $f(A_1) = f(A_2) = [1, 4]$, so

$$f(A_1 \cap A_2) = f(\varnothing) = \varnothing \subsetneq [1, 4] = f(A_1) \cap f(A_2).$$

Some further properties of direct and inverse images are given in Exercise 8.30.

If $f : X \to Y$ is a function, let

$$\mathcal{F}(f) \coloneqq \{f^{-1}(y) \mid y \in Y\}$$

be the set of fibers, and let

$$\mathcal{F}(f)^\circ \coloneqq \mathcal{F}(f) \setminus \{\varnothing\}$$

be the set of nonempty fibers of f. So $\mathcal{F}(f) = \mathcal{F}(f)^\circ$ if and only if all fibers are nonempty if and only if f is surjective.

PROPOSITION 8.47. *Let $f : X \to Y$ be a function.*

a) *The set $\mathcal{F}(f)^\circ$ of nonempty fibers of f is a partition of the domain X.*
b) *The associated equivalence relation $\sim_f$ on X is: $x_1 \sim_f x_2$ if and only if $f(x_1) = f(x_2)$.*

PROOF. a) If $x \in X$, then x lies in the fiber $f^{-1}(f(x))$, so the union of the nonempty fibers is all of X. If for $y_1, y_2 \in Y$ we have $x \in f^{-1}(y_1) \cap f^{-1}(y_2)$, then

$$y_1 = f(x) = y_2,$$

so distinct fibers are disjoint. We made sure to remove the empty set, so indeed $\mathcal{F}(f)^\circ$ is a partition of X.
b) The equivalence relation associated to a partition is the one such that two elements are equivalent if and only if they lie in the same element of the partition, so in this case $x_1 \sim_f x_2$ if and only if x_1 and x_2 lie in the same fiber over f, which means precisely that $f(x_1) = f(x_2)$. □

This is not just an example of an equivalence relation: in fact every equivalence relation on a set is of the form $\sim_f$ for a function f with domain X, and in a very specific way: if $\sim$ is an equivalence relation on a set X, let $\mathcal{P} = \{\mathfrak{c}(x) \mid x \in X\}$ be the associated partition, which is just the set of $\sim$-equivalence classes. Then we have a function

$$\mathfrak{c} : X \to \mathcal{P}$$

given by

$$x \mapsto \mathfrak{c}(x).$$

That is, we just map each element of x to the equivalence class that contains it. Then $\mathfrak{c}$ is a surjective function: every equivalence class is of the form $\mathfrak{c}(x)$, so it is mapped to by x (and possibly by other points). For this map, the fiber over $\mathfrak{c}(x)$ is the set of $y \in X$ such that $\mathfrak{c}(y) = \mathfrak{c}(x)$....which *is* $\mathfrak{c}(x)$, the equivalence class of x.

6. Composition and Inverse Functions

Perhaps the single most important property of functions is that they can, under the right circumstances, be *composed*.[8] For instance, in calculus, complicated functions are built up out of simple functions by plugging one function into another, e.g. $\sqrt{x^2+1}$, or $e^{\sin x}$, and the most important differentiation rule – the Chain Rule – tells how to find the derivative of a composition of two functions in terms of the derivatives of the original functions.

Let $f : X \to Y$ and $g : Y \to Z$: that is, the codomain of f is equal to the domain of g. Then we can define a new function $g \circ f : X \to Z$ by:

$$x \mapsto g(f(x)).$$

REMARK 8.48. *The expression $g \circ f$ means first perform f and then perform g. Thus function composition proceeds from right to left, counterintuitively at first. There was a time when this bothered mathematicians enough to suggest writing functions* on the right*, i.e., $(x)f$ rather than $f(x)$. But that time is past.*

REMARK 8.49. *The condition for composition can be somewhat relaxed: it is not necessary for the domain of g to equal the codomain of f. What is precisely necessary and sufficient is that for every $x \in X$, $f(x)$ lies in the domain of g, i.e.,*

$$f(X) \subseteq \text{Codomain}(g).$$

EXAMPLE 8.50. *The composition of functions is generally* not *commutative. In fact, if $g \circ f$ is defined, $f \circ g$ need not be defined at all. For instance, suppose $f : \mathbb{R} \to \mathbb{R}$ is the function which takes every rational number to 1 and every irrational number to 0 and $g : \{0,1\} \to \{a,b\}$ is the function $0 \mapsto b$, $1 \mapsto a$. Then $g \circ f : \mathbb{R} \to \{a,b\}$ is defined: it takes every rational number to a and every irrational number to b. But $f \circ g$ makes no sense at all:*

$$f(g(0)) = f(b) = ???.$$

Even when $g \circ f$ and $f \circ g$ are both defined – e.g. when $f, g : \mathbb{R} \to \mathbb{R}$, they need not be equal. This is again familiar from precalculus mathematics. If $f(x) = x^2$ and $g(x) = x + 1$, then

$$g(f(x)) = x^2 + 1, \text{ whereas } f(g(x)) = (x+1)^2 = x^2 + 2x + 1.$$

REMARK 8.51. *Those who have taken linear algebra will notice the analogy with the multiplication of matrices: if A is an $m \times n$ matrix and B is an $n \times p$ matrix, then the product AB is defined, an $m \times p$ matrix. But if $m \neq p$, the product BA is not defined. (In fact this is more than an analogy, since an $m \times n$ matrix A can be viewed as a linear transformation $L_A : \mathbb{R}^n \to \mathbb{R}^m$. Matrix multiplication is indeed a special case of composition of functions.)*

On the other hand, function composition is always **associative**: if $f : X \to Y$, $g : Y \to Z$ and $h : Z \to W$ are functions, then we have

$$(h \circ g) \circ f = h \circ (g \circ f).$$

The proof consists of observing that both sides map $x \in X$ to $h(g(f(x)))$.[9]

[8] This is a special case of the composition of relations, but since that was optional material, we proceed without assuming any knowledge of that material.

[9] As above, this provides a conceptual reason behind the associativity of matrix multiplication.

6.1. Basic facts about injectivity, surjectivity and composition.

Here we establish a small number of very important facts about how injectivity, surjectivity and bijectivity behave with respect to function composition. First:

THEOREM 8.52. *Let $f : X \to Y$ and $g : Y \to Z$ be two functions.*

a) *If f and g are injective, then so is $g \circ f$.*
b) *If f and g are surjective, then so is $g \circ f$.*
c) *If f and g are bijective, then so is $g \circ f$.*

PROOF. a) We must show that for all $x_1,\ x_2 \in X$, if $g(f(x_1)) = g(f(x_2))$, then $x_1 = x_2$. But put $y_1 = f(x_1)$ and $y_2 = f(x_2)$. Then $g(y_1) = g(y_2)$. Since g is assumed to be injective, this implies $f(x_1) = y_1 = y_2 = f(x_2)$. Since f is also assumed to be injective, this implies $x_1 = x_2$.
b) We must show that for all $z \in Z$, there exists at least one x in X such that $g(f(x)) = z$. Since $g : Y \to Z$ is surjective, there exists $y \in Y$ such that $g(y) = z$. Since $f : X \to Y$ is surjective, there exists $x \in X$ such that $f(x) = y$. Then $g(f(x)) = g(y) = z$.
c) Finally, if f and g are bijective, then f and g are both injective, so by part a) $g \circ f$ is injective. Similarly, f and g are both surjective, so by part b) $g \circ f$ is surjective. Thus $g \circ f$ is injective and surjective, i.e., bijective, qed. □

Now we wish to explore the other direction: suppose we know that $g \circ f$ is injective, surjective or bijective? What can we conclude about the "factor" functions f and g?

The following example shows that we need to be careful.

EXAMPLE 8.53. *Let $X = Z = \{0\}$, let $Y = \mathbb{R}$. Define $f : X \to Y$ be $f(0) = \pi$ (or your favorite real number; it would not change the outcome), and let f be the constant function which takes every real number y to 0: note that this is the unique function from $\mathbb{R}$ to $\{0\}$. We compute $g \circ f$: $g(f(0)) = g(\pi) = 0$. Thus $g \circ f$ is the identity function on X: in particular it is bijective. However, both f and g are far from being bijective: the range of f is only a single point $\{\pi\}$, so f is not surjective, whereas g maps every real number to 0, so is not injective.*

On the other hand, something is true: namely the "inside function" f is injective, and the outside function g is surjective. This is in fact a general phenomenon.

THEOREM 8.54. *Let $f : X \to Y$ and $g : Y \to Z$ be functions.*

a) *If $g \circ f$ is injective, then f is injective.*
b) *If $g \circ f$ is surjective, then g is surjective.*
c) *If $g \circ f$ is bijective, then f is injective and g is surjective.*

PROOF. a) We proceed by contraposition. If f is not injective, there are $x_1 \neq x_2$ in X such that $f(x_1) = f(x_2)$. But then $g(f(x_1)) = g(f(x_2))$, so that the distinct points x_1 and x_2 become equal under $g \circ f$: that is, $g \circ f$ is not injective.
b) Again by contraposition: suppose that g is not surjective: then there exists $z \in Z$ such that for no y in Y do we have $z = g(y)$. But then we certainly cannot have an $x \in X$ such that $z = g(f(x))$, because if so taking $y = f(x)$ shows that z is in the range of g, contradiction.
c) If $g \circ f$ is bijective, it is injective and surjective, so we apply parts a) and b). □

6.2. Inverse Functions.

Finally we come to the last piece of the puzzle: let $f : X \to Y$ be a function. We know that the inverse relation f^{-1} is a function if and only if f is injective and surjective. But there is another (very important) necessary and sufficient condition for invertibility in terms of function composition. Before stating it, recall that for a set X, the identity function 1_X is the function from X to X such that $1_X(x) = x$ for all $x \in X$. (Similarly $1_Y(y) = y$ for all $y \in Y$.)

We say that a function $g : Y \to X$ is the **compositional inverse** of $f : X \to Y$ if both of the following hold:

(CI1) $g \circ f = 1_X$: i.e., for all $x \in X$, $g(f(x)) = x$.
(CI2) $f \circ g = 1_Y$: i.e., for all $y \in Y$, $f(g(y)) = y$.

In other words, g is the compositional inverse f if applying one function and then the other – in either order! – brings us back where we started.

Just as the identity element for a binary operation is unique if it exists, the compositional inverse of a function is unique if it exists.

LEMMA 8.55. *A function $f : X \to Y$ has at most one compositional inverse: iif $g : Y \to X$ and $h : Y \to X$ are both compositional inverses of f, then $g = h$.*

PROOF. Since g is a compositional inverse of f, we have $g \circ f = 1_X$. Now apply $\circ h$ to both sides:

$$g = g \circ 1_Y = g \circ (f \circ h) = (g \circ f) \circ h = 1_X \circ h = h. \qquad \square$$

Two natural questions here are: when does a function $f : X \to Y$ have a compositional inverse g; and when g exists, how can we describe it in a simple way in terms of f? It turns out that both of these questions can be answered in terms of bijectivity and the inverse relation f^{-1}. Here is the crucial result:

THEOREM 8.56. *Let $f : X \to Y$ be a function.*

a) *The following are equivalent:*
   (i) *The function f is bijective.*
   (ii) *The inverse relation $f^{-1} : Y \to X$ is a function.*
   (iii) *The function f has a compositional inverse g.*
b) *When the equivalent conditions of part a) hold, then f^{-1} is the compositional inverse of f.*

PROOF. a) We already know that (i) $\iff$ (ii): this is Theorem 8.44 above. (ii) $\implies$ (iii): Suppose that the inverse relation f^{-1} is a function. We claim that it is the compositional inverse of f. Indeed:

Let $x \in X$. Then $(x, f(x))$ lies in the relation f, so $(f(x), x)$ lies in the relation f^{-1}, which means that $f^{-1}(f(x)) = x$.

Let $y \in Y$. Since (ii) $\iff$ (i), we know that f is bijective, so there is a unique $x \in X$ such that $y = f(x)$. Again this means that $(x, f(x))$ lies in the relation f, so $(y, x) = (f(x), x)$ lies in the relation f^{-1}, so $x = f^{-1}(y)$. This means that

$$f(f^{-1}(y)) = f(x) = y.$$

We have shown that

$$f^{-1} \circ f = 1_X \text{ and } f \circ f^{-1} = 1_Y,$$

so f^{-1} is the compositional inverse of f.
(iii) $\implies$ (i): Let g be a compositional inverse of f. Then we have $g \circ f = 1_X$. The identity function 1_X is injective, so by Theorem 8.54 the function f is injective. Similarly, we have $f \circ g = 1_Y$, and since the identity function 1_Y is surjective, by Theorem 8.54 the function f is surjective. Therefore f is bijective.[10]
b) In the proof of part a) we showed that if f^{-1} is a function then it is a compositional inverse of f. By Lemma 8.55 a function can have at most one compositional inverse, so f^{-1} is *the* compositional inverse of f. □

Exercise 8.46 sharpens Theorem 8.56: a function has a "left inverse" if and only if it is injective, and a function has a "right inverse" if and only if it is surjective.

In summary, for a function f, being bijective, having the inverse relation (obtained by "reversing all the arrows") be a function, and having another function g which undoes f by composition in either order, are all equivalent. Knowing this, we no longer need to speak of the "compositional inverse" of a bijective function as it is nothing other than the inverse relation f^{-1}.

7. Functions Between Finite Sets

EXAMPLE 8.57.

a) *The arctangent function* $\arctan : \mathbb{R} \to \mathbb{R}$ *is injective (Example 8.36) but not surjective: its image is* $(\frac{-\pi}{2}, \frac{\pi}{2})$.
b) *The function* $f : \mathbb{R} \to \mathbb{R}$ *given by* $f(x) = x^3 - x$ *is surjective (Theorem 8.37) but not injective: we have* $f(-1) = f(0) = f(1) = 0$.

In the above example, we saw that there are functions from $\mathbb{R}$ to $\mathbb{R}$ that are injective but not surjective and also functions from $\mathbb{R}$ to $\mathbb{R}$ that are surjective but not injective. In Exercise 8.37 you are asked to show that for every infinite set X, there is a function $f : X \to X$ that is injective but not surjective and also a function $g : X \to X$ that is surjective but not injective.

However, for functions from a finite set to itself, this is not possible. The following result proves this and in fact something mildly stronger.

THEOREM 8.58. *Let* $n \in \mathbb{N}$, *and let* X, Y *be finite sets with* $\#X = \#Y = n$. *For a function* $f : X \to Y$, *the following are equivalent:*

(i) *The function* f *is bijective.*
(ii) *The function* f *is injective.*
(iii) *The function* f *is surjective.*

PROOF. Certainly (i) $\implies$ (ii) and (i) $\implies$ (iii), so it suffices to show (ii) $\implies$ (i) and (iii) $\implies$ (i).
(ii) $\implies$ (i): Suppose f is injective. We may write $X = \{x_1, \ldots, x_n\}$; for $1 \leq i \leq n$, put $y_i := f(x_i)$. Since f is injective, the elements $y_1, \ldots, y_n$ are all distinct, so $\{y_1, \ldots, y_n\}$ is an n-element subset of Y. But $\#Y = n$, so

$$Y = \{y_1, \ldots, y_n\} = \{f(x_1), \ldots, f(x_n)\}$$

[10]Note that a very similar argument shows that g is also bijective.

and thus f is surjective. Being injective and surjective, f is bijective.
(iii) $\implies$ (i): Suppose f is surjective. We may write $Y = \{y_1, \ldots, y_n\}$. Since f is surjective, for $1 \leq i \leq n$ there is an element $x_i \in X$ such that $f(x_i) = y_i$. We claim that the list $x_1, \ldots, x_n$ is irredundant: if not, we remove each instance of the same element in the list after the first to get a shorter irreundant list, say $x_{i_1}, \ldots, x_{i_m}$ with $m < n$, which still has the property that $\{f(x_{i_1}), \ldots, f(x_{i_m})\} = Y$. (This is because we have only removed entries from the list that get mapped to the same element of Y as does an earlier entry on the list.) But this means that Y, a set with n elements, is the set associated to a finite list with $m < n$ elements, contradicting Exercise 1.5a). So $x_1, \ldots, x_n$ is irredundant. Since $\#X = n$, we must have $X = \{x_1, \ldots, x_n\}$. It follows that f is injective: for all $1 \leq i \neq j \leq n$ we have $f(x_i) = y_i \neq y_j = f(x_j)$. □

REMARK 8.59. *There is an analogue of Theorem 8.58 in linear algebra: if F is a field, V and W are F-vector spaces of equal, finite dimension, and $L : V \to W$ is a linear map, then L is injective $\iff$ L is surjective $\iff$ L is bijective.*

PROPOSITION 8.60. *Let X and Y be finite nonempty sets, and recall that Y^X denotes the set of all functions $f : X \to Y$. We have*

$$\#Y^X = (\#Y)^{\#X}.$$

PROOF. Let $m = \#X$, $n = \#Y$, and write

$$X = \{x_1, \ldots, x_m\}, \ Y = \{y_1, \ldots, y_n\}.$$

To define a function $f : \{x_1, \ldots, x_m\} \to \{y_1, \ldots, y_n\}$, we must map x_1 to one of the n elements $y_1, \ldots, y_n$, must map x_2 to one of the n elements $y_1, \ldots, y_n$, and so forth, finally mapping x_m to one of the n elements $y_1, \ldots, y_n$. None of these assignments places any restriction on another assignment, so by the Principle of Independent Choices (Proposition 3.4), we have

$$n \cdots n = n^m = (\#Y)^{\#X}$$

choices overall. □

Earlier we suggested that a finite list of length n with entries in a set X could be viewed as an element of the Cartesian product $X^n = \prod_{i=1}^n X$. It is also useful to view a finite list as a certain kind of function. Namely, for $n \in \mathbb{N}$, a finite list

$$\ell : x_1, \ldots, x_n$$

of length n with entries in a set X can be viewed as a function $f : [n] \to X$ just by putting $f(i) = x_i$. Putting these two observations together suggests identifying the Cartesian product X^n with the set $X^{[n]}$ of all functions $f : [n] \to X$, which is also reasonable, since there is an evident bijection between them: we map $(x_1, \ldots, x_n)$ to the function $i \mapsto x_i$. After identifying finite lists with functions, an irredundant finite list with entires in X is precisely an injective function $f : [n] \to X$.

THEOREM 8.61. *Let X be a nonempty set, and let Y be a finite set of size $n \in \mathbb{Z}^+$.*

a) *If X has more than n elements, there is no injective function $\iota : X \hookrightarrow Y$.*
b) *If X is finite of size $m \leq n$, then the number of injective functions $\iota : X \hookrightarrow Y$ is $P(n, m) = n(n-1)\cdots(n-m+1)$.*

PROOF. a) If X has more than n elements, then it admits a finite subset Z with size $n+1$. If $\iota : X \hookrightarrow Y$ is an injection, then $\iota|_Z : Z \hookrightarrow Y$ is also an injection. If we write the elements of Z as $z_1, \ldots, z_{n+1}$, then because ι is injective, then $\iota(z_1), \ldots, \iota(z_{n+1})$ are $n+1$ different elements of Y, a contradiction.
b) We go by induction on m. The base case is $m = 1$. Every function from a 1 element set to Y is injective, and the number of such functions is

$$(\#Y)^1 = n = P(n, 1).$$

Now let $m \geq 2$ and suppose that the number of injections from a finite set of size $m-1$ to Y is $P(n, m-1)$. Write $X = \{x_1, \ldots, x_{m-1}\}$. An injective function $\iota : X \to Y$ is obtained by mapping x_1 to any element $y_\bullet \in Y$ and then defining an injective function $\iota' : \{x_2, \ldots, x_m\} \to Y \setminus \{y_\bullet\}$, and conversely.[11] Therefore, the number of injective functions $\iota : X \to Y$ is

$$\begin{aligned}(\#Y)P(n-1, m-1) &= n \cdot (n-1)(n-2)\cdots((n-1)-(m-1)+1) \\ &= n(n-1)\cdots(n-m+1) = P(n, m).\end{aligned}$$

□

Theorem 8.61 is a restatement of the Pigeonhole Principle. It seems more basic and convincing: eventually the charm of introducing pigeons into a mathematical argument begins to fade. A similar reformulation of the Strong Pigeonhole Principle is given in Exercise 8.38.

7.1. Stirling Numbers. Now let us count the number of surjective functions $q : X \to Y$. The analogue of Theorem 8.61a) is similarly easy to establish:

PROPOSITION 8.62. *Let X be a finite set of size $m \in \mathbb{N}$, and let Y be a set. If Y has more than m elements, there is no surjective function $q : X \to Y$.*

PROOF. Since $m \geq 0$ and Y has more than m elements, Y is nonempty. If $m = 0$ then X is empty, and then

$$f(X) = f(\varnothing) = \varnothing \subsetneq Y,$$

so f is not surjective. So we may assume that $m \geq 1$ and write $X = \{x_1, \ldots, x_m\}$. If f is surjective, then every element of Y is of the form $f(x_i)$ for some $1 \leq i \leq m$. But this gives at most m different elements of Y, and Y has more than m elements, so f cannot be surjective. □

Now suppose we have finite sets X and Y with $\#X = m$ and $\#Y = n$ and such that $m \geq n \geq 1$. It is easy to see that there is a surjection $f : X \to Y$ in this case: indeed, if we write $X = \{x_1, \ldots, x_m\}$ and $Y = \{y_1, \ldots, y_n\}$, we can map $x_i \mapsto y_i$ for all $1 \leq i \leq n$ and map the remaining $m-n$ elements of X anywhere we like. However determining the number of surjections in this case is much more interesting than the corresponding result for injections (Theorem 8.61). One reason for this is that it is natural to build an injection with finite domain X "inductively" as in the proof of Theorem 8.61, i.e., by a sequence of injections on larger and larger subsets of X. However it does not seem possible to construct a surjection inductively (or recursively, or in any similar way).

We need a different idea. First, it should be clear that the number of surjective

[11] That is, given any $y_\bullet \in Y$ and any injective function $\iota' : \{x_2, \ldots, x_m\} \to Y \setminus \{y_\bullet\}$, extending ι to a function on X by mapping x_1 to $y_\bullet$ gives an injection.

functions from an m element set to an n element set is the number of surjective functions from $[m]$ to $[n]$. So let us put

$$\mathcal{S}(m,n) \coloneqq \{f : [m] \to [n] \mid f \text{ is surjective}\}.$$

If it helps, it would suffice to count the number of elements in the complementary $[n]^{[m]} \setminus \mathcal{S}(m,n)$. And here come the idea: for $1 \leq i \leq n$, let

$$A_i \coloneqq \{f : [m] \to [n] \mid i \notin f([m])\}$$

be the set of functions for which i does not lie in the image. A function fails to be surjective if and only if *some* $1 \leq i \leq n$ does not lie in the image, so

$$[n]^{[m]} \setminus \mathcal{S}(m,n) = \bigcup_{i=1}^{n} A_i.$$

Now the Inclusion-Exclusion Principle will solve the problem! Indeed, it tells us:

$$\#([n]^{[m]} \setminus \mathcal{S}(m,n)) = S_1(A) - S_2(A) + \ldots + (-1)^{n+1} S_N(A),$$

where for $1 \leq j \leq n$, $S_j(A)$ is the sum of the sizes of the sets $\bigcap_{i \in J} A_i$ as J ranges over all j-element subsets of $[n]$. For $1 \leq j \leq n$, let J be any j-element subset of $[n]$. Then

$$\bigcap_{i \in J} Y_i = \{f : [m] \to [n] \mid f([m]) \cap J = \varnothing\} = \{f : [m] \to ([n] \setminus J)\}$$

$$= (\#([n] \setminus J))^{\#[m]} = (n-j)^m.$$

We got the same answer for each j-element subset $J \subseteq [n]$ and there are $\binom{n}{j}$ such subsets, so overall we get

$$n^m - \#\mathcal{S}(m,n) = \#([n]^{[m]} \setminus \mathcal{S}(m,n)) = \sum_{j=1}^{n} (-1)^{j+1} \binom{n}{j} (n-j)^m.$$

Solving this for $\#\mathcal{S}(m,n)$, we get the following result.

THEOREM 8.63. *Let $m \geq n$ be positive integers. The number of surjecctions from a finite set with m elements to a finite set with n elements is*

$$\#\mathcal{S}(m,n) = \sum_{j=0}^{n} (-1)^j \binom{n}{j} (n-j)^m.$$

It turns out that Theorem 8.63 gives us, almost for free, the solution of another interesting counting problem. Namely, for positive integers $m \geq n$, let $\left\{ {m \atop n} \right\}$ denote the number of n element partitions of $[m]$: that is, the number of ways that we can express an m-element set as the disjoint union of n nonempty sets.

We notice that if instead we had $m < n$, then there would be no n-element partition of $[m]$: a disjoint union of n nonempty sets has size at least n. Conversely, for $m \geq n \geq 1$ there is certainly at least one n-element partition of $[m]$, e.g.

$$\mathcal{P} = \{\{1\}, \{2\}, \ldots, \{n-1\}, \{n, n+1, \ldots, m\}\}.$$

So for all $m \geq n \geq 1$, we have that $\left\{ {m \atop n} \right\}$ is a positive integer. The following result evaluates this positive integer.

THEOREM 8.64. *For integers $m \geq n \geq 1$, we have*

$$\left\{ {m \atop n} \right\} = \frac{\#\mathcal{S}(m,n)}{n!} = \frac{1}{n!} \sum_{j=0}^{n} (-1)^j \binom{n}{j} (n-j)^m.$$

PROOF. Let $f : [m] \to [n]$ be a surjection. Recall Proposition 8.47: for any function $f : X \to Y$, the nonempty fibers determine a partition $\mathcal{P}_f$ of X. Because f is moreover surjective, every fiber is nonempty and therefore in our case we get a partition $\mathcal{P}_f$ of $[m]$ with $\#[n] = n$ fibers.

To complete the proof, it suffices to show: for every n-element partition $\mathcal{P}$ of $[m]$, there are precisely $n!$ surjections $f : [m] \to [n]$ with $\mathcal{P} = \mathcal{P}_f$: if so, we have

$$\#\mathcal{S}(m,n) = n! \left\{ {m \atop n} \right\}.$$

The idea of this is simple: if we know the partition $\mathcal{P}$ then we know what the fibers should be as a set, so what remains is the order in which to put them. So let $\mathcal{P} = \{X_1, \ldots, X_n\}$ be an n-element partition of $[m]$. To give a surjection $f : [m] \to [n]$ with associated partition $\mathcal{P}_f = \mathcal{P}$ we need to assign to each X_i some $\iota(i) \in [n]$ such that every $j \in [n]$ is of the form $\iota(i)$ for a unique $i \in [n]$: then we get a surjection $f : [m] \to [n]$ that maps each element of X_i to $\iota(i)$, so the fiber over $\iota(i)$ is indeed X_i. The number of ways to do this is just the number of bijections $\iota : [n] \to [n]$, which by Theorem 8.58 is also the number of injections $\iota : [n] \to [n]$, which in turn by Theorem 8.61b) is $n!$. □

The numbers $\left\{ {m \atop n} \right\}$ are called **Stirling[12] numbers of the second kind**. They arise in combinatorics and also in the calculus of finite differences. One may notice that $\left\{ {m \atop n} \right\}$ is notationally similar to the binomial coefficient $\binom{m}{n}$. There are in fact some non-notational similarities as well: Exercises 8.39 and 8.40 concern the Stirling numbers; the former shows that they satisfy a recursion reminisent of (12).

7.2. Bell numbers. For $n \in \mathbb{N}$, we define the **nth Bell[13] number** B_n as the number of partitions of the n-element set $[n]$.

In Exercise 8.41 you are asked to show that for any finite set S with $\#S = n$, the unmber of partitions of S is B_n.

EXAMPLE 8.65.

a) *The unique partition of* $[0] = \varnothing$ *is* $\mathcal{P} = \varnothing$, *so* $B_0 = 1$.
b) *The unique partition of* $[1] = \{1\}$ *is* $\mathcal{P} = \{\{1\}\}$, *so* $B_1 = 1$.
c) *The partitions of* $[2] = \{1,2\}$ *are the discrete partition* $\{\{1\},\{2\}\}$ *and the trivial partition* $\{\{1,2\}\}$, *so* $B_2 = 2$.
d) *The partitions of* $[3] = \{1,2,3\}$ *are:*

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

so $B_3 = 5$.

[12]James Stirling (1692-1770) was a Scottish mathematician.

[13]Eric Temple Bell (1883-1960) was a Scottish born mathematician and science fiction writer who moved to the United States as a young adult.

Already for $n = 4$ it seems desirable to compute B_n in some other way besides actually writing down all the partitions. The following simple result gives a recursive formula for B_n.

THEOREM 8.66. *For all $n \in \mathbb{N}$, we have*

$$B_{n+1} = \sum_{k=0}^{n} \binom{n}{k} B_k. \tag{46}$$

PROOF. We will give a combinatorial proof.[14] That is, we know that B_{n+1} counts the number of partitions of $[n+1]$; so we will show that $\sum_{k=0}^{n} \binom{n}{k} B_k$ *also* counts the number of partitions of $[n+1]$.

For this, the key is to see how to sort the partitions of $[n+1]$ according to a parameter k that ranges from 0 to n. Here is one way to do it: if $\mathcal{P}$ is a partition of $[n+1]$, then there is a unique set (or "part") $S \in \mathcal{P}$ that contains, say, 1 as an element. How many other elements does S contain? It can be – aha – any number $0 \leq k \leq n$. So now we fix $0 \leq k \leq n$ and count the number of partitions of $[n+1]$ for which the part containing 1 has cardinality $k+1$ and thus has exactly k other elements. There are $\binom{n}{k}$ ways to choose these other elements – they can be anything other than 1 – and then we are left with a set of $n+1-(k+1) = n-k$ elements, for which there are B_{n-k} partitions. Thus overall the number of partitions of $[n+1]$ in which the part containing 1 has size $k+1$ is $\sum_{k=0}^{n} \binom{n}{k} B_{n-k}$.

This is not quite the formula we claimed...but don't panic. Let $K := n-k$, so also $k = n-K$. As k ranges from 0 to n, then so does K, so

$$B_{n+1} = \sum_{k=0}^{n} \binom{n}{k} B_{n-k} = \sum_{K=0}^{n} \binom{n}{n-K} B_K.$$

Since $\binom{n}{n-K} = \binom{n}{K}$, we get

$$B_{n+1} = \sum_{K=0}^{n} \binom{n}{K} B_K = \sum_{k=0}^{n} \binom{n}{k} B_k. \qquad \square$$

EXAMPLE 8.67. *By Example 8.65, we know that $B_0 = B_1 = 1$, $B_2 = 2$ and $B_3 = 5$. Let us use this and (46) to compute a few more values.*

a) *Taking $n = 3$ in (46), we get*

$$B_4 = \sum_{k=0}^{3} \binom{3}{k} B_k = B_0 + 3B_1 + 3B_2 + B_3 = 1 + 3 \cdot 1 + 3 \cdot 2 + 1 \cdot 5 = 15.$$

b) *Taking $n = 4$ in (46), we get*

$$\begin{aligned} B_5 &= \sum_{k=0}^{4} \binom{4}{k} B_k = B_0 + 4B_1 + 6B_2 + 4B_3 + B_4 \\ &= 1 + 4 \cdot 1 + 6 \cdot 2 + 4 \cdot 5 + 15 = 52. \end{aligned}$$

More Bell numbers are recorded at `https://oeis.org/A000110`*, along with other information about the sequence. The terms grow rapidly; e.g.*

$$B_{25} = 4638590332229999353.$$

[14] This is not so surprising, since after all B_n is *defined* as the cardinality of a finite set and at the moment we know nothing else about it.

The Bell numbers are also clearly related to the Stirling numbers: indeed, for $n, k \in \mathbb{Z}^+$, B_n counts all the partitions of $[n]$, whereas $\left\{ {n \atop k} \right\}$ counts the partitions $\mathcal{P}$ of $[n]$ into k parts: i.e., with $\#\mathcal{P} = k$. Since the number of parts of a partition of $[n]$ must be at least 1 and at most n, we have:

$$\forall n \in \mathbb{Z}^+,\ B_n = \sum_{k=1}^{n} \left\{ {n \atop k} \right\}. \tag{47}$$

In fact, Proposition 8.62 gives $\left\{ {n \atop k} \right\} = 0$ when $k > n$, so we can give a somewhat sneaky generalization:

$$\forall n \in \mathbb{Z}^+, \forall M \geq n,\ B_n = \sum_{k=1}^{M} \left\{ {n \atop k} \right\}. \tag{48}$$

We can take advantage of (48) and the formula for $\left\{ {n \atop k} \right\}$ given in Theorem 8.64 to give a remarkable infinite series representation for B_n due to G. Dobiński [**D077**].[15]

THEOREM 8.68 (Dobiński's Formula). *For all $n \in \mathbb{Z}^+$, we have*

$$B_n = \frac{1}{e} \sum_{k=0}^{\infty} \frac{k^n}{k!}. \tag{49}$$

PROOF. Let $n \in \mathbb{Z}^+$: We start with Theorem 8.64:

$$\left\{ {n \atop k} \right\} = \frac{1}{k!} \sum_{j=0}^{k} (-1)^j \binom{k}{j} (k-j)^n.$$

Making the substitution $j \mapsto k - j$ and writing out $\binom{k}{j} = \frac{k!}{j!(k-j)!}$, we get

$$\left\{ {n \atop k} \right\} = \sum_{j=0}^{k} (-1)^{k-j} \frac{j^n}{j!(k-j)!} = \sum_{j=1}^{k} (-1)^{k-j} \frac{j^{n-1}}{(j-1)!(k-j)!}.$$

Combining with (48) we get that for all $M \geq n$,

$$\begin{aligned} B_n = \sum_{k=1}^{M} \left\{ {n \atop k} \right\} &= \sum_{k=1}^{M} \sum_{j=1}^{k} (-1)^{k-j} \frac{j^{n-1}}{(j-1)!(k-j)!} \\ &= \sum_{j=1}^{M} \frac{j^{n-1}}{(j-1)!} \sum_{k=j}^{M} \frac{(-1)^{k-j}}{(k-j)!} \\ &= \sum_{j=1}^{M} \frac{j^{n-1}}{(j-1)!} \left(\sum_{s=0}^{M-r} \frac{(-1)^s}{s!} \right). \end{aligned}$$

Now we may take the limit as $M \to \infty$: the parenthesized expression approaches $\sum_{s=0}^{\infty} \frac{(-1)^s}{s!} = \frac{1}{e}$, so overall we get

$$B_n = \frac{1}{e} \sum_{j=1}^{\infty} \frac{j^{n-1}}{(j-1)!} = \frac{1}{e} \sum_{k=0}^{\infty} \frac{k^n}{k!}. \qquad \square$$

[15]Dobiński's paper was published in 1877, whereas Eric Temple Bell was born in 1883. We deduce that Bell was *not* the first to study Bell numbers. Many mathematical names are like this.

In Exercise 8.43 you are asked to derive the "exponential generating function" for the Bell numbers.

Before we move on, we want to indicate a very elegant approach to the Bell numbers due to G.-C. Rota [**Ro64**]. His approach uses some linear algebra. Let V be the $\mathbb{R}$-vector space of all polynomial functions $f : \mathbb{R} \to \mathbb{R}$: that is, there is $d \in \mathbb{N}$ and $a_0, \ldots, a_d \in \mathbb{R}$ such that

$$\forall x \in \mathbb{R},\ f(x) = a_d x^d + \ldots + a_1 x + a_0.$$

This is an infinite-dimensional vector space that comes with a well-known basis, the monomials $\{x^n\}_{n=0}^\infty$. (Our convention is that $x^0 = 1$.) However, any sequence $\{P_n\}_{n=0}^\infty$ in which the degree of P_n (i.e., the highest power of x that appears in P_n) is n for all $n \in \mathbb{N}$ is also a basis of V. This applies in particular to the **falling factorials**: we put $(x)_0 \coloneqq 1$ and

$$\forall n \in \mathbb{Z}^+, (x)_n \coloneqq x(x-1)\cdots(x-n+1).$$

Certainly $(x)_n$ has degree n. Since a linear functional $L : V \to \mathbb{R}$ is uniquely determined by its values on a basis, there is a unique such L such that

$$\forall n \in \mathbb{N}, L((x)_n) = 1.$$

Now for $x, n \in \mathbb{Z}^+$, consider a function $f : [n] \to [x]$. As in Proposition 8.47 above, f determines an equivalence relation on $[n]$ in which $i \sim j$ if and only if $f(i) = f(j)$; let $\pi = \pi_f$ be the associated partition. of $[n]$. The size of π is the size of the image of f; let us call this $N(\pi)$. The partition π_f does not uniquely determine f: to get f back we need to give an injective function from π_f to $[x]$, and the number of these is $(x)_{N(\pi)}$. As we have accounted for all x^n functions this way, we get the identity

$$\sum_\pi (x)_{N(\pi)} = x^n. \tag{50}$$

where the sum extends over all partitions of $[n]$. We may view both sides of (50) as polynomial functions of m, and the fact that it holds for all $m \in \mathbb{Z}^+$ implies that it holds as a polynomial identity. (If two polynomials are equal at all positive integers, then their difference is a polynomial with roots at all the positive integers; but a nonzero polynomial of degree d has at most d roots by the Root-Factor Theorem of high school algebra, so the difference is identitically zero.) Applying the linear functional L to this identity, we get:

$$\forall n \in \mathbb{Z}^+, B_n = L(x^n). \tag{51}$$

As an example of the usefulness of (51), we will rederive Theorem 8.66. We have

$$\forall n \in \mathbb{Z}^+,\ x(x-1)_n = (x)_{n+1},$$

so

$$L(x(x-1)_n) = 1 = L((x)_{n+1}).$$

But now we observe that the function $F : V \to \mathbb{R}$ by

$$F(p) \coloneqq L(xp(x-1)) - L(p(x))$$

is an $\mathbb{R}$-linear map. Since this linear map vanishes on the basis $\{(x)_n\}_{n=0}^\infty$ of V, it follows that it is identically zero: that is,

$$\forall p \in V, L(xp(x-1)) = L(p(x)).$$

Applying this with $p(x) = (x+1)^n$, we get

$$B_{n+1} = L(x^{n+1}) = L(x(x+1-1)^n) = L(x+1)^n = \sum_{k=0}^{n} \binom{n}{k} L(x^k) = \sum_{k=0}^{n} \binom{n}{k} B_k,$$

which is (46).

7.3. An application to ring theory. An **integral domain** is a nondegenerate ($1 \neq 0$) commutative ring R such that for all $x, y \in R$, if $xy = 0$ then $x = 0$ or $y = 0$, and a field is a nondegenerate commutative ring in which each nonzero element has a multiplicative inverse.

In a commutative ring R, an element x is a **zero divisor** if there is $0 \neq y \in R$ such that $xy = 0$. Because $0 \cdot 1 = 0$, in any nondegenerate ring the element 0 is a zero divisor. A nondegenerate commutative ring is an integral domain if and only if 0 is the only zero divisor.

PROPOSITION 8.69. *Let R be a nondegenerate (i.e., $0 \neq 1$) commmutative ring.*

a) *For an element $x \in R$, the following are equivalent:*
   (i) *The element x is* not *a zero divisor.*
   (ii) *The map $x\bullet : R \to R$ by $y \mapsto xy$ is an injection.*

b) *The following are equivalent:*
   (i) *The ring R is an integral domain.*
   (ii) *For all $x \in R \setminus \{0\}$, the map $x\bullet : R \to \mathbb{R}$ is injective.*

c) *For an element $x \in R$, the following are equivalent:*
   (i) *The element x is a unit in R: that is, there is $y \in R$ such that $xy = 1$.*
   (ii) *The map $x\bullet$ is a bijection.*
   (iii) *The map $x\bullet$ is a surjection.*

d) *The following are equivalent:*
   (i) *The ring R is a field.*
   (ii) *For all $x \in R \setminus \{0\}$, the map $x\bullet$ is a bijection.*
   (iiii) *For all $x \in R \setminus \{0\}$, the map $x\bullet$ is a surjection.*

PROOF. a) We show that $\neg$ (i) $\iff$ $\neg$ (ii): First suppose that x is a zero divisor. Then there is $y \in R \setminus \{0\}$ such that $xy = 0$. Thus

$$(x\bullet)(y) = xy = 0 = x \cdot 0 = (x\bullet)(0),$$

so $x\bullet$ is not injective. Now suppose that $x\bullet$ is not injective, so that there are $y_1 \neq y_2$ in R such that $xy_1 = (x\bullet)(y_1) = (x\bullet)(y_2) = xy_2$. It follows that $y_1 - y_2 \neq 0$ and $x(y_1 - y_2) = 0$, so x is not a zero divisor.
b) If R is an integral domain and $x \in R \setminus \{0\}$, then x is not a zero divisor, so by part a) the map $x\bullet$ is injective. Conversely, if for all $x \in R \setminus \{0\}$, the map $x\bullet$ is injective, then by part a) we have that no $x \in R \setminus \{0\}$ is a zero divisor, so R is an integral domain.
c) (i) $\implies$ (ii): If $y \in R$ is such that $xy = 1$, then $(x\bullet)$ and $(y\bullet)$ are inverse functions:

$$((x\bullet) \circ (y\bullet))(z) = (x\bullet)(yz) = xyz = z,$$
$$((y\bullet) \circ (x\bullet))(z) = (y\bullet)(xz) = yxz = xyz = z.$$

So $x\bullet$ is bijective by Theorem 8.56.
(ii) $\implies$ (iii) is immediate: every bijection is a surjection.
(iii) $\implies$ (i): If $x\bullet$ is a bijection, then 1 lies in its image: there is $y \in R$ such that

$1 = (x\bullet)(y) = xy$.
d) (i) $\implies$ (ii): If R is a field, then every nonzero $x \; in R$ is a unit, so by part c) we have that $(x\bullet)$ is injective.
(ii) $\implies$ (iii) is again immediate.
(iii) $\implies$ (i): If $(x\bullet)$ is surjective for all nonzero $x \in R$, then by part c) every nonzero $x \in R$ is a unit, so R is a field. □

THEOREM 8.70. *If R is a finite integral domain, then R is a field.*

PROOF. Let R be a finite integral domain, and let $x \in R \setminus \{0\}$. By Proposition 8.69b), the function $(x\bullet) : R \to R$ is injective. Since R is finite, Theorem 8.58 implies that $(x\bullet)$ is a bijection. Now Proposition 8.69d) implies that R is a field. □

Of course Theorem 8.70 fails if the word "finite" is removed from the statement, since the ring $\mathbb{Z}$ is an infinite integral domain that is not a field.

8. Exercises

EXERCISE 8.1. *Write down all* 16 *relations from the set $X = \{a, b\}$ to the set $Y = \{1, 2\}$.*

EXERCISE 8.2. *Let X and Y be nonempty sets, at least one of which is infinite. Show: $\mathcal{R}(X, Y)$ is infinite.*

EXERCISE 8.3. *Let R be a relation on the set X.*

a) *Show that the following are equivalent:*
   (i) *The relation R is both symmetric and anti-symmetric.*
   (ii) *We have $R \subseteq \Delta_X = \{(x, x) \mid x \in X\}$.*
b) *Show that the following are equivalent:*
(i) *The relation R is both an equivalence relation and a partial ordering.*
(ii) *We have $R = \Delta_X$.*
c) *Show: if X has at least two elements, then there is no relation on R that is both an equivalence relation and a total ordering.*

EXERCISE 8.4. *Which of the relations in Examples 1.1 through 1.15 are symmetric?*

EXERCISE 8.5. *Which of the relations in Examples 1.1 through 1.16 are anti-symmetric?*

EXERCISE 8.6. *Let X be a set.*

a) *If X has at most* 1 *element, show that every relation on X is transitive.*
b) *If X has at least* 2 *elements, show that there is a relation on X that is not transitive.*

EXERCISE 8.7. *Let R be a relation on X. Show the following are equivalent:*

(i) *The relation R is both symmetric and anti-symmetric.*
(ii) *The relation R is a subrelation of the equality relation.*

EXERCISE 8.8. *Show: a relation R on a set X is strongly anti-symmetric if and only if it is anti-reflexive and anti-symmetric.*

EXERCISE 8.9.

a) *Find a relation R on a set X that is reflexive and symmetric but not transitive.*

b) *Find a relation R on a set X that is reflexive and transitive but not symmetric.*
c) *Find a relation R on a set X that is symmetric and transitive but not reflexive.*

EXERCISE 8.10. *In practice, it rarely seems to be the case that a relation fails to be an equivalence relation only because it lacks the reflexive property. This exercise gives an explanation for this.*

Let R be a relation on a set X that is symmetric and transitive. Show that the following are equivalent:

(i) *The domain of R (recall that this is $\{x \in X \mid (x,y) \in R \text{ for some } y \in X\}$ is all of X.*
(ii) *The relation R is an equivalence relation.*

EXERCISE 8.11. *Which of the relations in Examples 1.1 through 1.16 are equivalence relations? Which are partial orderings?*

EXERCISE 8.12.

a) *Show that the composition of relations is associative: if*

$$R \subseteq W \times X,\ S \subseteq X \times Y,\ T \subseteq Y \times Z$$

are relations, then

$$(R \circ S) \circ T = R \circ (S \circ T).$$

b) *For a set X, we put $\Delta_X \coloneqq \{(x,x) \mid x \in X\}$. Show: for any relation $R \subseteq X \times Y$, we have*

$$R \circ \Delta_X = R,\ \Delta_Y \circ R = R.$$

EXERCISE 8.13. *Let $R \subseteq X \times Y$ and $\subseteq Y \times Z$ be relations. Show:*

$$(S \circ R)^{-1} = R^{-1} \circ S^{-1}.$$

EXERCISE 8.14. *Let X be a nonempty set.*

a) *Show that the empty relation on X satisfies $R \subseteq R^{(2)}$ but is not reflexive.*
b) *Suppose X has at least three elements. Show that the relation*

$$R \coloneqq (X \times X) \setminus \Delta = \{(x,y) \mid x \neq y\}$$

is anti-reflexive, but $R^{(2)} = X \times X$, so $R \subseteq R^{(2)}$.

EXERCISE 8.15. *Let R be a relation on a set X.*

a) *Suppose there are positive integers $m < n$ such that $R^{(m)} = R^{(n)}$. Show that the transitive closure of R is*

$$R_t = \bigcup_{i=0}^{n-1} R^{(i)}.$$

b) *Under the hypothesis of part a), suppose moreover that R is reflexive. Show that there is $N \in \mathbb{Z}^+$ such that $R^{(N)} = R^{(N+1)}$ and that for any such N (e.g. the least such N!) we have $R_t = R^{(N)}$.*

EXERCISE 8.16. *Let R be a relation on a set X.*

a) *Show: R is symmetric if and only if R_r is symmetric.*
b) *Show: if R is transitive, then R_r is transitive.*

c) *Show: The relation* $R \coloneqq (\{1,2,3\} \times \{1,2,3\}) \setminus \{(1,3)\}$ *on* $X = [3]$ *is not transitive, but* $R_s = [3] \times [3]$ *is.*
d) *Let* R *be the relation* $<$ *on* $\mathbb{R}$*. Show:* R *is transitive, but* R_s *is not transitive.*

EXERCISE 8.17. *Which of the following relations from* $\mathbb{R}$ *to* $\mathbb{R}$ *are functions?*

a) $R_1 = \{(x,y) \in \mathbb{R}^2 \mid x^2 = y^2\}$.
b) $R_2 = \{(x,y) \in \mathbb{R}^2 \mid x^3 = y^3\}$.
c) $R_3 = \{(x,y) \in \mathbb{R}^2 \mid e^x = y\}$.
d) $R_4 = \{(x,y) \in \mathbb{R}^2 \mid e^y = x\}$.

EXERCISE 8.18. *Let* X *be a set and let* $R \subseteq X \times X$ *be a relation on* X*. For a subset* Y *of* X*, we define a relation* R_Y *on* Y *by*

$$R_Y \coloneqq R \cap (Y \times Y).$$

That is, R_Y *consists of all elements* $(y_1, y_2) \in R$ *such that* $y_1, y_2 \in Y$*. We call* R_Y ***the restriction of R to Y****.*

a) *For each of the following properties, show that if* R *has that property, then so does* R_Y*:*
   (i) *Reflexivity.*
   (ii) *Symmetry.*
   (iii) *Anti-Symmetry.*
   (iv) *Transitivity.*
   (v) *Totality.*
b) *Deduce: if* R *is an equivalence relation, then* R_Y *is an equivalence relation on* Y*.*
c) *Deduce: if* R *is a partial ordering, then* R_Y *is a partial ordering on* Y*.*
d) *Deduce: if* R *is a total ordering, then* R_Y *is a total ordering on* Y*.*

EXERCISE 8.19. *Let* X *be the set of all functions* $f : \mathbb{R} \to (0, \infty)$*. We define a relation,* ***asymptotic equality*** *on* X*, as follows: we put* $f \sim g$ *if* $\lim_{x\to\infty} \frac{f(x)}{g(x)} = 1$*.*

a) *Show: asymptotic equality is an equivalence relation on* X*.*
b) *Let*

$$f(x) = a_n x^n + \ldots + a_1 x + a_0$$

be any polynomial function that such that $f(x) > 0$ *for all* $x \in \mathbb{R}$*. Show that* n *is an* even *integer,* $a_n > 0$ *and* $f \sim a_n x^n$*.*

EXERCISE 8.20. *Let* $f : X \to Y$ *be a function, and let* Z *be a subset of* X*. We define a relation* $f_{|Z} \subseteq Z \times Y$ *by*

$$f|_Z \coloneqq \{(z,y) \in Z \times Y \mid y = f(z)\}.$$

a) *Show:* $f|_Z : Z \to Y$ *is a function.*
b) *Show: if* f *is injective, then* $f|_Z : Z \to Y$ *is injective.*
c) *Give an example to show that if* f *is surjective, then* $f|_Z : Z \to Y$ *need not be surjective.*

EXERCISE 8.21. *Show: if* $f : \mathbb{R} \to \mathbb{R}$ *is a strictly decreasing function, then* f *is injective.*

EXERCISE 8.22. *Prove Theorem 8.35b).*

EXERCISE 8.23. *Let $n \geq 3$ be an odd integer. Exhibit a degree n polynomial function $P : \mathbb{R} \to \mathbb{R}$ – i.e.,*

$$P(x) = a_n x^n + \ldots + a_1 x + a_0,\ a_0, \ldots, a_n \in \mathbb{R},\ a_n \neq 0$$

that is not *injective.*

EXERCISE 8.24. *Complete the proof of Theorem 8.41b).*

EXERCISE 8.25. *Let $f : \mathbb{R} \to \mathbb{R}$ be continuous.*

a) *Show: if*
$$\lim_{x \to \infty} f(x) = \lim_{x \to -\infty} f(x) = \infty,$$
then there is $m \in \mathbb{R}$ such that $f(\mathbb{R}) = [m, \infty)$.

b) *Show: if*
$$\lim_{x \to \infty} f(x) = \lim_{x \to -\infty} f(x) = -\infty,$$
then there is $M \in \mathbb{R}$ such that $f(\mathbb{R}) = (-\infty, M]$.

EXERCISE 8.26. *Let $n \in \mathbb{Z}^+$*

a) *Show: if n is even then for all $x, y \in \mathbb{R}$, we have $x^n = y^n \iff x = \pm y$. Deduce that if if $x, y \geq 0$ then $x^n = y^n \iff x = y$.*

b) *Show: if n is odd, then for all $x, y \in \mathbb{R}$ we have $x^n = y^n \iff x = y$.*

c) *Do the results of parts a) and b) continue to hold in any number system satisfying the ordered field axioms?*

EXERCISE 8.27. *Complete the proof of Theorem 8.39 by showing that if a continuous function $f : \mathbb{R} \to \mathbb{R}$ satisfies $\lim_{x \to \infty} f(x) = \lim_{x \to -\infty} f(x) = -\infty$, then f is not injective.*

EXERCISE 8.28. *Let $f : X \to Y$.*

a) *Show: $f \circ 1_X = f$.*

b) *Show: $1_Y \circ f = f$.*

EXERCISE 8.29. *Show: for any set X, the identity function $1_X : X \to X$ by $1_X(x) = x$ is bijective.*

EXERCISE 8.30. *Let $f : X \to Y$ be a function.*

a) *Show: if $A_1 \subseteq A_2 \subseteq X$, then $f(A_1) \subseteq f(A_2)$.*

b) *Show: if $B_1 \subseteq B_2 \subseteq Y$, then $f^{-1}(B_1) \subseteq f^{-1}(B_2)$.*

c) *Show that the following are equivalent:*
   (i) *For all $A_1, A_2 \subseteq X$ we have $f(A_1 \cap A_2) = f(A_1) \cap f(A_2)$.*
   (ii) *The function f is injective.*

EXERCISE 8.31. *Let $f : X \to X$ be a function.*

a) *Show that the following are equivalent:*
   (i) *The function f is injective.*
   (ii) *For all $n \in \mathbb{N}$, the function $f^{\circ n}$ is injective.*

b) *Show that the following are equivalent:*
   (i) *The function f is surjective.*
   (ii) *For all $n \in \mathbb{N}$, the function $f^{\circ n}$ is surjective.*

c) *Show that the following are equivalent:*
   (i) *The function f is bijective.*
   (ii) *For all $n \in \mathbb{N}$, the function $f^{\circ n}$ is bijective.*

EXERCISE 8.32. *Let $f : X \to Y$ be a function.*

a) *Show: for all $A \subseteq X$ we have $f^{-1}(f(A)) \supseteq A$.*

b) *Show that the following are equivalent:*
   (i) *For all $A \subseteq X$ we have $f^{-1}(f(A)) = A$.*
   (ii) *The function f is injective.*

c) *Show: for all $B \subseteq Y$ we have $f(f^{-1}(B)) \subseteq B$.*

d) *Show that the following are equivalent:*
   (i) *For all $B \subseteq Y$ we have $f(f^{-1}(B)) = B$.*
   (ii) *The function f is surjective.*

EXERCISE 8.33. *Let $f : X \to X$ be a bijection, and let $Y \subseteq X$ be a subset such that $f(Y) = Y$. Show that $f(X \setminus Y) = X \setminus Y$ and that*

$$f|_{X \setminus Y} : (X \setminus Y) \to (X \setminus Y)$$

is a bijection.

EXERCISE 8.34. *Let $f : X \to Y$ and $g : Y \to X$ be functions. Suppose that* either *$g \circ f = 1_X$ or $f \circ g = 1_Y$. Also suppose that* either *f or g is bijective. (Thus we are assuming two things, in four possible ways.) Show: f and g are inverse functions.*

EXERCISE 8.35. *Let $f : X \to Y$ and $g : Y \to Z$ be bijective functions, so $g \circ f : X \to Z$ is also bijective and thus f^{-1}, g^{-1} and $(g \circ f)^{-1}$ also exist. Show the* ***Shoes and Socks*** *property:*

$$(g \circ f)^{-1} = f^{-1} \circ g^{-1}.$$

EXERCISE 8.36. *Let X be a set, and denote by $\mathcal{F}(X)$ the set of all functions $f : X \to X$. We view function composition as a binary operation on $\mathcal{F}(X)$:*

$$\circ : \mathcal{F}(X) \times \mathcal{F}(X) \to \mathcal{F}(X).$$

a) *Suppose that X has at most one element. Show: $\mathcal{F}(X) = \{1_X\}$. Deduce that for all $f, g \in \mathcal{F}(X)$, we have $f \circ g = g \circ f$.*

b) *Let $f \in \mathcal{F}(X) \setminus \{1_X\}$: thus, there is $x \in X$ such that $f(x) \neq x$. Show: there is $g \in \mathcal{F}(X)$ such that $g \circ f \neq f \circ g$.*
   (Suggestion: g can be taken to be a constant function.)

c) *Deduce: if X has at least two elements, $\circ$ is not commutative on $\mathcal{F}(X)$.*

EXERCISE 8.37. *Let X be an infinite set.*

a) *Show: there is a function $f : X \to X$ is injective and not surjective.*

b) *Show: there is a function $g : X \to X$ that is surjective and not injective.*

(Suggestion: use the fact that there is an injection $\iota : \mathbb{Z}^+ \hookrightarrow X$.)

EXERCISE 8.38 (Strong Pigeonhole Principle Reformulated). *Let X be a set, let Y be a finite set of size $n \in \mathbb{Z}^+$, and let $f : X \to Y$ be a function. Show: if for some $k \in \mathbb{Z}^+$ the set X has more than $(k-1)n$ elements, then at least one fiber of f has at least k elements.*

EXERCISE 8.39 (Recursion for Stirling Numbers).

a) *For all $m, n \in \mathbb{N}$, we may define $\left\{ {m \atop n} \right\}$ to be the number of n-element partitions of $[m]$, a non-negative integer. Show that the following are equivalent:*
   (i) *Either $m \geq n \geq 1$ or $m = n = 0$.*

(ii) *We have* $\left\{ {m \atop n} \right\} \geq 1$.

b) *Show: for all* $m \in \mathbb{N}$ *and* $n \in \mathbb{Z}^+$ *we have*

$$\left\{ {m+1 \atop n} \right\} = n \left\{ {m \atop n} \right\} + \left\{ {m \atop n-1} \right\}.$$

EXERCISE 8.40. *Let* $n \in \mathbb{Z}^+$.

a) *Show:* $\left\{ {n \atop n-1} \right\} = \binom{n}{2}$.

b) *Show:* $\left\{ {n \atop 2} \right\} = 2^{n-1} - 1$.

EXERCISE 8.41. a) *For a set* X, *let* $\mathrm{Part}(X)$ *be the set of partitions of* X. *Let* $f : X \to Y$ *be a bijection. Define an induced bijection*

$$\mathrm{Part}(f) : \mathrm{Part}(X) \to \mathrm{Part}(Y).$$

Your definition should be "canonical": no choices need to be made.

b) *Recall that* $B_n = \#\,\mathrm{Part}([n])$. *Let* X *be a finite set. Show:*

$$\#\,\mathrm{Part}(X) = B_{\#X}.$$

EXERCISE 8.42. *Show: for all* $n \in \mathbb{Z}^+$, *we have* $B_n \leq n!$. *(Suggestion: use induction.)*

The following exercise is for those with some analytic background (differential equations and complex analysis).

EXERCISE 8.43. *Define a function* $B(x) := \sum_{n=0}^{\infty} \frac{B_n}{n!} x^n$.

a) *Use Exercise 8.42 to show that the series defining* $B(x)$ *converges (at least) for all* $x \in (-1, 1)$, *so* B *determines an infinitely differentiable function on* $(-1, 1)$.

b) *Show:* $B(0) = 1$ *and* $B'(x) = e^x B(x)$. *(Suggestion: use Theorem 8.66.)*

c) *Let* $f(x) = e^{e^x - 1}$. *Show that* $f(0) = 1$ *and* $f'(x) = e^x f(x)$. *Deduce that* $f(x) = B(x)$ *for all* $x \in (-1, 1)$.

d) *The function* f *is an entire function in the sense of complex analysis. Deduce from this that the power series defining* $B(x)$ *converges for all* $x \in \mathbb{C}$ *and thus:*

$$\forall x \in \mathbb{C},\ B(x) = e^{e^x - 1}.$$

EXERCISE 8.44 (Factorization Principle). *Let* $f : X \to Z$ *and* $g : X \to Y$ *be functions.*

a) *Show that the following are equivalent:*

(i) *There is a function* $h : Y \to Z$ *such that* $f = h \circ g$.

(ii) *For all* $x_1, x_2 \in X$, *if* $g(x_1) = g(x_2)$, *then* $f(x_1) = f(x_2)$.

(Hint: If $y = g(x)$, *then we want to define* $h(y) = f(x)$. *For this to be well-defined, if also* $g(x) = y = g(x')$, *then we want* $f(x) = f(x')$. *Condition (ii) ensures this.)*

When the equivalent conditions of part a) hold, we say that ***f factors through g****.*

b) *Suppose that the conditions of part a) hold. Show that there is a* unique *function* $h : Y \to Z$ *such that* $f = h \circ g$ *if and only if* g *is surjective.*

EXERCISE 8.45. *Let* $f : X \to Y$ *and* $g : Y \to X$ *be functions such that* $g \circ f = 1_X$*: that is,*

$$\forall x \in X,\ g(f(x)) = x.$$

a) *Show by example that f and g need not be inverse functions.*
b) *Suppose that f is surjective. Show that f and g are inverse functions.*
c) *Suppose that g is injective. Show that f and g are inverse functions.*

EXERCISE 8.46. *Let $f : X \to Y$ be a function.*

a) *Show that the following are equivalent:*
   (i) *The function f has a **left inverse**: there is $g : Y \to X$ such that $g \circ f = 1_X$.*
   (ii) *The function f is injective.*
   (One way to do this is to apply Exercise 8.44a).)
b) *Show that the following are equivalent:*
   (i) *The function f has a **right inverse**: there is $g : Y \to X$ such that $f \circ g = 1_X$.*
   (ii) *The function f is surjective.*

EXERCISE 8.47. *For each of the following functions $f : \mathbb{R} \to \mathbb{R}$, determine whether f is injective and whether f is surjective:*

a)
$$f_1(x) = \prod_{i=1}^{2023} (x - i) = (x-1)(x-2)\cdots(x-2023).$$

b)
$$f_2(x) = \prod_{i=1}^{2024} (x - i) = (x-1)(x-2)\cdots(x-2024).$$

c)
$$f_3(x) = 7^{8^{9^x}}.$$

d)
$$f_4(x) = \arctan x.$$

CHAPTER 9

Applications

1. Dynamics

Let X be a nonempty set, and let $f : X \to X$ be a function. Then we can composite f with itself...repeatedly. For $n \in \mathbb{Z}^+$ we let $f^{\circ n} \coloneqq f \circ \cdots f$ be the n-fold composition of f with itself. We put $f^{\circ 0}(x) = x$ (that is, $f^{\circ 0} = 1_X$ is the identity function on X).

For $x \in X$, the **forward orbit of x under f** is

$$\overrightarrow{\mathcal{O}}_f(x) \coloneqq \{f^{\circ n}(x) \mid n \in \mathbb{N}\}.$$

This notation is a bit heavy; if it is clear what f is, we may abbreviate it to $\overrightarrow{\mathcal{O}}(x)$.

The relation "y lies in the forward orbit of x under f" is transitive: if z lies in the forward orbit of y and y lies in the forward orbit of x, then there are $n_1, n_2 \in \mathbb{N}$ such that $z = f^{\circ n_1}(y)$ and $y = f^{\circ n_2}(x)$, so

$$z = f^{\circ n_1}(y) = f^{\circ n_1}(f^{\circ n_2}x) = f^{\circ n_1 + n_2}(x),$$

so z lies in the forward orbit of x.

For $a, b \in X$, we put $a \sim_f b$ if the forward orbits intersect: $\overrightarrow{\mathcal{O}}(a) \cap \overrightarrow{\mathcal{O}}(b) \neq \varnothing$. More explicitly, this means there are $n_1, n_2 \in \mathbb{N}$ such that

$$f^{\circ n_1}(a) = f^{\circ n_2}(b).$$

Again, if f is understood, we write $a \sim b$ instead of $a \sim_f b$.

PROPOSITION 9.1. *Let X be a nonempty set, and let $f : X \to X$ be a function. The relation $\sim_f$ is an equivalence relation on X.*

PROOF. As usual, we must show that $\sim_f$ is reflexive, symmetric and transitive. As we will see, only the latter requires any real work.
Reflexivity: for all $x \in X$, we have $x \in \overrightarrow{\mathcal{O}}_f(x) \cap \overrightarrow{\mathcal{O}}_f(x)$.
Symmetry: evidently we have $\overrightarrow{\mathcal{O}}_f(x) \cap \overrightarrow{\mathcal{O}}_f(y) \neq \varnothing$ if and only if $\overrightarrow{\mathcal{O}}_f(y) \cap \overrightarrow{\mathcal{O}}_f(x) \neq \varnothing$.
Transitivity: Let $x, y, z \in X$, and suppose that $x \sim_f y$ and $y \sim_f z$. Then there are $a, b, c, d \in \mathbb{N}$ such that

$$f^{\circ a}(x) = f^{\circ b}(y) \text{ and } f^{\circ c}(y) = f^{\circ d}(z),$$

and then we have

$$f^{\circ(c+a)}(x) = f^{\circ c}(f^{\circ a}(x)) = f^{\circ c}(f^{\circ b}(y)) = f^{\circ(b+c)}(y)$$
$$= f^{\circ b}(f^{\circ c}(y)) = f^{\circ b}(f^{\circ d}(z)) = f^{\circ(b+d)}(z),$$

so $x \sim_f z$. □

For $f : X \to X$ and $x \in X$, let $\rfloor_f(x)$ be the $\sim_f$ equivalnce class of x. Then

$$f(\mathfrak{c}_f(x)) \subseteq \mathfrak{c}_f(x)).$$

We say that $x \in X$ is **preperiodic** for f if there are natural numbers $n_1 < n_2$ such that $f^{\circ n_1}(x) = f^{\circ n_2}(x)$. We claim that x is preperiodic for x if and only if the forward orbit $\overrightarrow{\mathcal{O}}_f(x)$ is finite. Indeed, if x is preperiodic, let n_1 be the least natural number for which there is $n_2 > n_1$ with $f^{\circ n_1}(x) = f^{\circ n_2}$. Then if $k := n_2 - n_1$, then the forward orbit $\overrightarrow{\mathcal{O}}_f(x)$ consists of the n_1 distinct elements $x, f(x), \ldots, f^{\circ(n_1-1)}(x)$ (they must be distinct by the minimality of n_1) followed by the k distinct elements

$$f^{\circ n_1}(x), f^{\circ(n_1+1)}(x), \ldots, f^{\circ(n_2-1)}(x)$$

repeated infinitely in a cycle. In particular we have

$$\#\overrightarrow{\mathcal{O}}_f(x) = n_2.$$

If x is not preperiodic, then the forward orbit sequence

$$x, f(x), f(f(x)), \ldots, f^{\circ n}(x), \ldots$$

consists of distinct elements.

In particular, if X is finite, then for all $x \in X$ the forward orbit $\overrightarrow{\mathcal{O}}_f(x)$ must be finite, so all points are preperiodic under all maps. If X is infinite then all, some or none of the points may be preperiodic.

A point $x \in X$ is **periodic** under f if it is preperiodic with $n_1 = 0$: equivalently, if there is $n \in \mathbb{Z}^+$ such that $f^{\circ n}(x) = x$. In this case, the k above is the least $n \in \mathbb{Z}^+$ for which $f^{\circ n}(x) = x$, and the orbit sequence is an infinitely repeated k-cycle. We say that x is **k-periodic**.

Example 9.2.

a) *Let X be any set, and let $1_X : X \to X$ be the identity map. Then every $x \in X$ is 1-periodic. (Conversely if $f : X \to X$ is such that every $x \in X$ is 1-periodic under f, then $f = 1_X$.) In this case, for each $x \in X$, the $\sim_f$ equivalence class of x is just $\mathfrak{c}(x) = \{x\}$.*

*In general we call a 1-periodic point a **fixed point** of f: this is $x \in X$ for which $f(x) = x$.*

b) *Let $X = \mathbb{Z}$ and let $\iota : \mathbb{Z} \to \mathbb{Z}$ by $\iota(n) = -n$. Then 0 is a fixed point of f and every $n \in \mathbb{Z} \setminus \{0\}$ is 2-periodic with orbit sequence*

$$n, -n, n, -n, \ldots.$$

In this class, the $\sim_f$ equivalence class of 0 is just $\{0\}$ while the $\sim_f$ equivalence class of a nonzero integer n is the two-element set $\{n, -n\}$.

c) *Let $X = \mathbb{N}$ and define $f : \mathbb{N} \to \mathbb{N}$ as follows: for $0 \leq n \leq 2020$, let $f(n) := n + 1$. For $n \geq 2021$, let $f(n) := 0$. Then every $n \in \mathbb{N}$ is preperiodic under f, and the periodic points are $0, \ldots, 2021$, which are all 2022-periodic. In this case there is just one $\sim_f$ equivalence class.*

Example 9.3. a) *Let $k \in \mathbb{Z}^+$, let $X = [k]$, and define $f : [k] \to [k]$ as follows: for $1 \leq i \leq k-1$ we put $f(i) = i+1$, and we put $f(k) = 1$. Then every $x \in X$ is k-periodic, and there is a single $\sim_f$-equivalence class.*

b) *Let $X = \mathbb{Z}$ and let $\tau : \mathbb{Z} \to \mathbb{Z}$ by $\tau(n) = n + 1$. For all $N \in \mathbb{Z}$ we have*

$$\overrightarrow{\mathcal{O}}_\tau(N) = \mathbb{Z}^{\geq N}.$$

In particular there are no preperiodic points under τ. In this case there is just one $\sim_f$ equivalence class.

PROPOSITION 9.4. *If $f : X \to X$ is injective, then every preperiodic point is periodic.*

PROOF. We prove the contrapositive: suppose that x is preperiodic for f but not periodic. Let $n_1 < n_2$ be as above: in particular $f^{\circ n_1}(x) = f^{\circ n_2}(x)$. Then $f^{\circ(n_1-1)}(x)$ and $f^{\circ(n_2-1)}(x)$ are distinct but map under f to the same point $f^{\circ n_1}(x) = f^{\circ n_2}(x)$, so f is not injective. □

1.1. Bijective Dynamics. Suppose now that $f : X \to X$ is *bijective*. (Recall that if X is finite, by Theorem 8.58 this holds if f is either injective or surjective.) In this case it is relatively easy to understand the $\sim_f$ equivalence classes, as we will now describe. Since f is bijective, it has an inverse function f^{-1}, and this allows us to define **backwards iterates** of f: e.g. $f^{\circ -2}(x) = f^{-1}(f^{-1}(x))$ In general, for $n \in \mathbb{Z}^{\leq 1}$ we put

$$f^{\circ n} := (f^{-1})^{\circ |n|}.$$

We can then define the **two-sided orbit sequence**

$$\ldots, f^{\circ -3}(x), f^{-1}(f^{-1}(x)), f^{-1}(x), x, f(x), f(f(x)), f^{\circ 3}(x), \ldots$$

and the **total orbit**

$$\overleftrightarrow{\mathcal{O}}_f(x) := \{f^{\circ n}(x) \mid n \in \mathbb{Z}\}.$$

PROPOSITION 9.5. *Let $f : X \to X$ be bijective, and let $x \in X$.*

a) *The following are equivalent:*
   (i) *The point x is preperiodic for f.*
   (ii) *The point x is periodic for f.*
   (iii) *The forward orbit $\overrightarrow{\mathcal{O}}_f(x)$ is finite.*
   (iv) *The total orbit $\overleftrightarrow{\mathcal{O}}_f(x)$ is finite.*

b) *Suppose that the equivalent conditions of part a) hold, and let k be the least positive integer such that $f^{\circ k}(x) = x$. Then $\#\overrightarrow{\mathcal{O}}_f(x) = \#\overleftrightarrow{\mathcal{O}}_f(x) = k$ and*

$$\overrightarrow{\mathcal{O}}_f(x) = \overleftrightarrow{\mathcal{O}}_f(x) = \{x, f(x), f(f(x)), \ldots, f^{\circ(k-1)}(x)\}.$$

Moreover the two-sided orbit sequence of x consists of the k-cycle

$$x, f(x), f(f(x)), \ldots, f^{\circ(k-1)}(x)$$

repeated infinitely in both directions.

c) *Suppose that the equivalent conditions of part a)* do not *hold. Then all elements of the two-sided orbit sequence are distinct.*

A good way to think about Proposition 9.5 is that if $f : X \to X$ is bijective, then on each $\sim_f$ equivalence class $\mathfrak{c}_f(x)$ the map $f : \mathfrak{c}_f(x) \to \mathfrak{c}_f(x)$ looks either like the map of Example 9.3a) – i.e., it is a k-cycle for some $k \in \mathbb{Z}^+$ – or like the map of Example 9.3b), which one can think of as a "bi-infinite cycle." In this case for each $x \in X$, the function restricted to $\mathfrak{c}_f(x)$ induces a bijection on $\mathfrak{c}_f(x)$.

Let us focus on the case in which X is finite: say $X = \{x_1, \ldots, x_n\}$ has size n. If $f : X \to X$ is a bijection, then Theorem 9.5 gives us a partition of X, say

$$X \coloneqq \coprod_{i=1}^{r} Y_i,$$

where each Y_i is one full orbit under f. We may order the Y_i's so that if $k_1 \coloneqq \#Y_1, \ldots, k_r \coloneqq \#Y_r$ then $k_1 \geq \ldots \geq k_r$; moreover the Sum Theorem gives us

$$k_1 + \ldots + k_r = n.$$

We call the tuple $(k_1, \ldots, k_r)$ the **cycle type** of the bijection f. For instance, if $f : [n] \to [n]$ is the map

$$1 \mapsto 2 \mapsto \ldots \mapsto n \mapsto 1,$$

then f has cycle type (n), while the identity map on $[n]$ has cycle type $(1, 1, \ldots, 1)$.

We are now in a position to complete an argument that we left unfinished long ago: namely, that if one is trying to prove the equivalence of N statements using N basic implications, then the implications must be arranged in a circle. We already showed that to establsh the equivalence, we need each of the n statements to appear exactly once as a hypothesis of an implication and also to appear exactly once as a conclusion of an implication, and thus we get a bijection $f : [N] \to [N]$ by putting $f(i) = j$ if and only if $P_i \implies P_j$. Then our basic implications suffice to show that P_i and P_j are equivalent if and only $i \sim_f j$. So we have established the equivalence of all n statements if and only if we have exactly one $\sim_f$-equivalence class if and only if we have exactly one cycle.

1.2. Injective Dynamics. Now suppose (only) that $f : X \to X$ is injective. Then for all $x \in X$, there is at most one $x_{-1} \in X$ such that $f(x_{-1}) = x$, so we can attempt to iterate the map backwards, but we may have a stopping point. Namely, for $n \in \mathbb{Z}^{<0}$ we put $f^{\circ n}(x) = y$ if $f^{\circ |n|}(y) = x$. By Exercise 8.31, the function $f^{\circ |n|}$ is also injective, so there is at most one such $y \in Y$. There are three possibilities:

(i) The point x is preperiodic for f. Then by Proposition 9.4, since f is injective, x is periodic, we have that $f^{\circ n}(x)$ is defined for all $n < 0$, and $\mathfrak{c}_f(x) = \overrightarrow{\mathcal{O}}_f(x)$ is a k-cycle.
(ii) The point x is not preperiodic for f and for all $n \in \mathbb{Z}^+$ there is $y \in X$ with $f^{\circ n}(y) = x$. Then $f^{\circ n}(x)$ is defined for all n, and $\mathfrak{c}_f(x)$ is the **bi-infinite cycle**

$$\ldots f^{\circ -3}(x), f^{-1}(f^{-1}(x)), f^{-1}(x), x, f(x), f(f(x)), f^{\circ 3}(x), \ldots.$$

(iii) The point x is not preperiodic for f and there is some $N \leq 0$ such that $f^{\circ N}(x)$ is defined but $f^{\circ (N-1)}(x)$. The equivalence class $\mathfrak{c}_f(x)$ is the **singly infinite cycle**

$$f^{\circ N}(x), \ldots, f^{\circ -1}(x), x, f(x), f(f(x)), \ldots, f^{\circ n}(x), \ldots.$$

In this case, the first point in the cycle, $f^{\circ N}(x)$ does not lie in the image of f, while all the other points do. In fact, there is exactly one $\sim_f$ equivalence class that is a singly infinite cycle for each $x \in X \setminus f(X)$, namely

$$x, f(x), f(f(x)), \ldots, f^{\circ n}(x), \ldots.$$

THEOREM 9.6 (Dedekind-Schröder-Bernstein). *Let $f : X \to Y$ and $g : Y \to X$ be injections. Then there is a bijection from X to Y.*

PROOF. Put $Z \coloneqq X \coprod Y$, the disjoint union of X and Y. We define a function

$$\Phi : Z \to Z$$

by: if $x \in X$, then $\Phi(x) \coloneqq f(x)$, while if $y \in Y$, then $\Phi(y) \coloneqq g(y)$. Then the function Φ is also injective: if $x \in X$ is such that $x = \Phi(z)$ for some $z \in Z$, then we must have $z \in Y$ and $x = g(z)$; since g is injective, there is at most one $z \in Y$ with this property. Similarly, if $y \in Y$ is such that $y = \Phi(z)$ for some $z \in Z$, then we must have $z \in X$ and $y = f(x)$; since f is injective, there is at most one $x \in X$ with this property. Moreover, Φ is bijective if and only if both f and g are bijective.

If Φ is bijective, then we are really done: each of f and g^{-1} is a bijection from X to Y. So we may assume that Φ is not bijective. As mentioned just above, each element $z \in Z \setminus \Phi(Z)$ is the first element of a singly infinite cycle under Φ.

Supose first that $z \in X$. Then we may write $z = x_1 \in X$, $y_1 = \Phi(x_1) \in Y$, $x_2 = \Phi(y_1) \in X$, and so forth: in general, having defined $x_1, \ldots, x_n \in X$ and $y_1, \ldots, y_n \in Y$ we put $y_{n+1} = \Phi(x_n) \in Y$ and $x_{n+1} = \Phi(y_n) \in X$. With this notation, the singly infinite cycle starting at z is

$$x_1, y_1, x_2, y_2, \ldots, x_n, y_n, \ldots.$$

So we redefine Φ on $\mathfrak{c}_\Phi(z)$ by putting $\tilde{\Phi}(x_n) = y_n$ and $\tilde{\Phi}(y_n) = x_n$ for all $n \in \mathbb{Z}^+$. Instead of one singly infinite cycle, we now have infinitely many 2-cycles. In particular the restriction of $\tilde{\Phi}$ to $\mathfrak{c}_\Phi(z)$ is now bijective, more precisely it consists of two mutually inverse bijections from $\{x_n \mid n \in \mathbb{Z}^+\}$ to $\{y_n \mid n \in \mathbb{Z}^+\}$.

Now suppose that $z \in Y$. We do the same thing with the roles of X and Y interchanged: write $z = y_1 \in Y$, $x_1 = \Phi(y_1) \in X$, $y_2 = \Phi(x_1) \in Y$, and so forth. Now the singly infinite cycle starting at z is

$$y_1, x_1, y_2, x_2, \ldots$$

As above, redefine Φ on this cycle to be $\tilde{\Phi}$, which maps $y_n \mapsto x_n$ and $x_n \mapsto y_n$.

The new map $\tilde{\Phi} : Z \to Z$ is obtained by adjusting Φ on each singly infinite cycle as above. The map $\tilde{\Phi}$ is now a bijection, since it induces a bijection from each Φ-equivalence class to itself. The map $\tilde{\Phi}$ still has the property that for all $x \in X$, we have $\tilde{\Phi}(x) \in Y$ and for all $y \in Y$ we have $\tilde{\Phi}(y) \in X$. Thus if we let $\tilde{f} : X \to Y$ by $\tilde{f}(x) \coloneqq \tilde{\Phi}(x)$ and $\tilde{g} : Y \to X$ by $\tilde{g}(y) \coloneqq \tilde{\Phi}(y)$, then $\tilde{f} : X \to Y$ and $\tilde{g} : Y \to X$ are injections. Since $\tilde{\Phi}$ is a bijection, each of $\tilde{f} : X \to Y$ and $\tilde{g} : Y \to X$ are bijections, so we get two bijections from X to Y, namely $\tilde{f}$ and $\tilde{g}^{-1}$. □

This proof of Theorem 9.6 is an elegant application of our cycle analysis of an injective map $f : X \to X$. The idea of putting together f and g into one injection is a very clever one, and we did not try to motivate it but rather just did it. It is probably worth covering up this proof and thinking anew about how one might prove the result. If $f : X \to Y$ is injective but not surjective, then there is $y \in Y \setminus f(X)$. We need to somehow change f so that y gets mapped to. The one thing that presents itself to us is that because we also have the function g, we can consider $x \coloneqq g(y)$. We can then redefine f by putting $\tilde{f}(x) \coloneqq y$, but this screws something else up: since f was injective, whatever $f(x) \in Y$ was previously, it is no longer mapped to by any element of X. So if we put $y_2 \coloneqq f(x)$, then we could adjust again by having $x_2 \coloneqq g(y_2)$ map to y_2. And now we continue. It is not at all clear upon first glance that this redefinition process eventually succeeds in making f into a bijection, but with some thought it can be made to work. The

above proof, which defines the injective function Φ and analyzes its cycles, is one way of doing so.

1.3. Non-Injective Dynamics. For a non-injective function $f : X \to X$, the study of its iterates $f^{\circ n}$ and the equivalence relation $\sim_f$ is, to say the least, much more challenging.

EXAMPLE 9.7. *Let $C : \mathbb{Z}^+ \to \mathbb{Z}^+$ be the **Collatz function**, defined as follows:*
- *If $n = 2k-1$ is odd, then $C(n) = 3n+1 = 6k-2$.*
- *If $n = 2k$ is even, then $C(n) = \frac{n}{2} = k$.*

The number $n = 1$ is 3-periodic: under C we have $1 \mapsto 4 \mapsto 2 \mapsto 1$.
We consider some forward orbits of points under C: if we reach any of 1, 2 or 4 we stop, because then we enter the above 3-cycle. We also stop if we reach any previously reached value of n.

$$3 \mapsto 10 \mapsto 5 \mapsto 16 \mapsto 8 \mapsto 4.$$

$$6 \mapsto 3.$$

$$7 \mapsto 22 \mapsto 11 \mapsto 34 \mapsto 17 \mapsto 52 \mapsto 26 \mapsto 13 \mapsto 40 \mapsto 20 \mapsto 10.$$

$$9 \mapsto 28 \mapsto 14 \mapsto 7.$$

$$12 \mapsto 6.$$

In fact we may as well consider only odd n because any even n will get its factors of 2 removed one by one and then have an odd value of n in its forward orbit. And then in place of writing down any even number we may as write down the corresponding odd number obtained by dividing by the largest 2^k such that $2^k \mid n$. So e.g.

$$15 \mapsto 23 \mapsto 35 \mapsto 53 \mapsto 5 \mapsto 1.$$

*Try it yourself. You will eventually come to believe that there is only one $\sim_C$-equivalence class, or in other words, for all $n \in \mathbb{Z}^+$ there is some $N \in \mathbb{Z}^+$ such that $f^{\circ N}(n) = 1$. Whether this is true is unknown. It is called the **Collatz Conjecture** after Lothar Collatz, who first conjectured that there is only one $\sim_C$-equivalence class in 1937.*

Let X be nonempty, and let $f : X \to X$ be a function. Of course we have $f(X) \subseteq X$. Applying f to this relation, we get

$$f^{\circ 2}(X) = f(f(X)) \subseteq f(X),$$

and similarly for all $n \in \mathbb{N}$ we get

$$X = f^{\circ 0}(X) \supseteq f(X) \supseteq f^{\circ 2}(X) \supseteq f^{\circ 3}(X) \supseteq \ldots \supseteq f^{\circ n}(X) \supseteq \ldots.$$

We define the **essential image**

$$f^\infty(X) \coloneqq \bigcap_{n=0}^{\infty} f^{\circ n}(X).$$

If f is surjective, then $f^{\circ n}(X) = X$ for all $n \in \mathbb{N}$, and thus $f^\infty(X) = X$. So the interesting case is when f is not surjective, in which case $f^\infty(X) \subsetneq X$. The following example shows that the essential image can be empty when X is infinite.

EXAMPLE 9.8. *Let $f : \mathbb{N} \to \mathbb{N}$ by $f(n) = n+1$. Then for all $n \in \mathbb{N}$ we have $f^{\circ n}(\mathbb{N}) = \mathbb{Z}^{\geq n}$, so*

$$f^\infty(\mathbb{N}) = \bigcap_{n=0}^{\infty} \mathbb{Z}^{\geq n} = \varnothing.$$

Now suppose X is finite and $f : X \to X$ is not surjective (hence, by Theorem 8.58, also not injective). For $x \in X$ we have $f^{\circ n}(x) \in f^{\circ n}(X)$, so certainly $f^{\circ n}(X)$ is nonempty for all $n \in \mathbb{N}$. Thus we have a nested sequence of nonempty subsets of the finite set X. Any such sequence must stabilize: there is $N \in \mathbb{Z}^+$ such that $f^{\circ m}(X) = f^{\circ n}(X)$ for all $m, n \geq N$, and then $f^\infty(X) = f^{\circ N}(X) \neq \varnothing$. Thus when X is finite, the essential image of f is the image of $f^{\circ n}$ for all sufficiently large n.

THEOREM 9.9. *Let X be finite nonempty, and let $f : X \to X$ be a function. The essential image $f^\infty(X)$ is the set of periodic points for f.*

PROOF. A point x is periodic for f if and only if its forward orbit $\overrightarrow{\mathcal{O}}_f(x)$ is a finite cycle. In this case we have

$$f(\overrightarrow{\mathcal{O}}_f(x)) = \overrightarrow{\mathcal{O}}_f(x)$$

and thus also for all $n \in \mathbb{Z}^+$ we have

$$f^{\circ n}(\overrightarrow{\mathcal{O}}_f(x)) = \overrightarrow{\mathcal{O}}_f(x),$$

so $\overrightarrow{\mathcal{O}}_f(x)$ lies in the essential image $f^\infty(X)$.

Conversely, suppose that x lies in the essential image $f^\infty(X)$. Let $\#X = n$. Then there is y in x such that $f^{\circ(n+1)}(y) = x$. Since $f^{\circ 0}(y), f^{\circ 1}(y), \ldots, f^{\circ n}(y)$ consists of more than $\#X$ elements, by the Pigeonhole Principle there must be $0 \leq i < j \leq n$ such that $f^{\circ i}(y) = f^{\circ j}(y) = z$, say. This means that the point z is periodic under f, and hence so is every point in its forward orbit, including x. □

In Exercise 9.1, you are asked to show that under the hypotheses of Theorem 9.9, if the essential image $f^\infty(X)$ consists of a single point x, then X has a unique $\sim_f$-equivalence class and x is a fixed point for f, i.e., $f(x) = x$.

2. Congruences

2.1. The Ring $\mathbb{Z}/N\mathbb{Z}$. Let $N \in \mathbb{Z}^+$. Let $\mathbb{Z}/N\mathbb{Z}$ denote the set of equivalence classes under the relation of congruence modulo N. For $a \in \mathbb{Z}$, we write $a \pmod N$ for the equivalence class $\mathfrak{c}(a)$ of a: it consists of all $b \in \mathbb{Z}$ such that $b = a + cN$ for some $c \in \mathbb{Z}$. As we saw in the discussion following Proposition 8.47, we have a map

$$\mathfrak{c} : \mathbb{Z} \to \mathbb{Z}/N\mathbb{Z},\ a \mapsto a \pmod N.$$

The big idea here is to use the binary operations $+$ and $\cdot$ on $\mathbb{Z}$ and the map $\mathfrak{c}$ to define binary operations $+$ and $\cdot$ on $\mathbb{Z}/N\mathbb{Z}$ and then to show that the good properties of the binary operations on $\mathbb{Z}$ propagate under $\mathfrak{c}$ to good properties of the binary operations on $\mathbb{Z}/N\mathbb{Z}$.

To start with, let $a \pmod N, b \pmod N$ be two elements of $\mathbb{Z}/N\mathbb{Z}$. We **would like to** define

$$(a \pmod N) + (b \pmod N) \coloneqq a + b \pmod N.$$

However, there is something to check here. The catchphrase here is that we need to check that this operation is **well-defined**: what does that mean? The point is that $a \pmod N$ is *associated* to the integer a but not the same as a: e.g. we have

$$1 \pmod 3 = 7 \pmod 3 = -11 \pmod 3$$

and

$$2 \pmod 3 = 14 \pmod 3 = -28 \pmod 3.$$

Thus we are trying to define an addition operation on equvialence classes of integers by *selecting* an integer in each equivalence class and adding those selected integers. So what needs to be checked is that the answer does not depend upon the integers that we selected. So for instance we want to put

$$(1 \pmod 3) + (2 \pmod 3) = 3 \pmod 3$$

but also $1 \pmod 3 = 7 \pmod 3$ and $2 \pmod 3 = 14 \pmod 3$, so our definition would also give

$$(3 \pmod 3) = (1 \pmod 3) + (2 \pmod 3) = (7 \pmod 3) + (14 \pmod 3)$$
$$= 7 + 14 \mod 3 = 21 \pmod 3,$$

so for the definition to make sense we need $3 \pmod 3 = 21 \mod 3$. Happily this is indeed the case, since $3 \mid (21 - 3)$. That was just an example: we must check that, in general, no matter which representatives a and b we choose for our two congruence classes, the congruence class $a + b \pmod N$ of the sum is the same.

This is a case in which understanding the difficulty is most of the battle: in fact it is pretty easy to check this. Namely, suppose $a_1 \equiv a_2 \pmod N$ and $b_1 \equiv b_2 \pmod N$. Then $N \mid (a_1 - a_2)$ and $N \mid (b_1 - b_2)$, so

$$N \mid (a_1 - a_2) + (b_1 - b_2) = (a_1 + b_1) - (a_2 + b_2),$$

and thus $a_1 + b_1 \pmod N = a_2 + b_2 \pmod N$.

We have an entirely parallel discussion for multiplication: we wish to define

$$(a \pmod N) \cdot (b \pmod N) \coloneqq (a \cdot b) \pmod N,$$

but we need to check that this is well-defined. So, if $a_1 \equiv a_2 \pmod N$ and $b_1 \equiv b_2 \pmod N$, then $N \mid (a_1 - a_2)$ and $N \mid (b_1 - b_2)$, so

$$N \mid (a_1 - a_2)b_1 + (b_1 - b_2)a_2 = a_1b_1 - a_2b_2 + (a_2b_1 - a_2b_1) = a_1b_1 - a_2b_2,$$

so

$$a_1b_1 \equiv a_2b_2 \pmod N.$$

PROPOSITION 9.10. *Let $N \in \mathbb{Z}^+$. The operations of $+$ and $\cdot$ defined above make $\mathbb{Z}/N\mathbb{Z}$ into a commutative ring.*

PROOF. We have a fairly lengthy list of properties to check, but all the verifications go in the same way. We will do some and leave the rest to the reader.

• Commutativity of $+$: Let $X, Y \in \mathbb{Z}/N\mathbb{Z}$. Then $X = x \pmod N$ and $Y = y \pmod N$ for some $x, y \in \mathbb{Z}$. Since addition in $\mathbb{Z}$ is commutative, we have

$$X + Y = (x \pmod N) + (y \pmod N) = x + y \pmod N$$
$$= y + x \pmod N = (y \pmod N) + (x \pmod N) = Y + X.$$

• Associativity of $+$: Let $X, Y, Z \in \mathbb{Z}/N\mathbb{Z}$. Then there are $x, y, z \in \mathbb{Z}$ such that

$$X = x \pmod N,\ Y = y \pmod N,\ Z = z \pmod N.$$

Since addition in $\mathbb{Z}$ is associative, we have

$$(X + Y) + Z = ((x \pmod N) + (y \pmod N)) + (z \pmod N)$$
$$= (x + y \pmod N) + (z \pmod N) = ((x + y) + z \pmod N)$$
$$= (x + (y + z) \pmod N) = (x \pmod N) + (y + z) \pmod N)$$
$$= (x \pmod N) + ((y \pmod N) + (z \pmod N)) = X + (Y + Z).$$

• Identity element for addition: We claim that, since 0 is the additive identity in $\mathbb{Z}$, the class $0 + \pmod N$ is the additive identity in $\mathbb{Z}/N\mathbb{Z}$. Indeed, for all $X = x \pmod N$, we have

$$(0 \pmod N) + X = (0 \pmod N) + (x \pmod N)$$

$$= (0 + x) \pmod N = x \pmod N = X.$$

• Inverses for addition: Let $X = x \pmod N \in \mathbb{Z}/N\mathbb{Z}$. We claim that $Y = -x \pmod N$ is the additive inverse of X. Indeed we have

$$X + Y = (x \pmod N) + (-x \pmod N) = (x - x) \pmod N = 0 \pmod N.$$

In Exercise 9.2 you are asked to check the commutativity of $\cdot$, the associativity of $\cdot$, that $1 \pmod N$ is a multiplicative identity, and the distributive property. □

EXAMPLE 9.11. *We claim that if n is an odd integer, then $n^2 \equiv 1 \pmod 8$. Indeed, if n is odd then $n = 2k + 1$, so*

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

We observe that for any $k \in \mathbb{Z}$, exactly one of k and $k+1$ is even, so $k(k+1) = 2A$ for some $A \in \mathbb{Z}$, and thus

$$n^2 = 4(2A) + 1 = 8A + 1.$$

This shows that $n^2 \equiv 1 \pmod 8$.

Let us try to think of this fully in terms of the ring $\mathbb{Z}/8\mathbb{Z}$ rather than a statement about integers and remainders. First of all, $n^2 \pmod 8$ depends only on $n \pmod 8$. (This is a special case of the well-definedness of multiplication modulo n.) Moreover, we claim that under the map $\mathfrak{c} : \mathbb{Z} \to \mathbb{Z}/8\mathbb{Z}$, the image of the odd integers is

$$\{1 \pmod 8, 3 \pmod 8, 5 \pmod 8, 7 \pmod 8\}.$$

that all of these classes lie in the image is clear, and every other element of $\mathbb{Z}/8\mathbb{Z}$ is of the form $n \pmod 8$ for some even n. But again we have to remember that the elements of $\mathbb{Z}/N\mathbb{Z}$ are not equal to integers but only represented by them, so we must check that we cannot *have an even integer n and and an odd integer m such that $m \pmod 8 = n \pmod 8$. If so, then $8 \mod m - n$, so $2 \mod m - n$, so m and n have the same parity, a contradiction.*

Having established all this, we just need to check that

$$(1 \pmod 8)^2 = 1 \pmod 8,$$

$$(3 \pmod 8)^2 = 9 \pmod 8 = 1 \pmod 8,$$

$$(5 \pmod 8)^2 = 25 \pmod 8 = 1 \pmod 8,$$

$$(7 \pmod 8)^2 = 49 \pmod 8 = 1 \pmod 8.$$

As a statement about the ring $\mathbb{Z}/8\mathbb{Z}$, this is not hard to prove but still somewhat surprising: it says that in this ring the element $1 \pmod 8$ has four square roots.

In the above example we saw that it makes sense to talk about the parity of an element of $\mathbb{Z}/8\mathbb{Z}$: more precisely, any two elements of the same congruence class modulo 8 have the same parity. This means that the map

$$\mathbb{Z}/8\mathbb{Z} \to \mathbb{Z}/2\mathbb{Z},\ n \pmod 8 \mapsto n \pmod 2$$

is well-defined. More generally, for $N \in \mathbb{Z}^+$ we can *try* to define a parity map on congruence classes modulo N:

$$\mathbb{Z}/N\mathbb{Z} \to \mathbb{Z}/2\mathbb{Z},\ n \pmod N \mapsto n \pmod 2.$$

Is this well-defined?

The answer is that it is if and only if N is even. Indeed, if N is even and $a \pmod N \equiv b \pmod N$, then $N \mid a-b$, and since $2 \mid N$ we have also $2 \mid a-b$, so $a \pmod 2 = b \pmod 2$. The same argument proves something more general.

PROPOSITION 9.12. *Let $M, N \in \mathbb{Z}^+$ with $M \mid N$.*

a) *We have a well-defined map $q : \mathbb{Z}/N\mathbb{Z} \to \mathbb{Z}/M\mathbb{Z}$ given by*

$$q : a \pmod N \mapsto a \pmod M.$$

b) *The map q is surjective.*

c) *For every $Y = y + \pmod M \in \mathbb{Z}/M\mathbb{Z}$, the fiber $q^{-1}(Y)$ consists of all $x \pmod N$ with $x \equiv y \pmod M$. The fiber $q^{-1}(Y)$ consists of precisely $\frac{M}{N}$ elements.*

PROOF. a) Indeed, if $M \mid N$ and $a \pmod N \equiv b \pmod N$, then $N \mid a-b$, so also $M \mid a-b$ and thus $a \pmod M = b \pmod M$.
b) Every element of $\mathbb{Z}/M\mathbb{Z}$ is of the form $y \pmod M$ for some $y \in \mathbb{Z}$. We have $q(y \pmod N) = y \pmod M$.
c) For $X = x \pmod N$, we have $q(X) = Y$ if and only if $x \pmod M = y \pmod M$ if and only if $M \mid x - y$ if and only if

$$x = y + aM$$

for some $a \in \mathbb{Z}$. The question is how many distinct classes modulo N this yields. Writing $a = qN/M + r$ with $q \in \mathbb{Z}$ and $0 \le r < M/N$, we get

$$x = y + (qN/M + r)M = y + qN + rM \equiv y + rM \pmod N,$$

we see that the class modulo N depends only on the the remainder of a modulo N/M. Conversely, any two of the numbers the numbers $y, y+M, \ldots y+(N/M-1)M$ differ from each other by less than N so yield distinct congruence classes modulo N. It follows that the fiber $q^{-1}(Y)$ consists of the $\frac{M}{N}$ distinct elements

$$y \pmod N, y + M \pmod N, \ldots, y + N - M \pmod N. \qquad \square$$

On the other hand our "map" $a \pmod N \to a \pmod 2$ is **not** well-defined when N is odd (and thus it is not actually a map at all: the whole thing just doesn't work). As a concrete example, it doesn't make sense to say that the class $1 \pmod 3$ is odd, because $1 \pmod 3 = 4 \pmod 3$. More generally, if N is odd, then $a + N \pmod N = a \pmod N$ but $a + N \pmod 2 = a + 1 \pmod 2 \neq a \pmod 2$.

EXAMPLE 9.13. *A **Pythagorean triple** is a triple (x, y, z) of positive integers such that $x^2 + y^2 = z^2$. You surely know some Pythagorean triples, e.g. $(3, 4, 5)$, $(5, 12, 13)$, $(7, 24, 25)$, $(8, 15, 17)$. In Exercise 9.3 you are asked to check that for any integers $0 < u < v$, we have that $(v^2 - u^2, 2uv, v^2 + u^2)$ is a Pythagorean triple, so there are infinitely many Pythagorean triples.*

Here will prove a congruential fact about Pythagorean triples: for any Pythagorean triple, we have $60 \mid xyz$. Because 60 is the product of the pairwise coprime integers 3, 4 and 5, it suffices to show that $3 \mid xyz$, $4 \mid xyz$ and $5 \mid xyz$: then

$\mathrm{lcm}(3,4,5) = 60 \mid xyz$.
Step 1: More precisely, we will show that if $x^2 + y^2 = z^2$ *then at least one of* x *and* y *is divisible by* 4.

If $x^2 + y^2 = z^2$, *then* x *and* y *cannot both be odd, for then*

$$z^2 = x^2 + y^2 \equiv 1 + 1 = 2 \pmod 4,$$

whereas the squares modulo 4 *are* $0 \pmod 4$ *and* $1 \pmod 4$. *Because of the symmetry in* $x^2 + y^2 = z^2$ *we may assume that* x *is even.*
Case 1: Suppose that y *is odd. If* $4 \nmid x$, *then* $x^2 \equiv 4 \pmod 8$ *(indeed* $(2 \pmod 8)^2 = (6 \pmod 8)^2 = 4 \pmod 8$*), so* $z^2 \pmod 8 = x^2 + y^2 \pmod 8 = 4 + 1 \pmod 8 = 5 \pmod 8$, *a contradiction, since* 1 *is the only odd square in* $\mathbb{Z}/8\mathbb{Z}$.
Case 2: Suppose that y *is also even. Here we will need to work in* $\mathbb{Z}/16\mathbb{Z}$, *in which case a little calculation shows that the even squares are* $0 \pmod{16}$ *and* $4 \pmod{16}$. *So if both* x *and* y *are even and neither is divisible by* 4 *then* $x^2 \pmod{16} = y^2 \pmod{16} = 4$, *so* $z^2 = x^2 + y^2 \equiv 4 + 4 \pmod{16} = 8 \pmod{16}$, *a contradiction.*
Step 2: Similarly, we claim that if $x^2 + y^2 = z^2$, *then at least one of* x *and* y *is divisible by* 3.

The squares in $\mathbb{Z}/3\mathbb{Z}$ *are*

$$0 \pmod 3 = (0 \pmod 3)^2,$$

$$1 \pmod 3 = (1 \pmod 3)^2 = (2 \pmod 3)^2.$$

So if neither x *nor* y *is divisible by* 3, *then we would have*

$$z^2 = x^2 + y^2 \equiv 1 + 1 \pmod 3 = 2 \pmod 3,$$

a contradiction.
Step 3: We will suppose that $x^2 + y^2 = z^2$, *that neither* x *nor* y *is divisible by* 5, *and show that* z *must be divisible by* 5. *The squares in* $\mathbb{Z}/5\mathbb{Z}$ *are*

$$0 \pmod 5 = (0 \pmod 5)^2,$$

$$1 \pmod 5 = (1 \pmod 5)^2 = (4 \pmod 5)^2,$$

$$4 \pmod 5 = (2 \pmod 5)^2 = (3 \pmod 5)^2.$$

So the possibilities for $x^2 + y^2$ *modulo* 5 *are* $1 + 1 \pmod 5 = 2 \pmod 5$, $4 + 4 \pmod 5 = 3 \pmod 5$ *and* $1 + 4 \pmod 5 = 0 \pmod 5$. *Of the three, only the last is a square modulo* 5, *so it must be that* $z^2 \equiv 0 \pmod 5$, *which as we have seen, means that* $z \equiv 0 \pmod 5$ *and thus* $5 \mid z$.

PROPOSITION 9.14. *Let* $x \in \mathbb{Z}$, *and let* $N \in \mathbb{Z}^+$. *The following are equivalent:*

(i) *The element* $X \coloneqq x \pmod N \in \mathbb{Z}/N\mathbb{Z}$ *is a unit in the ring* $\mathbb{Z}/N\mathbb{Z}$: *i.e., it has a multiplicative inverse.*
(ii) *We have* $\gcd(x, N) = 1$.

PROOF. (i) $\implies$ (ii) Let $Y = y \pmod N$ be such that $XY = 1 \pmod N$: that is, $xy \equiv 1 \pmod N$, so $N \mid xy - 1$. Let $d = \gcd(X, N)$. Then $d \mid x$ and $d \mid N = xy - 1$, so $d \mid y(x) - (xy - 1) = 1$. Thus $d = 1$.
(ii) $\implies$ (i): By Theorem 5.13 there are integers $a, b \in \mathbb{Z}$ such that $ax + bN = 1$. It follows that if $A \coloneqq a \pmod N$, then

$$AX = (a \pmod N)(x \pmod N) = 1 - bN \pmod N = 1 \pmod N,$$

so $A = X^{-1}$. □

2.2. The Chinese Remainder Theorem.

PROPOSITION 9.15. *Let N_1 and N_2 be positive integers, and consider the map*

$$\Phi : \mathbb{Z}/N_1N_2\mathbb{Z} \to \mathbb{Z}/N_1\mathbb{Z}\times\mathbb{Z}/N_2\mathbb{Z},\ x \pmod{N_1N_2} \mapsto (x \pmod{N_1}, x \pmod{N_2}),$$

which is well-defined by Proposition 9.12. The following are equivalent:

(i) *The map Φ is injective.*
(ii) *The map Φ is surjective.*
(iii) *The map Φ is bijective.*
(iv) *We have $\gcd(N_1, N_2) = 1$: equivalently, $\operatorname{lcm}(N_1, N_2) = N_1N_2$.*

PROOF. Since $\mathbb{Z}/N_1N_2\mathbb{Z}$ and $\mathbb{Z}/N_1\mathbb{Z}\times\mathbb{Z}/N_2\mathbb{Z}$ are both finite sets of size N_1N_2, the equivalence of (i), (ii) and (iii) follows from Theorem 8.58. To complete the proof if suffices to show (i) $\iff$ (iv).
(i) $\implies$ (iv): We argue by contraposition: suppose that $\gcd(N_1, N_2) > 1$. Put $d \coloneqq \operatorname{lcm}(N_1, N_2)$. Then $1 \leq d < N_1N_2$, so $0 \pmod{N_1N_2} \neq d \pmod{N_1N_2}$. But since $N_1 \mid d$ and $N_2 \mid d$ we have $d \pmod{N_1} = 0 \pmod{N_1}$ and $d \pmod{N_2} = 0 \pmod{N_2}$. Therefore

$$\Phi(d \pmod{N_1N_2}) = \Phi(0 \pmod{N_1N_2}),$$

so Φ is not injective.
(iv) $\implies$ (i): Let $x, y \in \mathbb{Z}$ and suppose that $\Phi(x \pmod{N_1N_2}) = \Phi(y \pmod{N_1N_2})$. This means:

$$x \pmod{N_1} = y \pmod{N_1},\ x \pmod{N_2} = y \pmod{N_2}.$$

Thus $N_1 \mid (x - y)$ and $N_2 \mid (x - y)$, we have $N_1N_2 = \operatorname{lcm}(N_1N_2) \mid (x - y)$, so $x \pmod{N_1N_2} = y \pmod{N_1N_2}$. □

THEOREM 9.16 (Chinese Remainder Theorem). *Let N_1 and N_2 be coprime positive integers.*

a) *For all $a, b \in \mathbb{Z}$ there is a $c \in \mathbb{Z}$ such that $c \equiv a \pmod{N_1}$ and $c \equiv b \pmod{N_2}$.*
b) *The set of all integers c such that $c \equiv a \pmod{N_1}$ and $c \equiv b \pmod{N_2}$ consists of one full congruence class modulo N_1N_2.*

PROOF. Since N_1 and N_2 are coprime, the map Φ of Proposition 9.15 is surjective, so there is $C \in \mathbb{Z}/N_1N_2\mathbb{Z}$ such that $\Phi(C) = (a \pmod{N_1}, b \mod N_2)$. Like any element of $\mathbb{Z}/N\mathbb{Z}$, the class C is of the form $c \pmod{N_1N_2}$ for some $c \in \mathbb{Z}$. This provides an integer c such that $c \equiv a \pmod{N_1}$ and $c \equiv b \pmod{N_2}$ and also shows that any integer that is congruent to c modulo N_1N_2 works just as well. The injectivity of Φ means that if $d \in \mathbb{Z}$ is such that $d \pmod{N_1N_2} \neq c \pmod{N_1N_2}$ then either $d \not\equiv a \pmod{N_1}$ or $d \not\equiv b \pmod{N_2}$. □

Our proof of Theorem 9.16 is somewhat sneaky: we deduce the surjectivity of the map Φ from its injectivity, which is easier to show directly. One should still ask: if you are given actual coprime N_1, N_2 and integers a and b, how do you *actually* go about finding an integer c that is congruent to a modulo N_1 and to b modulo N_2?

Here is a procedure for this: start by taking $c \coloneqq a$. Thus certainly $c \equiv a \pmod{N_1}$, but probably $c \not\equiv b \pmod{N_2}$. If we adjust c by any multiple of N_1, then we will retain the first congruence. So let $c_2 \coloneqq c + N_1$, $c_3 \coloneqq c_2 + N_1 = c + 2N_1$, and so forth. I claim that eventually we will hit c_n such that $c_n \equiv b \pmod{N_2}$.

For instance, suppose that we want an integer that is congruent to 1 modulo 3 and also to 2 modulo 5. Start with $c = c_1 = 1$. Then $c_1 \not\equiv 2 \pmod 5$. so we take $c_2 = c_1 + 3 = 4$. Again $c_2 \not\equiv 2 \pmod 5$, so we take $c_3 = c_2 + 3 = 7$. Now $c_3 \equiv 2 \pmod 5$, so our answer is $c = 7$, which is uniquely determined modulo 15.

To justify the claim in general: because $\gcd(N_1, N_2) = 1$, the class $A := N_1 \pmod{N_2}$ has a multiplicative inverse, say $B = b \pmod{N_2}$. Thus $bN_1 \equiv 1 \pmod{N_2}$. The functions

$$\bullet A : \mathbb{Z}/N_2\mathbb{Z} \to \mathbb{Z}/N_2\mathbb{Z},\ x \pmod{N_2} \mapsto N_1 x \pmod{N_2}$$

and

$$\bullet B : \mathbb{Z}/N_2\mathbb{Z} \to \mathbb{Z}/N_2\mathbb{Z},\ x \pmod{N_2} \mapsto bx \pmod{N_2}$$

are mutually inverse:

$$x \pmod{N_2} \mapsto N_1 x \pmod{N_2} \mapsto bN_1 x \pmod{N_2} = 1 \cdot x \pmod{N_2} = x \pmod{N_2}$$

$$x \pmod{N_2} \mapsto bx \pmod{N_2} \mapsto N_1 bx \pmod{N_2} = 1 \cdot x \pmod{N_2} = x \pmod{N_2},$$

hence $\bullet A$ is bijective. This means that every congruence class modulo N_2 is of the form $xN_1 \pmod{N_2}$ for some integer x (which we can take to be positive if we like by adding a large enough multiple of N_2). In particular there is $n \in \mathbb{Z}^+$ such that $nN_1 \equiv b - a \pmod{N_2}$, so $a + nN_1 \equiv b \pmod{N_2}$.

A generalization of Proposition 9.15 to the case where N_1 and N_2 need not be coprime is given in Exercise 9.5. More useful than this is Exercise 9.6 (which can be proved as a consequence of Exercise 9.5 or also directly), which says that two congruence classes $a \mod N_1$ and $b \pmod{N_2}$ are "compatible" in the sense that there is some $c \in \mathbb{Z}$ such that $c \equiv a \pmod{N_1}$ and $c \equiv b \pmod{N_2}$ if and only if $a \pmod{\gcd(N_1 N_2)} = b \pmod{\gcd(N_1 N_2)}$.

The following result extends the Chinese Remainder Theorem to pairwise coprime moduli $N_1, \ldots, N_k$.

THEOREM 9.17 (r-Fold Chinese Remainder Theorem). *Let $r \in \mathbb{Z}^+$, and let $N_1, \ldots, N_r$ be pairwise coprime positive integers.*

a) *The map $\Phi : \mathbb{Z}/N_1 \cdots N_r\mathbb{Z} \to \prod_{i=1}^r \mathbb{Z}/N_i\mathbb{Z}$ given by*
$$x \pmod{N_1 \cdots N_r} \mapsto (x \pmod{N_1}, \ldots, x \pmod{N_r})$$
is a bijection.

b) *For all $a_1, \ldots, a_r \in \mathbb{Z}$, there is $c \in \mathbb{Z}$ such that $c \equiv a_i \pmod{N_i}$ for all $1 \le i \le r$.*

c) *The set of all integers c such that $c \equiv a_i \pmod{N_i}$ for all $1 \le i \le r$ consists of one full congruence class modulo $N_1 \cdots N_r$.*

You are asked to prove Theorem 9.17 in Exercise 9.7.

Let $N \ge 2$, and let $N = p_1^{a_1} \cdots p_r^{a_r}$ be its standard form prime factorization: thus $p_1 < \ldots < p_r$ are primes and $a_1, \ldots, a_r$ are positive integers. Then we can apply Theorem 9.17 with $N_1 = p_1^{a_1}, \ldots, N_r = p_r^{a_r}$. In this way most considerations about congruences modulo N (and even about the ring $\mathbb{Z}/N\mathbb{Z}$) get reduced to considerations about congruences modulo prime powers p^a (and even to the rings $\mathbb{Z}/p^a\mathbb{Z}$).

Here is one example: for $N \in \mathbb{Z}^+$, let $\varphi(N)$ denote the number of units in the ring $\mathbb{Z}/N\mathbb{Z}$: i.e., the number of elements X that have a multiplicative inverse Y (i.e., so that $XY = 1$). By Proposition 9.14, a class $X = x \pmod N$ is a unit if and only if $\gcd(x, N) = 1$, so we also have that

$$\varphi(N) = \#\{1 \leq x \leq N \mid \gcd(x, N) = 1\}.$$

The function φ was first considered by Euler.[1]

LEMMA 9.18. *Let $N_1, N_2 \in \mathbb{Z}^+$. For $x \in \mathbb{Z}$, the following are equivalent:*

(i) *The class $X = x \pmod{N_1N_2}$ is a unit in $\mathbb{Z}/N_1N_2\mathbb{Z}$.*
(ii) *For $i = 1, 2$, the class $X_i = x \pmod{N_i}$ is a unit in $\mathbb{Z}/N_i\mathbb{Z}$.*

PROOF. For $n \in \mathbb{Z}$, we have $\gcd(n, N_1N_2) = 1$ if and only if $\gcd(n, N_1) = 1$ and $\gcd(n, N_2) = 1$. The result follows from this and Proposition 9.14. □

Suppose now that N_1 and N_2 are coprime positive integers. Then Lemma 9.18 shows that for all $x \in \mathbb{Z}$, we have that $X = x \pmod{N_1N_2}$ is a unit in $\mathbb{Z}/N_1N_2\mathbb{Z}$ if and only if both components of $\Phi(X)$ are units. In other words, if we denote by $U(N)$ the set of units in $\mathbb{Z}/N\mathbb{Z}$, then Φ induces a bijection

$$\Phi : U(N_1N_2) \to U(N_1) \times U(N_2),$$

and we conclude:

PROPOSITION 9.19. *If N_1 and N_2 are coprime positive integers then*

$$\varphi(N_1N_2) = \varphi(N_1)\varphi(N_2). \tag{52}$$

THEOREM 9.20. *Let $N = p_1^{a_1} \cdots p_r^{a_r}$ be the standard form factorization of the positive integer N. Then we have*

$$\varphi(p_1^{a_1} \cdots p_n^{a_n}) = \prod_{i=1}^{r} p_i^{a_i - 1}(p_i - 1). \tag{53}$$

PROOF. Step 1: In Exercise 9.8 you are asked to extend Proposition 9.19 to show that if $N_1, \ldots, N_r$ are pairwise coprime, then $\varphi(N_1 \cdots N_r) = \prod_{i=1}^r \varphi(N_i)$. (This is yet another instance of the kind of easy induction proof discussed in §7.8.) So we have

$$\varphi(p_1^{a_1} \cdots p_r^{a_r}) = \prod_{i=1}^{r} \varphi(p_i^{a_i}).$$

Thus it remains to compute $\varphi(p^a)$ for any prime power p^a, which we do in Step 2. Step 2: Let $N = p^a$ be a prime power. To compute $\varphi(p^a)$ it suffices to count the elements of $\{1 \leq n \leq p^a \mid \gcd(n, p^a) = 1\}$. But $\gcd(n, p^a) = 1$ if and only if $\gcd(n, p) = 1$: so we need to count the number of integers between 1 and p^a that are *not* multiples of p. Well, this is the total number of integers from 1 to p^a – which is of course p^a – minus the number of integers that are multiples of p. The multiples of p in this interval are $p, 2p, \ldots, p^a = p \cdot 1, p \cdot 2, \ldots, p \cdot p^{a-1}$, so there are p^{a-1} of them. Conclusion:

$$\varphi(p^a) = p^a - p^{a-1} = p^{a-1}(p - 1).$$ □

[1]Its official name is the **totient**, but so far as I have seen, in 21st century life it is much more common to call it the "Euler phi function."

2.3. Wilson's Theorem and Fermat's Two Squares Theorem.

THEOREM 9.21. *For all prime numbers p, the ring $\mathbb{Z}/p\mathbb{Z}$ is a field.*

PROOF. FIRST PROOF: To show that a commutative ring R is a field, we must show that every nonzero element x of R has a multiplicative inverse: i.e., there is $y \in R$ such that $xy = 1$. So let $X = x \pmod p$ be a nonzero element of $\mathbb{Z}/p\mathbb{Z}$. This means that $x \pmod p \neq 0 \pmod p$, i.e., $p \nmid x$. Since p is prime, this means that $\gcd(p, x) = 1$, so X has a multiplicative inverse by Proposition 9.14.
SECOND PROOF: Since $\mathbb{Z}/p\mathbb{Z}$ is a finite commutative ring, by Theorem 8.70 it suffices to show that it is an integral domain. But I claim that this is a reformulation of Euclid's Lemma. Indeed, let $X = x \pmod p$, $Y = y \pmod p$ be nonzero elements of $\mathbb{Z}/p\mathbb{Z}$. Then $p \nmid x$ and $p \nmid y$, so – by Euclid's Lemma in contrapositive form – $p \nmid xy$. It follows that $XY = xy \pmod p \neq 0$. □

Conversely, if $N > 1$ is not prime, then the ring $\mathbb{Z}/N\mathbb{Z}$ is *not* a field. Indeed, because N is not prime there are integers $1 < a, b < N$ such that $ab = N$, and then $A = a \pmod N$ and $B = b \pmod N$ are such that

$$AB = ab \pmod N = N \pmod N = 0 \pmod N$$

but $A, B \neq 0 \pmod N$ because $N \nmid a$ and $N \nmid b$.

COROLLARY 9.22. *Let p be a prime number.*

a) *There is exactly one $X \in \mathbb{Z}/2\mathbb{Z}$ such that $X^2 = 1$, namely $X = 1 \pmod 2$.*
b) *If $p > 2$, there are exactly 2 elements $X \in \mathbb{Z}/p\mathbb{Z}$ such that $X^2 = 1$, namely $X_1 = 1 \pmod p$ and $X_2 = -1 \pmod p$.*

PROOF. a) We have $(0 \pmod 2)^2 = 0 \pmod 2$ and $(1 \pmod 2)^2 = 1 \pmod 2$.
b) First we observe that since $p > 2$, the classes $1 \pmod p$ and $-1 \pmod p$ are distinct elements of $\mathbb{Z}/p\mathbb{Z}$: if not, we would have $p \mid (1 - (-1) = 2$. Now let F be any field, and let $X \in F$ be such that $X^2 = 1$. Then $0 = X^2 - 1 = (X + 1)(X - 1)$. In a field F the product of two elements can only be zero if one of them is zero, so we get $X + 1 = 0$ or $X - 1 = 0$, i.e., $X = 1$ or $X = -1$. □

This corollary shows for instance that the fact that the ring $\mathbb{Z}/8\mathbb{Z}$ has four square roots of 1 prevents it from being a field. However e.g. the ring $\mathbb{Z}/4\mathbb{Z}$ is not a field but has only two square roots of 1.

You might enjoy investigating how many square roots of 1 there are in the ring $\mathbb{Z}/N\mathbb{Z}$. The answer depends on the prime factorization of N.

THEOREM 9.23 (Fermat's Little Theorem). *Let p be a prime number. For all $x \in \mathbb{Z}$, we have*

$$x^p \equiv x \pmod p.$$

PROOF. Case 1: If $x \equiv 0 \pmod p$, then $x^p \equiv 0^p \equiv 0 \equiv x \pmod p$.
Case 2: Next suppose that $x \not\equiv 0 \pmod p$. Since $\mathbb{Z}/p\mathbb{Z}$ is a field and

$$X \coloneqq x \pmod p$$

is a nonzero element of $\mathbb{Z}/p\mathbb{Z}$, by Proposition 8.69d) the map $X\bullet : \mathbb{Z}/p\mathbb{Z} \to \mathbb{Z}/p\mathbb{Z}$ given by $Y \mapsto XY$ is a bijection. Certainly

$$(\bullet X)(0 \pmod p) = x \cdot 0 \pmod p = 0 \pmod p,$$

so also

$$\bullet X : \mathbb{Z}/p\mathbb{Z} \setminus \{0 \pmod p\} \to \mathbb{Z}/p\mathbb{Z} \setminus \{0 \pmod p\}$$

is a bijection (cf. Exercise 8.33). For $1 \leq i \leq p-1$, let $Y_i \coloneqq i \pmod p$, so

$$\mathbb{Z}/p\mathbb{Z} \setminus \{0 \pmod p\} = \{Y_0, \ldots, Y_{p-1}\}.$$

It follows that the finite lists

$$\ell_1 : Y_0, \ldots, Y_{p-1}$$

and

$$\ell_2 : XY_0, \ldots, XY_{p-1}$$

are both irredundant and consist of all elements of $\mathbb{Z}/p\mathbb{Z} \setminus \{0 \pmod p\}$, so the second list is obtained from the first list by reordering the elements. Therefore the product of all the elements of ℓ_1 is equal to the product of all the elements of ℓ_2:

$$Y_0 \cdots Y_{p-1} = (XY_0) \cdots (XY_{p-1}) = X^{p-1}(Y_0 \cdots Y_{p-1}). \tag{54}$$

The element $Y_0 \cdots Y_{p-1}$ is a product of nonzero elements in a field, thus it has a multiplicative inverse, and multiplying (54) by this inverse, we get

$$1 \pmod p = X^{p-1} = (x \pmod p)^{p-1} = x^{p-1} \pmod p,$$

so $p \mid x^p - 1$. It follows that $p \mid x(x^p - 1) = x^p - x$, so $x^p \equiv x \pmod p$. □

Theorem 9.24 (Wilson's Theorem). *For a prime number p, we have*

$$(p-1)! \equiv -1 \pmod p. \tag{55}$$

Proof. We have $(2-1)! = 1 \equiv -1 \pmod 2$, so we may assume that $p \geq 3$.

Every nonzero element of the field $\mathbb{Z}/p\mathbb{Z}$ is of the form $X = x \pmod p$ for a unique integer X with $1 \leq X \leq p-1$, so if we write $(\mathbb{Z}/p\mathbb{Z})^\times$ for the set of nonzero elements of $\mathbb{Z}/p\mathbb{Z}$, then (55) is equivalent to the statement that the product of all the elements of $(\mathbb{Z}/p\mathbb{Z})^\times$ is $-1 \pmod p$. Henceforth, since we will be working throughout with elements of the field $\mathbb{Z}/p\mathbb{Z}$, we will just write -1 for $-1 \pmod p$, and for $x \in (\mathbb{Z}/p\mathbb{Z})^\times$, we will write x^{-1} for the inverse of x in $\mathbb{Z}/p\mathbb{Z}$.

By Theorem 9.21, for r $x \in (\mathbb{Z}/p\mathbb{Z})^\times$, we also have $x^{-1} \in \mathbb{Z}/p\mathbb{Z}$: that is, every element of $(\mathbb{Z}/p\mathbb{Z})^\times$ has a multiplicative inverse. The key idea of the proof is to evaluate the product $\prod_{x \in (\mathbb{Z}/p\mathbb{Z})^\times} x$ by pairing off a factor x with its inverse x^{-1} whenever possible. If $x \neq x^{-1}$, then both factors appear in the product, and since $x \cdot x^{-1} = 1$, they cancel each other out. Therefore the product $\prod_{x \in (\mathbb{Z}/p\mathbb{Z})^\times} x$ is equal to the product over $x \in (\mathbb{Z}/p\mathbb{Z})^\times$ such that $x = x^{-1}$. We have $x = x^{-1}$ if and only if $x^2 = 1$, which by Corollary 9.22 holds if and only if $x = \pm 1$, which are distinct elements of $\mathbb{Z}/p\mathbb{Z}$ since we've assumed $p \geq 3$. It follows that

$$(p-1)! \pmod p = \prod_{x \in (\mathbb{Z}/p\mathbb{Z})^\times} x = 1 \cdot -1 = -1.$$ □

The fields $\mathbb{Z}/p\mathbb{Z}$ can behave quite differently from the field $\mathbb{R}$ of real numbers. In particular $\mathbb{R}$ and all of its subfields are ordered fields, so as we saw in §4.2, for all $x \in \mathbb{R}$ we have $x^2 \geq 0$. The field $\mathbb{C}$ of complex numbers cannot be ordered because $i^2 = -1$ and by Proposition 4.4a) in any ordered field we have $-1 < 0$.

In any ordered field F it is also true that if $x_1, \ldots, x_n$ are elements of F that are not all zero, then $x_1^2 + \ldots + x_n^2 > 0$. (Indeed, each term is non-negative and

at least one term is strictly positive, so their sum is strictly positive.) This shows that the field $\mathbb{Z}/p\mathbb{Z}$ cannot be ordered because in this field

$$0 = 1 + \ldots + 1 \ (p \text{ times}) = 1^2 + \ldots + 1^2 \ (p \text{ times}).$$

Looking over what we've just done, one might be inspired to ask the question: for which prime numbers p is there an element $x \in \mathbb{Z}/p\mathbb{Z}$ such that $x^2 = -1$? In other words, in which fields $\mathbb{Z}/p\mathbb{Z}$ is -1 a square? For any given prime p, we can determine all the squares in $\mathbb{Z}/p\mathbb{Z}$ just by reducing each of $1^2, 2^2, \ldots, (p-1)^2$ modulo p. Let us record what happens for small values of p:

EXAMPLE 9.25. *In any field $\mathbb{Z}/p\mathbb{Z}$, we have $0 = 0^2$. We record the nonzero squares for the following values of p:*

a) *For $p = 2$ or $p = 3$, the only nonzero square is 1.*
b) *For $p = 5$, the nonzero squares are 1 and $4 = -1$.*
c) *For $p = 7$, the nonzero squares are 1, 2 and 4.*
d) *For $p = 11$, the nonzero squares are 1, 3, 4 5 and 9.*
e) *For $p = 13$, the nonzero squares are 1, 3, 4, 9, 10 and $12 = -1$.*

Even this small amount of computation suggests some patterns. First it seems that for each $p \geq 3$, exactly half of the $\frac{p-1}{2}$ nonzero elements of $\mathbb{Z}/p\mathbb{Z}$ are squares. This is true and not hard to show: every nonzero square in $\mathbb{Z}/p\mathbb{Z}$ is of course of the form x^2 for some $x \in (\mathbb{Z}/p\mathbb{Z})^\times := \mathbb{Z}/p\mathbb{Z} \setminus \{0\}$, and moreover we have $x^2 = y^2$ if and only if $(\frac{x}{y})^2 = 1$ if and only if $y = x$ or $y = -x$. Therefore the squaring function

$$x \in (\mathbb{Z}/p\mathbb{Z})^\times \mapsto x^2 \in (\mathbb{Z}/p\mathbb{Z})^\times$$

is exactly two-to-one – every value is assumed twice – so the cardinality of its image is half the cardinality of its domain, which is $\frac{p-1}{2}$.

We move on now to a more interesting observation: sometimes -1 is a square in $\mathbb{Z}/p\mathbb{Z}$ and sometimes it isn't. You may or may not have a guess based on the data given, but if not it is no problem to compute further: one calculates (either by hand or with computer assistance) that among primes $2 < p < 100$, we have that -1 is a square in $\mathbb{Z}/p\mathbb{Z}$ if

$$p \in \{5, 13, 17, 29, 37, 41, 53, 61, 73, 89, 97\}$$

and that -1 is *not* a square in $\mathbb{Z}/p\mathbb{Z}$

$$p \in \{3, 7, 11, 19, 23, 31, 43, 47, 59, 67, 71, 79, 83\}.$$

At this point it is hard not to notice that all the primes in the the first list are congruent to 1 modulo 4 and all the primes in the second list are congruent to 3 modulo 4. We are now going to prove that this is the case for all odd primes. The proof we give will draw us into a much deeper number-theoretic question. Namely, we say that an integer N is **a sum of two squares** if there are $x, y \in \mathbb{Z}$ such that $x^2 + y^2 = N$. Evidently then N must be non-negative; as $0 = 0^2 + 0^2$, we lose nothing by restricting to positive integers N.

LEMMA 9.26. *Suppose that $N \in \mathbb{Z}^+$ is a sum of two squares: there are $x, y \in \mathbb{Z}$ such that $x^2 + y^2 = N$.*

a) *N is* not *congruent to 3 modulo 4.*
b) *If N is even, then also $\frac{N}{2}$ is a sum of two squares.*

PROOF. a) Since $0^2 \equiv 2^2 \equiv 0 \pmod 4$ and $1^2 \equiv 3^2 \equiv 1 \pmod 4$, the squares in $\mathbb{Z}/4\mathbb{Z}$ are 0 and 1. Thus the elements of $\mathbb{Z}/4\mathbb{Z}$ of the form $x^2 + y^2$ are 0, 1 and 2, but not 3.
b) If N is even and $x^2 + y^2 = N$, then $x \equiv y \pmod 2$, so $\frac{x+y}{2}, \frac{x-y}{2} \in \mathbb{Z}$ and

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

THEOREM 9.27. *Let $p \geq 3$ be a prime number. The following are equivalent:*

(i) *We have $p \equiv 1 \pmod 4$.*
(ii) *There is $x \in \mathbb{Z}$ such that $p \mid x^2 + 1$.*
(iii) *There are $x, y \in \mathbb{Z}$ with $\gcd(p, xy) = 1$ such that $p \mid x^2 + y^2$.*

PROOF. (i) $\implies$ (ii): Until further notice, let p be any odd prime. A subset S of $\mathbb{Z}$ is a **reduced residue system modulo p** if it has exactly $p-1$ elements, if no $x \in S$ is divisible by p and if for all $x, y \in S$ with $x \neq y$ then $x \pmod p \neq y \pmod p$. Then every element of $(\mathbb{Z}/p\mathbb{Z})^\times$ is of the form $x \pmod p$ for a unique $x \in S$, so for any two reduced residue systems S and T modulo p we have

$$\prod_{x \in S} x \equiv \prod_{y \in T} \pmod p.$$

The most evident reduced residue system modulo p is

$$S := \{1, 2, \ldots, p-1\};$$

since $\prod_{x \in S} x = (p-1)!$, we see that Wilson's Theorem can be reformulated as follows: for any reduced residue system T modulo p we have

$$\prod_{y \in T} y \equiv -1 \pmod p.$$

The second most evident reduced residue system modulo p is the set

$$T := \{-(\frac{p-1}{2}), -(\frac{p-1}{2}) + 1, \ldots, -1, 1, 2, \ldots, \frac{p-1}{2}\}$$

of nonzero integers of absolute value less than $\frac{p}{2}$. There are $p-1$ such integers, none of them is 0 modulo p, and the difference between the greatest and the least element of T is $\frac{p-1}{2} - (-\frac{p-1}{2}) = p-1$, so no two elements of T can differ by a multiple of p and T is a reduced residue system. Applying Wilson's Theorem to T, we find that

$$-1 \equiv \prod_{y \in T} y \equiv (-1)^{\frac{p-1}{2}} (\frac{p-1}{2}!)^2 \pmod p.$$

Now taking $p \equiv 1 \pmod 4$, we get that $(-1)^{\frac{p-1}{2}} = 1$, so

$$(\frac{p-1}{2}!)^2 \equiv -1 \pmod p,$$

so the conclusion of (ii) holds with $x := \frac{p-1}{2}!$.
(ii) $\implies$ (iii): If $p \mid x^2 + y^2$, then $p \mid x \iff p \mid y$. By assumption we have $p \mid x^2 + 1$ and certainly p does not divide 1, so also $p \nmid x$.
(iii) $\implies$ (i): For any prime number p **complete residue system modulo p** is a subset $S \subseteq \mathbb{Z}$ such that every integer is congruent modulo p to a unique element

of S. Thus every complete residue syssem modulo p is obtained from a reduced residue system by adjoining some integer that is divisible by p, so for any $p \geq 3$,

$$\{-(\frac{p-1}{2}), -(\frac{p-1}{2})+1, \ldots, -1, 0, 1, \ldots \frac{p-1}{2}\}$$

is a complete residue system modulo p in which each element has absolute value less than $p/2$.

Seeking a contradiction, we assume there is some prime $p \equiv 3 \pmod 4$ and integers x and y such that $p \mid x^2 + y^2$ and neither x nor y is divisible by p. Let p be the least prime satisfying all of these conditions. By the above paragraph, there are integers X and Y such that $x \equiv X \pmod p$, $y \equiv Y \pmod p$ and $|X|, |Y| < \frac{p}{2}$. Then $p \mid X^2 + Y^2$. Also X and Y are not divisible by p hence are not 0, so $X^2 + Y^2 > 0$, and also

$$X^2 + Y^2 < (\frac{p}{2})^2 + (\frac{p}{2})^2 = \frac{p^2}{2},$$

so there is $1 \leq m < \frac{p}{2}$ such that

$$X^2 + Y^2 = mp.$$

Suppose $2 \mid m$. As in the proof of Lemma 9.26a), we have $X \equiv Y \pmod 2$ and

$$\left(\frac{X+Y}{2}\right)^2 + \left(\frac{X-Y}{2}\right)^2 = \frac{m}{2}p.$$

If $p \mid \frac{X+Y}{2}$, then also $p \mid \frac{X-Y}{2}$, so $p \mid \frac{X+Y}{2} + \frac{X-Y}{2} = X$, a contradiction. Thus by repeated application of Lemma 9.26a) we may assume that m is odd and write

$$m = p_1 \cdots p_r q_1 \cdots p_s$$

with not necessarily distinct primes $p_1, \ldots, p_r, q_1, \ldots, q_s < p$ such that $p_i \equiv 1 \pmod 4$ for all $1 \leq i \leq r$ and $q_j \equiv 3 \pmod 4$ for all $1 \leq j \leq s$. Then for any $1 \leq j \leq s$ we have $q_j \mid X^2 + Y^2$; by the minimality of p, this implies that we may write $X = pX_1$, $Y = pY_1$ for $X_1, Y_1 \in \mathbb{Z}$ and then we have

$$X_1^2 + Y_1^2 = \frac{m}{q_j^2}p.$$

Continuing in this way we can remove all of the primes q_j, getting a representation

$$C^2 + D^2 = p_1 \cdots p_r p.$$

Then

$$C^2 + D^2 \equiv p_1 \cdots p_r p \equiv 1 \cdot 1 \cdots 1 \cdot 3 \equiv 3 \pmod 4,$$

contradicting Lemma 9.26b). □

The proof of Theorem 9.27 gives a bit more than the statement advertises: if $p \equiv 1 \pmod 4$, then not only is there $x \in \mathbb{Z}$ such that $x^2 \equiv -1 \pmod p$, but we may explicitly take $x := \frac{p-1}{2}!$. This came from applying Wilson's Theorem to the reduced residue system consisting of integers x with $0 < |x| \leq \frac{p-1}{2}$. For a prime $p \equiv 3 \pmod 4$, one can still make such an argument, but the answer is different: see Exercise 9.10.

PROPOSITION 9.28 (Euler). *Let $N \in \mathbb{Z}^+$ and let p be a prime divisor of N. If both N and p are sums of two squares, so is $\frac{N}{p}$.*

PROOF. There are integers a, b, u, v such that

$$N = u^2 + v^2 \text{ and } p = a^2 + b^2.$$

Then

$$p \mid u^2(a^2 + b^2) - b^2(u^2 + v^2) = a^2u^2 - b^2v^2 = (au + bv)(au - bv),$$

so $p \mid au + bv$ or $p \mid au - bv$. Since $p = a^2 + b^2 = a^2 + (-b)^2$, we may replace b with $-b$ if necessary and thereby assume that $p \mid au + bv$. We must have $p \nmid ab$: as above, if $p = a^2 + b^2$ divdes one of a and b, it divides both, and then $p^2 \mid p$, a contradiction. So we have

$$(a/b)^2 \equiv -1 \pmod{p},$$

where by $\frac{1}{b}$ we mean the inverse of b in the field $\mathbb{Z}/p\mathbb{Z}$. It follows that

$$(b/a)^2 \equiv (-1)^{-1} \equiv -1 \pmod{p}.$$

Now we have

$$au + bv \equiv 0 \pmod{p};$$

multiplying this congruence through by b/a, we get

$$0 \equiv bu + \frac{b^2}{a}v \equiv bu + (\frac{b}{a})^2 av \equiv bu - av \pmod{p}.$$

So we have

$$\frac{au + bv}{p}, \ \frac{bu - av}{p} \in \mathbb{Z}$$

and thus

$$\left(\frac{au + bv}{p}\right)^2 + \left(\frac{bu - av}{p}\right)^2 =$$

$$\frac{1}{p^2}\left(a^2u^2 + 2abuv + b^2v^2 + b^2u^2 - 2abuv + a^2v^2\right)$$

$$= \frac{1}{p^2}\left(a^2(u^2 + v^2) + b^2(u^2 + v^2)\right) = \frac{1}{p^2}(a^2 + b^2)(u^2 + v^2) = \frac{Np}{p^2} = \frac{N}{p}$$

is a sum of two squares. □

We can now characterize the set of integers that are sums of two squares. This is arguably the first "non-ancient" number-theoretic theorem, by which I mean that it goes safely beyond what the ancient Greeks knew. The result was conjectured by the French mathematician Albert Girard in 1625, although in regards to its proof he was only able to contribute the (extremely easy) Lemma 9.26a). A proof was first given by Pierre de Fermat, in a later he wote to Marin Mersenne dated December 25, 1640. Because of this, the result is sometimes called 'Fermat's Christmas Theorem."

The Two Squares Theorem is an archetypical number-theoretic result in that the simplicity of its statement is charming and inviting but somewhat misleading: there are now dozens if not hundreds of proofs known, but to understand – let alone find! – any of them is much more challenging than apprehending or even guessing the result. There are some quite natural proofs that use more advanced concepts, such as unique factorization in certain "higher arithmetic rings" (rings that contain $\mathbb{Z}$ but also certain irrational quantities, such as $\sqrt{-1}$) or the theory of binary quadratic forms. See [**Cl-NT**, Chapter 3] for a proof of the former type that is still at the undergraduate level. The proof that we will give is sort of "neo-classical": it follows the same basic "infinite descent" strategy as Fermat's proof

and includes one of the pieces of a proof given by Euler in the 18th century, but it leans on a more carefully structured strong inductive argument in roughly the same way as Lindemann-Zermelo's proof of the Fundamental Theorem of Arithmetic.

THEOREM 9.29 (Fermat Two Squares Theorem). *Let N be a positive integer, and write*

$$N = 2^a p_1^{a_1} \cdots p_r^{a_r} q_1^{b_1} \cdots q_s^{b_s} \tag{56}$$

with $ar, s, \in \mathbb{N}$, $a_1, \ldots, a_r, b_1, \ldots, b_s \in \mathbb{Z}^+$, $p_1 < \ldots < p_r$ primes congruent to 1 modulo 4 and $q_1 < \ldots q_s$ primes congruent to 3 modulo 4. Then N is a sum of two squares if and only if b_j is even for all $0 \leq j \leq s$.

PROOF. Step 1: Let $p \equiv 1 \pmod 4$ be a prime. We will show that p is a sum of two squares. By strong induction, we may assume that every prime q with $q \equiv 1 \pmod 4$ and $q < p$ is a sum of two squares. (We actually don't need a base case here, e.g. because we may formulate the strong inductive hypothesis as "if n is a prime that is 1 modulo 4..." so that it vacuously applies to $n = 1$. But if you prefer: the smallest prime that is 1 mod 4 is $5 = 2^2 + 1^2$.) By Theorem 9.27 there are $x, y \in \mathbb{Z}$ and $1 \leq m < \frac{p}{2}$ such that

$$x^2 + y^2 = mp.$$

If m is even, then using Lemma 9.26 we get that $\frac{m}{2}p$ is a sum of two squares, so repeatedly applying this result we may assume that m is odd. If m is divisible by a prime $q \equiv 3 \pmod 4$, then by Theorem 9.27 this implies that both x and y are divisible by q, and then dividing through by q^2 gives $\frac{m}{q}p$ as a sum of two squares. Appplying this argument repeatedly, we get a representation

$$X^2 + Y^2 = p_1 \cdots p_r p$$

with each $p_i \equiv 1 \pmod 4$ and $p_i < p$. By our inductive hypothesis, each p_i is then a sum of two squares, so by Proposition 9.28 we get a representation of $(p_1 \cdots p_r / p_i)p$ as a sum of two squares; continuing in this way, we eventually get a representation of p as a sum of two squares.

Step 2: One directly checks the **Diophantus-Brahmagupta-Fibonacci identity**

$$\forall a, b, c, d \in \mathbb{R},\ (a^2 + b^2)(c^2 + d^2) = (ac - bd)^2 + (ad + bc)^2.$$

This identity implies that if $N_1, N_2 \in \mathbb{Z}^+$ are each sums of two squares, then so is $N_1 N_2$. After Step 1, we know that every prime $p \equiv 1 \pmod 4$ is a sum of two squares. Certainly $2 = 1^2 + 1^2$ is a sum of two sqares; and for any prime $q \equiv 3 \pmod 4$ trivially we have $q^2 = q^2 + 0^2$ is a sum of two squares. A positive integer N is a product of integers of this form so long as for every prime $q \equiv \pmod 4$, the prime q divides N with even multipiicity: that is, there is some $a \in \mathbb{Z}^+$ such that $p^{2a} \mid N$ and $p^{2a+1} \nmid N$. This shows that if N is written as in (56) and b_j is even for all $1 \leq j \leq s$ then N is a sum of two squares.

Step 3: Finally, we must show that if $N \in \mathbb{Z}^+$ is a sum of two squares, then every prime $q \equiv 3 \pmod 4$ divides N to even multipliciity. This inclues the case in which $q \nmid N$, so we may assume that $q \mid N$. Then $q \mid x^2 + y^2 = N$, so by Theorem 9.27 both x and y are divisible by q, so $\frac{N}{q^2} = (x/q)^2 + (y/q)^2$ is again a sum of two squares. If now $\frac{N}{q^2}$ is not divisible by q then $q^2 \mid N$ and $q^3 \nmid N$, so q divides N to even multiplicity. If $\frac{N}{q^2}$ is divisible by q, then the above argument applies to show that $q^4 \mid N$ and $\frac{N}{q^4}$ is a sum of two squares. The positive ineteger N can't

be infinitely divisible by q, so after some number k of steps, we find that $\frac{N}{q^{2k}}$ is not divisible by q, so q divides N with even multiplicity $2k$. □

3. Graph Theory

3.1. Some Basic Terminology. The notion of a graph comes in several natural variations. In this text we will only consider graphs that are **simple** and **undirected**. Such a graph is given by a set V, of **vertices**, together with a set $E \subseteq 2^V$ of two-element subsets $\{x, y\}$ of V, called **edges**.

The upshot of this is: given any pair of distinct vertices, either there is an edge between them or there isn't. The data of this is captured in the **adjacency relation** on V: for $x, y \in V$ we sa that x is **adjacent** to y and write $x \sim y$ if $\{x, y\} \in E$. To be sure, if we know the adjacency relation on V, we know the graph: the edge set is then

$$\{\{x, y\} \mid x \sim y\}.$$

The adjacency relation is clearly symmetric. It is also anti-reflexive: since E consists of two-element subsets of V, for no $x \in V$ do we have $x \sim x$. Conversely, given any relation $\sim$ on a set V that is anti-reflexive and symmetric, if we put

$$E \coloneqq \{\{x, y\} \mid x \sim y\}$$

then (V, E) is a graph.

Actually it is reasonable to "change the rules" on the definition of a graph in several different ways. Namely:

• If one eliminates the requirement of anti-reflexivity, one gets graphs in which a vertex is allowed to be adjacent to itself: we call this a "loop."
• If one eliminates the requirement of symmetry, one gets graphs in which we consider **directed edges**: we may have an edge from x to y but not an edge from y to x. This variation is called **directed graphs** or **digraphs**.
• Finally, one can also consider the situation in which for any ordered pair (x, y) of vertices we attach a *set* of edges, thus possibly having multiple edges running between the same pair of vertices.

Each of these variation (and even each possible combination of these variations) is very natural for certain kinds of problems, and each has been widely studied. Because we are only giving a brief introduction to graph theory here, it seems best to focus on our attention on one specific definition, and so we will.

A graph $G = (V, E)$ is **finite** if V and E are both finite. Notice that if V is finite of size n, then since E is a set of 2-element subsets of V, E must also be finite and indeed $\#E \leq \binom{n}{2}$. If the graph is finite, it is most convenient to take $V = [n] = \{1, \ldots, n\}$, and then the idea is that for each of the $\binom{n}{2} = \frac{n(n-1)}{2}$ 2-element subsets of $[n]$, we get to independently decide whether to include the edge or not. So the number of graphs with vertex set $[n]$ is $2^{\frac{n(n-1)}{2}}$.

A graph $G = (V, E)$ is **locally finite** if for all $v \in V$, the set of vertices adjacent to v is finite. In this case, we define the **degree** $\deg(v)$ of v to be $\#\{w \in V \mid v \sim w\}$,

i.e., the number of vertices adjacent to v. (More generally, we can define $\deg(v)$ in the same way whenever there are only finitely many vertices adjacent to v.

EXAMPLE 9.30. *We define a graph $G = (\mathbb{Z}, E)$ as follows: for $x, y \in \mathbb{Z}$ we have $x \sim y$ if and only if $|x - y| = 1$. In other words, for all $n \in \mathbb{Z}$, we have that n is adjacent precisely to $n-1$ and to $n+1$. This graph is infinite but locally finite, and every vertex has degree 2. We call this graph the* ***doubly infinite path****.*

PROPOSITION 9.31. *Let $G = (V, E)$ be a finite graph.*

a) *We have*
$$\sum_{v \in V} \deg(v) = 2\#E.$$
b) *The number of odd degree vertices is even.*

PROOF. a) We have
$$\sum_{v \in V} \deg(v) = \sum_{v \in V} \#\{e \in E \mid v \in e\}.$$
Every $e \in E$ is of the form $\{v, w\}$ for some vertices $v \neq w$ so contributes exactly twice to $\sum_{v \in V} \#\{e \in E \mid v \in e\}$, and thus
$$\sum_{v \in V} \#\{e \in E \mid v \in e\} = 2\#E.$$
b) Let $V_e \subseteq V$ be the subset of vertices of even degree, and let $V_o \subseteq V$ be the subset of vertices of odd degree, so $V = V_e \coprod V_o$. Then going modulo 2 we get
$$0 \equiv 2\#E \equiv \sum_{v \in V} \deg(v) \pmod 2$$
$$\equiv \sum_{v \in V_e} \deg(v) + \sum_{v \in V_o} \deg(v) \equiv \sum_{v \in V_e} 0 + \sum_{v \in V_o} 1$$
$$\equiv \#V_o \pmod 2.$$
□

A vertex v in a graph is **isolated** if it has degree zero, i.e., if there are no vertices adjacent to v.

EXAMPLE 9.32.

a) *If (V, E) is a graph with $\#V = 1$, then the unique vertex of G is isolated.*
b) *For any set V, we define the* ***empty graph*** *$(V, \varnothing)$: it has no edges. It is the unique graph on vertex set V for which every vertex is isolated.*

A vertex v in a graph is **pendant** if it has degree 1, i.e., there is exactly one edge coming out of V.

A **finite walk** in a graph G is a finite list
$$\ell : v_0, \ldots, v_n$$
of vertices such that $v_i \sim v_{i+1}$ for all $0 \leq n - 1$: here $n \in \mathbb{N}$. That is, each vertex is adjacent to the next. We say that the walk $v_0, \ldots, v_n$ has **length** n. (Thus the length of the walk is the number of "steps," or in other words the number of edges $\{v_0, v_1\}, \ldots, \{v_{n-1}, v_n\}$: this is one less than the number of vertices.) We say that the walk is **from** v_0 **to** v_n.

Notice that we allow $\ell : v_0$: this is a walk of length 0 from v_0 to itself. (This may seem silly; we will see shortly why this is a useful definition.)

A **finite circuit** is a finite walk in which $v_n = v_0$, i.e., we end where we started. A **finite path** is a finite walk in which all of the vertices $v_0, \ldots, v_n$ are distinct. A **cycle** is a finite circuit $v_0, \ldots, v_{n-1}, v_n = v_0$ of length $n \geq 3$ in which $v_0, \ldots, v_{n-1}$ are all distinct.

Why did we require $n \geq 3$ in the above definition? There is no circuit of length 1 since no edge runs from a vertex to itself. A circuit of length 2 is of the form v_0, v_1, v_0 with $v_0 \sim v_1$: that is, it consists of the same edge traversed twice, in opposite directions. Any graph with edges evidently has a circuit of length 2, but not every graph has cycles according to our definition.

LEMMA 9.33. *Let $G = (V, E)$ be a graph.*

a) *Let v, w be distinct vertices of G. The following are equivalent:*
   (i) *There is a path from v to w.*
   (ii) *There is a walk from v to w.*

b) *Let*
$$\ell : v = a_0, a_1, \ldots, a_n = w$$
be a walk from v to w. If for some $m \leq n$ we have that $a_0, \ldots, a_m$ are all distinct vertices, then there is a path frrom v to w of the form $v = a_0, \ldots, a_m, b_{m+1}, \ldots, b_n = w$.

c) *Let $\approx$ be the relation on V given by $x \approx y$ if and only if there is a walk from x to y. Then $\approx$ is an equivalence relation on V.*

PROOF. a) (i) $\implies$ (ii): Since every path is a walk, this is immediate.
(ii) $\implies$ (i): Let
$$\ell : v = v_0, v_1, \ldots, v_n = w$$
be a walk from v to w. What prevents ℓ from being a path is vertices appearing more than once. Let i be the least index such that there is $j > i$ with $v_i = v_j$. Then $v_i, v_{i+1}, \ldots, v_j = v_i$ is a circuit, so as for walks in real life, if you come back where you started you would have walked more efficiently. Consider instead
$$\ell' : v = v_0, v_1, \ldots, v_i, v_{j+1}, \ldots, v_n = w.$$
Since $v_i = v_j \sim v_{j+1}$, this is still a walk from v to w, of length $n - (j - i) < n$. If ℓ' is still not a path, we can repeat the process, getting another walk from v to w of yet smaller length, and so forth. Since a walk between distinct vertices must have length at least 1, this process must terminate after finitely many steps, yielding a path from v to w.
b) In part a) we gave a specific procedure that constructs a path from v to w given a walk from v to w. This procedure involves removing vertices that have already appeared earlier in the list, so if the vertices $a_0, \ldots, a_m$ are distinct, then none of these vertices get removed.
c) Reflexivity: For all $v \in V$, v is a walk of length 0 from v to itself.[2]

[2]Good thing we allowed walks of length 0.

Symmetry: For all $v, w \in V$, if $v = v_0, v_1, \ldots, v_n = w$ is a walk from v to w, then reverse it:

$$w = v_n, v_{n-1}, \ldots, v_1, v_0 = v$$

gives a walk from w to v.
Transitivity: If $x, y, z \in V$ and we have a walk

$$\ell_1 : x = a_0, a_1, \ldots, a_m = y$$

from x to y and also a walk

$$\ell_2 : y = b_0, \ldots, b_n = z$$

from y to z, then

$$x = a_0, \ldots, a_m = b_0, b_1, \ldots, b_n = z$$

is a walk from x to z. □

For a graph $G = (V, E)$, the $\approx$-equivalence classes are called the **connected components** of G. We denote the connected component of $v \in V$ by $\mathfrak{c}(v)$. A graph G is **connected** if there is exactly one $\approx$-equivalence class.[3] By Lemma 9.33, a graph is connected if and only if for every pair of distinct vertices $v \neq w$, there is a walk from v to w.

3.2. Cycles.

PROPOSITION 9.34. *Let $G = (V, E)$ be a finite graph.*

a) *The following are equivalent:*
   (i) *Every vertex $x \in V$ has even degree.*
   (ii) *The edge set E is a disjoint union of cycles.*

b) *If G has no cycles and at least one edge, then it has at least two pendant (i.e., of degree one) vertices.*

PROOF. a) (i) $\implies$ (ii): We go by induction on the number of edges. The base case, $\#E = 0$, is immediate: E is the empty union of cycles! Suposing that E is nonempty and that every vertex has even degree, it will suffice to find a single cycle C in G. For then, if we remove all the edges in that cycle, then we reduce the degree of every vertex in the cycle by 2 and leave all other vertex degrees unchanged, so we maintain the condition that every vertex has even degree while decreasing the number of edges. Thus induction applies.

If G is any finite graph with n vertices, then every path involves at most n vertices so has length at most $n-1$. Therefore paths of maximal length must exist, and this maximal length is positive iff $E \neq \varnothing$.

Let $P : x_0, \ldots, x_k$ be a path of maximal length in G. Since $e \coloneqq \{x_0, x_1\} \in E$, the vertex x_0 has positive degree.But also by assumption v_0 has even degree, so there is another edge $e \neq e' = \{x_0, y\} \in E$. Since the path P has maximal length, $y, x_0, \ldots, x_k$ cannot be a path, which means that $y = x_i$ for some $2 \leq i \leq k$. Thus $y, x_0, x_1, \ldots, x_k$ is a cycle in G.
(ii) $\implies$ (i): Since every cycle in G contributes 2 to the degree of each vertex it contains, this is clear.
b) As above, let $P : x_0, \ldots, x_k$ be a path in G of maximal length. Since G has at least one edge, we have $k \geq 1$. We claim that the initial vertex x_0 and the final vertex x_k are both pendant, which will complete the proof. Moreover it suffices to

[3]Thus the empty graph $G = (\varnothing, \varnothing)$ *is not* connected according to our definintion.

prove that the final vertex of a path of maximal length is pendant, for then applying this to the reversed path $\overline{P} : x_k, x_{k-1}, \ldots, x_0$, we conclude that x_0 is pendant.

Seeking a contradiction, we suppose that there is some vertex $y \neq x_{k-1}$ such that $x_{k-1} \sim y$. The argument is now much the same as the one we made in part a) above: we must have $y = x_i$ for some $0 \leq i \leq k-2$ for otherwise $x_0, \ldots, x_k, y$ is a longer path than P, and thus $x_i, \ldots, x_k, y$ is a cycle in G. □

Let $G = (V, E)$ be a finite graph. An **Eulerian circuit** in G is a circuit

$$C : x_0, \ldots, x_n = x_0$$

in which every edge of G appears exactly once: precisely, for each $e \in G$, there is a unique $0 \leq i \leq n-1$ with $e = \{x_i, x_{i+1}\}$. In a graph without isolated vertices, any walk in which each edge appears must visit every vertex, and thus that graph is connected. On the other hand, a finite graph admits an Eulerian circuit if and only if the graph obtained by removing all the isoslated vertices admits an Eulerian circuit (and indeed, these two graphs have precisely the same Eulerian circuits). So when searching for Eulerian circuits, we may as well assume our graph is connected.

THEOREM 9.35. *For a connected finite graph G, the following are equivalent:*

(i) *The graph G admits an Eulerian circuit.*
(ii) *Every vertex in G has even degree.*

PROOF. (i) $\implies$ (ii): Let $C : x_0, \ldots, x_n$ be an Eulerian circuit in G. Let $v \in V$ and let $e \in E$ be an edge that contains v, so $e = \{x_i, x_{i+1}\}$ for a unique $0 \leq i \leq n-1$. We call e an **incoming edge** for v if $v = x_{i+1}$ and an **outgoing edge** for v if $v = x_i$. If $e = \{x_i, x_{i+1}\}$ is an outgoing edge for v, then $e' = \{x_{i-1}, x_i\}$ is an incoming edge for v. (When $i = 0$, we put $x_{-1} = x_{n-1}$.) The mapping $e \mapsto e'$ gives a bijection from the set of outgoing edges for v to the set of incoming edges for v. Since these sets are disjoint, we have partitioned the set of all vertices adjacent to v into two finite sets of the same size, and therefore the set of vertices adjacent to v has even size.
(ii) $\implies$ (i): Suppose that every vertex in the connected finite graph G has even degree. Then by Proposition 9.34a), E is a disjoint union of cycles. We go by induction on the number n of cycles. The base case is $n = 1$, in which case G is itself a cycle, which gives an Eulerian circuit. Suppose now that $n \geq 2$ and that the result holds for all connected graphs in which every vertex has even degree for which the edge set is a disjoint union of $n-1$ cycles, and write

$$G = C_1 \coprod \ldots \coprod C_n$$

as a disjoint union of n cycles. Choose an edge e_2 that does not lie in C_1 but contains a vertex that does lie in C_1. (This must be possible: either C_1 contains all the vertices of G, in which case every vertex e not in C_1 has this property, or there are vertices not in C_1, in which case since G is connected, some edge must run between a vertex in C_1 and a vertex not in C_1.) After reordering the cycles $C_2, \ldots, C_n$ if necessary, we let C_2 be the cycle containing e_2. Then the graph G_2 with vertex set all vertices lying in C_1 or C_2 and edge set $C_1 \coprod C_2$ is connected, and every vertex of G_2 has even degree. We can continue in this manner until we obtain a subgraph G_{n-1} of G in which the edge set is $C_1 \coprod \cdots \coprod C_{n-1}$, and the vertex set is all the vertices lying in any of these edges, and the graph G_{n-1} is connected and has each vertex of even degree. Since every edge of G that does not lie in C_n must

lie in G_{n-1} and G is connected, there must a vertex v that lies in both G_{n-1} and in C_n. By induction, there is an Eulerian circuit in G_{n-1}. If a finite graph admits an Eulerian circuit, then it admits an Eulerian circuit starting and ending at any given vertex, so there is an Eulerian circuit

$$P_1 : v = x_0, \ldots, x_n = v.$$

The edges of C_n can be written as

$$P_2 : v = y_0, y_1, \ldots, y_m = v,$$

and then

$$P_1 \cdot P_2 : v = x_0, \ldots, x_n = v = y_0, y_1, \ldots, y_m = v$$

is an Eulerian circuit in G. □

An **Eulerian walk** in a finite graph is a walk in which each edge appears exactly once. The results for Eulerian walks are similar to those for Eulerian circuits; they are developed in Exercise E.E.

PROPOSITION 9.36. *Let $G = (V, E)$ be a connected graph. The following are equivalent:*

(i) *For all $v, w \in V$, there is a* unique *path from v to w.*
(ii) *The graph G has no cycles.*

A connected graph satisfying these equivalent conditions is called a ***tree****.*

PROOF. (i) $\implies$ (ii): We use the contrapositive. Suppose that G has a cycle $v_0, v_1, \ldots, v_k = v_0$ with $k \geq 3$. Then $v_0, \ldots, v_{k-1}$ and v_0, v_{k-1} are both paths from v_0 to v_{k-1}, and they are different: the first path has length $k - 1 \geq 2$, while the second path has length 1.
(ii) $\implies$ (i): Again, we use the contrapositive: suppose that for some vertices v, w there are two diferent paths from v to w. We must have $v \neq w$ because the only path from a vertex to itself is the path of length 0. If

$$\ell_1 : v = a_0, a_1, \ldots, a_m = w \text{ and } \ell_2 : v = b_0, b_1, \ldots, b_n = w$$

are two different paths from v to w there must be some $i \leq \min(m-1, n-1)$ such that $a_i = b_i$ and $a_{i+1} \neq b_{i+1}$. (Without loss of generality we may assume that $m \leq n$. If $m < n$ it is not possible for $a_i = b_i$ for all $0 \leq i \leq m$ because then $b_m = a_m = w$ but also $b_n = w$, so ℓ_2 contains a repeated vertex, which is not allowed.) Letting $v' = a_i = b_i$ we replace our paths from v to w with paths

$$\ell_1' : v' = a_i, a_{i+1}, \ldots, a_m = w, \text{ and } \ell_2' : v' = b_i, b_{i+1}, \ldots, b_n = w,$$

and then we reduce to a situation where we have two different paths from v' to w for which the first directed edge in the path is already different. To keep the notation simple, we will assume that our originally chosen paths ℓ_1 and ℓ_2 from v to w have that property.

Since $v = a_0, \ldots, a_m = w, b_{n-1}, \ldots, b_1$ is a walk from v to b_1 such that a_0 and a_1 are distinct, by Lemma 9.33b) there is a path

$$P : v = c_0, c_1 = a_1, c_2, \ldots, c_k = b_1$$

from v to b_1. Since $a_1 \neq b_1$, this path has length $k \geq 2$. Then

$$v = c_0, c_1, \ldots, c_k = b_1, v$$

is a $(k+1)$-cycle. □

The graph with one vertex (and no edges, necessarily) is a tree. There is essentially (more precisely, "up to isomorphism," a concept we will not formalize here) only one tree with two vertices: we connect the two vertices with an edge, getting a path of length 2. There is essentially only one tree with three vertices: a path of length 3. At four vertices, things get more interesting: in addition to the path of length 4 there is a "star" obtained by putting one vertex in the middle and drawing edges between it and each of the other three vertices. The middle vertex has degree 3, so this is different from any path. How many trees are there on five vertices?

Trees are a very interesting class of graphs, because they have a certain "Goldilocks" property. Namely, suppose that we start with a graph $G = (V, E)$ and add edges to it. If G is connected, then the new graph must still be connected, because any path in the old graph is still a path in the new graph. However, subtracting edges from a connected graph may ruin connectedness (if $\#V > 1$ and we remove enough of the edges – e.g. all of them – then we get a disconnected graph). Now suppose that we start with a graph $G = (V, E)$ and remove edges. If G has no cycles, then the new graph still must have no cycles. But if we add edges, then we may create cycles. So a tree has, in a sense, a "just right" number of edges compared to the number of vertices. Our next theorem gives a sense in which this is true. First we need a preliminary result.

THEOREM 9.37. *Let $T = (V, E)$ be a finite tree. Then $\#V = 1 + \#E$. That is, the number of vertices is one more than the number of edges.*

PROOF. We go by induction on n, the number of vertices of our finite tree T. Base Case ($n = 1$): If there is one vertex, there are no edges, and $1 = 1 + 0$: OK. Induction Step: Let $n \in \mathbb{Z}^+$, assume that every finite tree with n vertices has $n-1$ edges, and let T be a finite tree with $n+1$ vertices. By Proposition 9.34b) we have a pendant vertex v, so that there is exactly one edge e coming out of v. Here is the crux of the entire proof: we may **prune** the tree by removing the pendant vertex v and its edge e. This leaves us with a tree T'. Certainly T' has no cycles: if we start with a graph with no cycles and remove stuff, certainly we have no cycles. Moreover a pendant vertex cannot occur in the middle of a path (i.e., as neither the initial or terminal vertex of the path), because being a middle vertex in a path means there are at least two edges coming out of it. If v_1 and v_2 are distinct vertices in T', then whatever path connected them in T cannot include the removed pendant vertex v, so it still gives a path in T'. The tree T' has $n-1$ vertices, so by induction it has $n-2$ edges. On the other hand, clearly T' has one less edge than T, so T must have $n-2+1-n-1$ edges. We're done. □

REMARK 9.38. *Theorem 9.37 provides a nice example of a certain kind of induction proof, in which we have a problem of discrete "complexity" and we solve it by showing that we can always "reduce the complexity" at every step. In fact it is this application of induction that motivated us to include Theorem 9.37 in this text. At first it was placed in Chapter 7, as an application of induction. This result together with Ramsey's Theorem was the genesis of our decision to include some general coverage of graph theory.*

For a finite graph $G = (V, E)$ we define the **Euler characteristic**

$$\chi(G) = \#V - \#E.$$

Thus Theorem 9.37 states that the Euler characteristic of a tree is 1.

Theorem 9.37 raises several interesting questions.

QUESTION 1.

a) *Is the converse of Theorem 9.37 true? That is, must a finite graph G with $\chi(G) = 1$ be a tree?*
b) *Can we add edges to a tree and still get a tree? More precisely, if If $T = (V, E)$ is a tree, and $G = (V, E')$ with $E' \supsetneq E$, can G be a tree?*
c) *Can we subtract edges from a tree and still get a tree? More preicsely, if $T = (V, E)$ is a tree, and $G = (V, E')$ with $E' \subsetneq E$, can G be a tree?*

To answer the first question, we introduce the coproduct of graphs. Let $\{G_i = (V_i, E_i)\}_{i \in I}$ be an indexed family of graphs: that is, we have a set I and for every $i \in I$ we have a graph $G_i = (V_i, E_i)$. We define the graph

$$G = \coprod_{i \in I} G_i$$

to be the graph with vertex set

$$V = \coprod_{i \in I} V_i,$$

i.e., the disjoint union of the vertex sets V_i. Formally, we put $V \coloneqq \bigcup_{i \in I} V_i \times \{i\}$ as a subset of $(\bigcup_{i \in I} V_i) \times I$. We define the edge set to be

$$E \coloneqq \bigcup_{i \in I} \{\{(x_i, i),\ (y_i, i)\} \mid i \in I \text{ and } (x_i, y_i) \in E_i\}.$$

In other words, every edge in E runs between two vertices in the same subset V_i of the disjoint union, and the edges running between vertices in $V_i \times \{i\}$ are precisely the same as the edges of V_i. This is a good way of producing disconnected graphs: in general, every graph is the coproduct of its connected components.

Now let T be a finite tree, let C_n be an n-cycle, and let $G \coloneqq T \coprod C_n$: that is, to T we add n more vertices that we connect to each other in a cycle, and we add no further edges. Then G is not a tree: it has two connected components, T and C_n, but

$$\#V(G) = \#V(T) + n,\ \#E(G) = \#V(T) + n,$$

so

$$\#V(G) - \#E(G) = (\#V(T) + n) - (\#E(T) + n) = \#V(T) - \#(T) = 1.$$

An edge e of a connected graph $G = (V, E)$ is called a bridge if removing it gives a disconnected graph.

LEMMA 9.39. *For a graph $G = (V, E)$ and an edge $e \in e$, the following are equivalent:*

(i) *There is a cycle in G containing e.*
(ii) *The edge e is* not *a bridge.*

PROOF. Let $G' \coloneqq (V, E \setminus \{e\})$ be the graph obtained by removing the edge e. (i) $\implies$ (ii): If $e = \{v, w\}$ is part of a cycle, then the rest of the cycle still gives a path P from v to w in G'. Thus in any walk between two vertices of G, we can

replace the edge e with the path P and get a walk from v to w in G'. Thus G' is still connected, so e is not a bridge.
(ii) $\implies$ (i): If $e = v, w$ is not a bridge, then G' is connected, so there is a path

$$w = a_0, a_1, \dots, a_n = v$$

from w to v in G'. Snce we removed the edge e, which is the unique path of length 1 from w to v, we must have $n \geq 2$. Thus

$$v, w = a_0, a_1, \dots, a_n = v$$

is a cycle in G containing e. □

LEMMA 9.40. *If $T = (V, E)$ is a tree, then for all $e \in E$, the graph $G' \coloneqq (V, E \setminus \{e\})$ has two connected components. In particular, every edge of a tree is a bridge.*

PROOF. Let $e = \{v, w\}$. Because T is a tree, the directed edge v, w is the unique path from v to w, so there is no path from v to w in G'. Thus v and w lie in different connected components of G', so e is a bridge in T.

It remains to show that for every $x \in V$, there is either a path from x to v in G' or a path from x to w in G': this will give that $\mathfrak{c}(v)$ and $\mathfrak{c}(w)$ are the two connected components of G'. To see this, let P be the unique path from x to v in T. If this path does not contain the edge e, then P is still a path from x to v in G'. If P does contain the edge e, then P must be of the form

$$x = a_0, \dots, a_{n-2}, w, v.$$

(The edge cannot occur anywhere else in the path than at the end, because otherwise the vertex v would appear twice in the path.) Then

$$x = a_0, \dots, a_{n-2}, w$$

is a path from x to w in G'. □

If we add an edge e to a tree $T = (V, E)$, we get a graph $G' = (V, E \cup \{e\})$ in which the edge e is not a bridge, so by Lemma 9.39, there is a cycle in G' containing e and thus G' is no longer a tree. Thus the answer to question 1b) is **no**: if we add even a single edge to a tree, we get a cycle (so still more do we get a cycle if we add more than one edge). Lemma 9.40 shows that the answer to question 1c) is also **no**: if we remove even a single edge from a tree , then we get a disconnected graph that is not a tree (so still more do we get a disconnected graph if we remove more than one edge).

These results have some interesting consequences:

COROLLARY 9.41. *Let $G = (V, E)$ be a finite connected graph. Then:*

a) *We have $\chi(G) \leq 1$, with equality if and only if G is a tree.*
b) *Write $\chi(G) = 1 - c$ with $c \in \mathbb{N}$. Then there is a subset $E' \subseteq E$ with $\#E \setminus E' = c$ such that $G' \coloneqq (V, E')$ is a tree.*
c) *If $e = \{v, w\}$ is a bridge in G, then $G' \coloneqq (V, E \setminus \{e\})$ has two connected components, $\mathfrak{c}(v)$ and $\mathfrak{c}(w)$.*

PROOF. The basic idea behind both parts is this: if G is a tree, then by Theorem 9.37 we have $\chi(G) = 1$. Otherwise G is connected and not a tree, so G contains a cycle, and by Lemma 9.39 we can remove any edge of this cycle to get

a connected graph G_1 with $\chi(G_1) = \chi(G) + 1$. If G_1 is a tree, then $\chi(G_1) = 1$ and thus $\chi(G) = 1 - 1$. If not, then as above we can remove an edge from a cycle in G_1 to get a graph G_2 with $\chi(G_2) = \chi(G_1) + 1 = \chi(G) + 2$. This process cannot continue indefinitely because we have only finitely many edges, so there must be some $c \in \mathbb{N}$ such that after removing precisely c edges we get a graph G' that is a tree. Thus

$$\chi(G) = \chi(G') - c = 1 - c.$$

c) By definition of a bridge, removing e disconnects G. So there fcannot be a path P from v to w in G', because then in any path in G we could replace e with P and get a path in G', contradicting the disconnectedness of G'. So $\mathfrak{c}(v) \neq \mathfrak{c}(w)$, and we need to show that these are the only connected components. But we showed this in Lemma 9.40 when G is a tree. If G' is not a tree, then by part b) we can remove some edges to get a tree T', and *then* when we remove $e = \{v, w\}$, every vertex in the resulting graph T'' is connected to either v or w by a path in T'', hence certainly every vertex in G' is connected to either v or to w by a path in G'. □

For a connected graph $G = (V, E)$ a subset $E' \subseteq E$ such that $G' \coloneqq (V, E')$ is called a **spanning tree** of G. Corollary 9.41 shows that every finite connected graph admits a spanning tree. In fact every infinite connected spanning graph admits a spanning tree too, but this requires transfinite methods: see e.g. [**MSE**].

It is interesting to ask for the number of spanning trees in a finite connected graph. We will take only the first step in this direction. For $n \in \mathbb{Z}^+$, the **complete graph** K_n is the graph with vertex set $[n]$ and edge set

$$\{Y \subseteq [n] \mid \#Y = 2\},$$

the set of all 2 element subsets of $[n]$: thus there are $\binom{n}{2}$ edges overall. It is called "complete" because it has all possible edges for a graph with vertex set $[n]$. Clearly K_n is connected: indeed every pair of distinct vertices are connected by an edge. The graph K_1 has no edges; the graph K_2 is a path of length 2; the graph K_3 is a 3-cycle. For $n \geq 4$, K_n properly contains an n-cycle, so is not a tree.

A spanning tree in K_n is just a tree T with vertex set $[n]$. For $n \in \mathbb{Z}^+$, we define the quantity $\mathcal{T}_n$ to be the number of trees with vertex set $[n]$, a.k.a. the number of spanning trees in K_n.

Example 9.42.

a) *Since K_1 and K_2 are themselves trees, they have unique spanning trees: $\mathcal{T}_1 = \mathcal{T}_2 = 1$.*

b) *The graph K_3 is a 3-cycle, with Euler characteristic 0. So we must remove 1 edge to get a tree, and it is clear (e.g. by symmetry) that removing any one edge will give us a spanning tree. Since there are three edges in all, we get $\mathcal{T}_3 = 3$.*

c) *The graph K_4 has*

$$\chi(K_4) = \#V(K_4) - \#E(K_4) = 4 - \binom{4}{2} = -2 = 1 - 3.$$

Thus we need to remove 3 edges from K_4 to get a spanning tree. There are $\binom{6}{3} = 20$ three element subsets of the set $E(K_4)$ of edges. If S is a 3-element subset of the vertex set, we can delete three edge so as to get a 3-cycle with vertex set S and one more isolated vertex: this is a way of

removing 3 *edges that* does not *yield a tree. Since there are* $\binom{4}{3} = 4$ *such sets* S*, this gives* 4 *three element subsets of* E_4 *that* do not *yield spanning trees. For any* 3 *element subset* T *of* $E(K_4)$*, the vertices comprising the edges of* T *cannot be pairwise disjoint, because that would yield six different vertices, whereas we only have* 4*. So we must have edges* $e_1 = \{x, y\}$ *and* $e_2 = \{y, z\}$ *in* T *for some vertices* x, y, z*. If the third edge of* T *is* $\{z, x\}$ *then the edges of* T *form a* 3*-cycle, which as above happens* 4 *times and does not yield a tree. In every other case the remaining edge* e_3 *must contain the unique vertex* $w \in [4] \setminus \{x, y, z\}$*; thus the fourth vertex is connected to the other three and* $([4], T)$ *is a tree. Conclusion:*

$$\mathcal{T}_4 = 20 - 4 = 16.$$

d) *The graph* K_5 *has*

$$\chi(K_5) = \#V(K_5) - \#E(K_5) = 5 - \binom{5}{2} = -5 = 1 - 6,$$

so to get a spanning tree we need to remove 6 *edges from our set of* 10*, which we can do in* $\binom{10}{6} = 210$ *possible ways. Figuring out how many of these yield trees seems to require either a rather lengthy consideration of cases or a significant new idea. It turns out that* $\mathcal{T}_5 = 125$*.*

The preceding considerations ought to give us some appreciation for the following remarkable result.

THEOREM 9.43 (Cayley). *For all* $n \in \mathbb{Z}^+$*, the number of spanning trees of the complete graph* K_n *is* $\mathcal{T}_n = n^{n-2}$*.*

PROOF. By Example 9.42, may assume $n \geq 3$. Since n^{n-2} is rather evidently the cardinality of a certain finite set – namely, the set of all functions from a set with $n-2$ elements to a set with n elements, or equivalently, the set of all finite lists of length $n-2$ with entries in $[n] = \{1, \ldots, n\}$ – a good general idea is to show this by finding a bijection from the set of trees on $[n]$ to the set of finite lists of length $n-2$ on $[n]$. Indeed this can – and will – be done, but it is really not obvious how to proceed. We follow a procedure due to Prüfer.

Step 1: Let $T = ([n], E)$ be a tree. We build the **Prüfer code** $C(T) \in [n]^{n-2}$ as follows: Since T is a finite tree with $n \geq 3$ vertices, it has at least two pendant vertices. The first element of the Prüfer code is the unique vertex that is adjacent to the *minimal* pendant vertex. Here minimal just means that every vertex is a number between 1 and n, and we just choose the smallest among the pendant vertices. Moreover, precisely because the vertex is pendant, it is adjacent to a unique vertex, so it makes sense to write down that vertex: it is a number $x_1 \in [n]$.

Then we prune the tree by removing the minimal pendant vertex and the unique edge containing it. This leaves us with a tree T_2 with $n-1$ vertices, which are the elements of $[n]$. If $n = 3$ then T_2 is a tree with two vertices – i.e., an edge – and we stop: the Prüfer code is simply $C(T) = (x_1)$. Otherwise we continue as before: we identify the minimal pendant vertex of T_2, and we take x_2 to be the unique vertex adjacent to it. We continue in this manner, pruning the tree $n-2$ times, until we get a tree T_{n-2} with exactly two vertices, and then we stop. The Prüfer code is

$$T(C) = (x_1, \ldots, x_{n-2}).$$

Step 2: Thus we have defined a map $C : \mathcal{T}_n \to [n]^{n-2}$, and we need to show that it is a bijection. As usual, the best way to show that a map is a bijection is to find the inverse function. That is, we need to find a **decoding function**

$$D : [n]^{n-2} \to \mathcal{T}_n$$

that assigns to each sequence $(x_1, \ldots, x_{n-2})$ with elements in $[n]$ a tree on $[n]$, in such a way that for all trees T on $[n]$ we have $D(C(T)) = T$ and for all sequences $(x_1, \ldots, x_{n-2}) \in [n]^{n-2}$ we have $C(D(x_1, \ldots, x_{n-2})) = (x_1, \ldots, x_{n-2})$. In order to do this gracefully we begin with an observation about the Prüfer code:
I. In the tree T, let d_i be the degree of the ith vertex. Then the number of times that i occurs in the finite list is $d_i - 1$.

Indeed, in a finite tree with at least three vertices, no vertex adjacent to a pendant vertex is also pendant (if so, the tree would consist of a single edge), so every pendant vertex i appears $0 = 1 - 1 = d_i - 1$ time in the Prüfer code. For the rest, we can go by induction on the number of vertices: a tree with 3 vertices has exactly one non-pendant vertex, and its Prüfer code consists precisely of that nonpendant vertex. Assuming the claim holds for all trees with $n \geq 3$ vertices, if we take a tree with $n + 1$ vertices and prune it, then we remove the minimal pendant vertex, write down its adjacent vertex x_1 in the Prüfer code, and then remove the corresponding edge, which leaves a tree with n vertices in which the degree of x_1 has been decreased by 1 and none of the degrees of the other nonpendant vertices have changed. So the result follows by induction.

Now we define the decoding function: let $\ell_{n-2} : (x_1, \ldots, x_{n-2})$ be a list of length $n - 2$ with entries in $[n]$. Put $S_n := [n]$. We start with the edgeless graph on $[n]$ and in each of $n - 1$ stage we add an edge; at the end we will get a tree.
Stage 1: We add the edge $e_1 = \{x_1, y_n\}$, where y_n is the least element of $[n] \setminus \{x_1, \ldots, x_{n-2}\}$. Also we put $\ell_{n-3} : (x_2, \ldots, x_{n-2})$ and $S_{n-1} := [n] \setminus \{y_n\}$.
Stage 2: We add the edge $e_2 = \{x_2, y_{n-1}\}$, where y_{n-1} is the least element of $S_{n-1} \setminus \{x_2, \ldots, x_{n-2}\}$.
Also we put $\ell_{n-4} : (x_3, \ldots, x_{n-2})$ and $S_{n-2} := S_{n-1} \setminus \{y_{n-1}\} = [n] \setminus \{y_{n-1}, y_n\}$.
$\vdots$
Stage $n - 2$: We add the edge $e_{n-2} = \{x_{n-2}, y_3\}$, where y_3 is the least element of $S_3 \setminus \{x_{n-2}\}$. We are left with $S_2 = S_3 \setminus \{y_3\} = [n] \setminus \{y_3, \ldots, y_n\}$.
Stage $n-1$: S_2 consists of two elements y_1 and y_2, and add the edge $e_{n-1} = \{y_1, y_2\}$.
Now we can see that for all $2 \leq k \leq n$ we have $S_k = \{y_1, y_2, \ldots, y_k\}$, a subset of $[n]$ of size k. Finally, we put

$$D(x_1, \ldots, x_{n-2}) = ([n], \{e_1, \ldots, e_{n-1}\}).$$

Our first order of business is to show that $D(x_1, \ldots, x_{n-2})$ is a tree. We'll show by induction on k that for all $2 \leq k \leq n$ the graph $(S_k, \{e_k, \ldots, e_{n-1}\})$ is a tree: taking $k = n$ gives the desired conclusion. In the base case $k = 2$ we have the edge $\{y_{n-1}, y_{n-2}\}$ on $S_2 = \{y_{n-1}, y_n\}$, so this is certainly a tree. Let $2 \leq k \leq n - 1$ and assume that $G_k := (S_k, \{e_{n-k+1}, \ldots, e_{n-1}\})$ is a tree. Then the graph G_{k+1} is G_k together with the additional vertex y_{k+1} and the additional edge $e_{k+1} = (x_{n-k}, y_{k+1})$. If we start with a tree, add a new vertex and connect it to any edge of the tree, we get a new tree. So by induction $G_n = ([n], \{e_1, \ldots, e_{n-1}\})$ is a tree.
Step 3: It remains to show that C and D are inverse functions.

To see that for any tree T on $[n]$ we have $D(C(T)) = T$, we observe that the

edges $e_1, \ldots, e_{n-1}$ generated by the decoding procedure are precisely the edges we remove as we prune the tree in the encoding procedure.

Similarly, let $\ell : (x_1, \ldots, x_{n-2}) \in [n]^{n-2}$. We first observe that in the tree $T(\ell)$, the degree of the vertex $i \in [n]$ is one more than the number of times that i appears in the sequence ℓ, since we get one edge containing i for each instance of i in the sequence ℓ together with one more edge containing i because each $i \in [n]$ is y_k for a unique $k \in [n]$. Using this observation, you can (I hope) convince yourself that the edges $e_1, \ldots, e_{n-1}$ generated in the decoding procedure are precisely the edges we remove as we prune the tree $D(\ell)$ using the encoding procedure. □

3.3. Ramsey's Theorem. In this section, by a "graph" $G = (V, E)$ we will always mean a simple, undirected graph. If G is moreover finite, then we have $\#V = n$ for some $n \in \mathbb{N}$. (The case of $n = 0$, i.e., no vertices, is allowed – however, there is certainly nothing going on there.) In this case we gain in concreteness and lose nothing in return by assuming that $V = [n]$. A **clique** in a graph G is a subset $C \subseteq V$ of vertices such that for all $x, y \in C$, if $x \neq y$ then $x \sim y$: that is, any pair of distinct vertices in the clique are adjacent. An **independent set** in G is a subset $I \subseteq V$ of vertices such that for all $x, y \in I$, we *do not* have $x \sim y$: that is, no pair of vertices in I are adjacent.

For any graph $G = (V, E)$ we define its **complement** $\overline{G} = (V, \overline{E})$. That is, the vertex set of $\overline{G}$ is the same as the vertex set of G, but for distinct $x, y \in V$, we put

$$\{x, y\} \in \overline{E} \iff \{x, y\} \notin E.$$

That is, the edge set $\overline{E}$ is the complement of the edge set E inside the set $\binom{V}{2}$ of 2-element subsets of V. An immediate – but important – observation is that for all subsets $S \subseteq V$, we have that S is a clique in G if and only if S is an independent set in $\overline{G}$, and similarly S is an indepenent set in G if and only if S is a clique in $\overline{G}$.

PROPOSITION 9.44. *Let $G = (V, E)$ be a finite graph with $\#V \geq 6$. Then at least one of the following holds:*

a) *There is a 3-element clique C in G.*
b) *There is a 3-element independent set I in G.*

PROOF. We claim that this result is an equivalent reformulation of Proposition 6.14. First of all, we may as well assume that $V = [6]$, because if V has more than six vertices we can just look at edges running between any six of the vertices to get the desired conclusion. Now if $G = ([6], E)$ is a graph, then we can view it as a model of six people at a party, where for $1 \leq i \neq j \leq 6$, we say that person P_i knows person P_j if and only if $i \sim j$. (And conversely: given six people at a party, their mutual knowing / not knowing each other determines a graph with vertex set [6].) So Proposition 6.14 applies to tell us that either there are three people who all know each other – in which case the corresponding 3-element subset of [6] forms a clique – or there are three people none of whom know each other – in which case the corresponding 3-element subset of [6] forms an independent subset. □

In this reformulation, Exercise 6.14 asks you to construct a graph with vertex set [5] for which there is neither a 3-element clique nor a 3-element independent set, so 6 is the minimum number of vertices for this conclusion to hold.

This graph-theoretic reformulation of Proposition 6.14 suggests a vast generalization: for any integers $a, b \geq 1$ we can ask whether it is true that for sufficiently large n, any graph with n vertices must have an a-element clique or a b-element independent subset, and if so, what is the smallest $n = \#V$ that ensures this. Let us define the **Ramsey number** $R(a, b)$ to be this minimum n if such an n exists, and ∞ otherwise. We get a very succinct reformulation of Proposition 6.14, namely:

$$R(3, 3) = 6. \tag{57}$$

For $n \in \mathbb{Z}^+$, we have $R(a, b) \leq n$ if and only if every graph $G = ([n], E)$ has either an a-element clique or a b-element independent set: if this holds for all graphs with n vertices, then it certainly holds for graphs with more than n vertices because we can just consider any n of the vertices of a larger graph.

In the following results we will allow ourselves to use ∞ in inequalities. This goes as follows: for all $x \in \mathbb{R}$ we decree that $x \leq \infty$ is true, as is $\infty \leq \infty$, while for all $x \in \mathbb{R}$ we decree that $\infty \leq x$ is false.

The next few results are *less interesting* than Proposition 6.14, but we include them so as to be systematic.

PROPOSITION 9.45. *For all $a, b \in \mathbb{Z}^{\geq 2}$ we have $R(a, b) \geq \max(a, b)$.*

PROOF. Let $G = ([n], E)$ be a finite graph. If $C \subseteq [n]$ is a clique, then $\#C \leq n$, to have an a element clique we need at least n vertices. Similarly, if $I \subseteq [n]$ is an independent set, then $\#I \leq n$, so to have a b element independent set we need at least n vertices. □

PROPOSITION 9.46. *Let $a, b \in \mathbb{Z}^+$.*

a) *We have $R(a, 1) = R(1, b) = 1$.*
b) *We have $R(a, 2) = a$.*
c) *We have $R(2, b) = b$.*

PROOF. a) If v is a vertex of a graph, then the singleton set $\{v\}$ is, trivially, both a 1-element clique and a 1-element independent set, so $R(a, 1) = R(1, b) = 1$.
b) By Proposition 9.45 we have $R(a, 2) \geq a$. Suppose that $G = ([n], E)$ is a finite graph on at least a vertices. The only way for it not to have a 2-element independent set is for all pairs of distinct vertices to be adjacent (we say that G is the **complete graph** on n vertices), in which case it has an a-element clique.
b) By Proposition 9.45 we have $R(2, b) \geq b$. Suppose that $G = ([n], E)$ is a finite graph on at least b vertices. The only way for it not to have a 2-element clique is for no vertices to be adjacent whatsoever (we say that G is the **empty graph** on n vertices), in which case it has a b-element independent set. □

PROPOSITION 9.47. *For all $a, b \in \mathbb{Z}^+$ we have $R(a, b) = R(b, a)$.*

PROOF. In other words, we must prove that for $n \in \mathbb{Z}^+$, if every graph on n vertices has either an a-element clique or a b-element independent set, then every graph on n vertices has either a b-element clique or an a-element independent set, and conversely.

Suppose that every graph on n vertices has either an a-element clique or a b-element independent set, and let $G = ([n], E)$ be a graph. Then its complement

$\overline{G} = ([n], \overline{E})$ has either an a-element clique or a b-element independent set, so G has either a b-element clique or an a-element independent set.

The assertion that if every graph on n vertices has either a b-element clique or an a-element independent set then every graph on n vertices has either an a-element clique or a b-element independent set follows from the previous paragraph by interchanging a and b. □

THEOREM 9.48. *Let $a, b \in \mathbb{Z}^{\geq 2}$.*

a) *(Greenwood-Gleason) We have*

$$R(a,b) \leq R(a-1,b) + R(a,b-1). \tag{58}$$

b) *(Ramsey) It follows that $R(a,b) < \infty$.*

PROOF. a) Let $n := R(a-1,b) + R(a,b-1)$. It suffices to show that every graph $G = ([n], E)$ has an a-element clique or a b-element independent set. Let

$$S := \{2 \leq i \leq n \mid \ 1 \text{ is adjacent to } i\}$$

and

$$T := \{2 \leq j \leq n \mid \ 1 \text{ is not adjacent to } j\},$$

so $S \coprod T = \{2, \ldots, n\}$. It follows that either $\#S \geq R(a-1,b)$ or $\#T \geq R(a,b-1)$: for if not we have $\#S \leq R(a-1,b) - 1$ and $\#T \leq R(a,b-1) - 1$, so

$$n - 1 = \#S + \#T \leq R(a-1,b) + R(a,b-1) - 2 = n - 2,$$

a contradiction.

Case 1: If $\#S \geq R(a-1,b)$, then among the vertices in S we have a clique C of size $a-1$ or an independent set I of size b. In the latter case we are done. In the former case, since S consists of vertices adjacent to 1, $C \cup \{1\}$ is a clique of size a.

Case 2: If $T \geq R(a,b-1)$ then we have a clique C of size a or an independent set I of size $b-1$. In the former case we are done. In the latter case, since T consists of vertices *not* adjacent to 1, $I \cup \{1\}$ is an independent set of size b.

b) We go by induction on $N := a + b$.

BASE CASE: Among $a, b \in \mathbb{Z}^{\geq 3}$, we have $N = 6$ if and only if $a = b = 3$, and we know that $R(3,3) = 6$.[4]

INDUCTIVE STEP: Suppose that for some $N \geq 3$ and all $a, b \in \mathbb{Z}^{\geq 3}$ with $a + b = N$ we have $R(a,b) < \infty$. In light of Proposition 9.46 we know that $R(a,2)$ and $R(2,b)$ are finite, so we may assume that $R(a,b) < \infty$ for all $a, b \in \mathbb{Z}^{\geq 2}$ such that $a + b = N$. Now suppose that $a, b \in \mathbb{Z}^{\geq 3}$ are such that $a + b = N + 1$. Then

$$(a-1) + b = a + (b-1) = N,$$

so we know that $R(a-1,b)$ and $R(a,b-1)$ are both finite, so

$$R(a,b) \leq R(a-1,b) + R(a,b-1)$$

is also finite. □

In Exercise 9.11 you are asked to show that for all $a, b \in \mathbb{Z}^+$ we have

$$R(a,b) \leq \binom{a+b-2}{a-1}.$$

[4]In fact, using (58) and Proposition 9.46 we get $R(3,3) \leq R(2,3) + R(3,2) = 3 + 3 = 6$.

It is a very difficult problem to determine the Ramsey numbers $R(a,b)$ for $a,b \geq 3$. In fact (counting $R(a,b) = R(b,a)$ as one value), only 9 values are known! We handled one of them: $R(3,3) = 6$. The others are recorded in the following result.

THEOREM 9.49 (Known Ramsey Numbers).

a) *(Greenwood-Gleason* [**GG55**]*) We have* $R(3,4) = R(4,3) = 9$.
b) *(Greenwood-Gleason* [**GG55**]*) We have* $R(3,5) = R(5,3) = 14$.
c) *(Kéry* [**Ke64**]*) We have* $R(3,6) = R(6,3) = 18$.
d) *(Graver-Yackel* [**GY68**]*) We have* $R(3,7) = R(7,3) = 23$.
e) *(McKay-Min* [**MM92**]*) We have* $R(3,8) = R(8,3) = 28$.
f) *(Grinstead-Roberts* [**GR82**]*) We have* $R(3,9) = R(9,3) = 36$.
g) *(Greenwood-Gleason* [**GG55**]*) We have* $R(4,4) = 18$.
h) *(McKay-Radziszowski* [**MR95**]*) We have* $R(4,5) = R(5,4) = 25$.

Exercises 9.12, 9.13 and 9.14 together show that $R(3,4) = 9$ and $R(4,4) \leq 18$. (To be sure, these exercises are somewhere between challenging and unfair.) The other values are too difficult to treat here.

3.4. More Ramsey and Schur. Recall that in §6.5.2 we showed that no matter how we color each positive integer either red or blue, there are always integers $1 \leq a \leq b \leq c \leq 5$ such that $c = a + b$ and a, b, c are either all red or all blue; we also showed that 5 is the least positive integer for which this holds. Note that to establish this result we evidently don't need to consider 2-colorings of all of $\mathbb{Z}^+$: it is enough to consider 2-colorings of $[5] = \{1,2,3,4,5\}$. There are $2^5 = 32$ such colorings, so we didn't need any real "technique" to verify this: we could have just checked all 32 possible cases. (Of course what we did was easier and better!)

We can extend the notion of a 2-coloring to more colors. We can define for instance a 3-coloring of a set X to be a function

$$\mathfrak{c} : X \to \{\text{red}, \text{blue}, \text{yellow}\}.$$

We can now ask the same problem: is there an $N \in \mathbb{Z}^+$ such that for every 3-coloring $\mathfrak{c}$ of $[N] = \{1,2,\ldots,N\}$ there is a monochromatic Schur triple (a,b,c) with $c \leq N$?

It turns out the answer is **yes**:

PROPOSITION 9.50. *The least* $N \in \mathbb{Z}^+$ *such that every* 3*-coloring of* $[N]$ *has a monochromatic Schur triple* (a,b,c) *with* $c \leq N$ *is* $N = 14$.

We will not prove Proposition 9.50 here. Let us observe that a computer would have no trouble verifying this, since the number of 3-colorings of $[14] = \{1,2,\ldots,14\}$ is $3^{14} = 4782969 \approx 4.78 \times 10^6$. This is small enough for a computer to range over all of them and find the monochromatic Schur triple by brute force. Similarly, a computer would have no trouble looking through the 3-colorings of [13] until it finds one for which there is no monochromatic Schur triple.

More generally, for a positive integer k, let us define a **k-coloring** of a set X to be a function

$$\mathfrak{c} : X \to [k] = \{1,2,\ldots,k\}.$$

Notice that here instead of using actual colors like red, blue and yellow, we are identifying the k different colors with the numbers $1, 2, \ldots, k$. Let us pause for a moment to observe that this is completely permissible in our mathematical context – all we care about the colors is that they are k distinct mathematical objects – and how ridiculous it would be to identify colors with positive integers in virtually any non-mathematical context. Then there is a beautiful general theorem here due to I. Schur [**Sch16**].

THEOREM 9.51 (Schur, 1916). *Let $k \in \mathbb{Z}^+$. There a positive integer N such that for any k-coloring $\mathfrak{c}$ of $\mathbb{Z}^+$ there is a monochromatic Schur triple (a, b, c) with $c \leq N$. The least such positive integer N is called the* ***kth Schur number*** *$S(k)$.*

We will deduce Schur's Theorem from a multi-color generalization of Ramsey's Theorem. For $a, b \in \mathbb{Z}^+$, recall that $R(a, b)$ is the least positive integer n such that every graph on n-vertices has either a a-element clique or a b-element independent set. But already in §6.5.2 we gave an alternative "colorful" interpretation: namely, $R(a, b)$ is the least $n \in \mathbb{Z}^+$ such that if we start with the complete graph K_n on n vertices and color each of its vertices either red or blue, then there is either a red clique of size a or a blue clique of size b. (The equivalence is just because we can color an edge red if it is was included in the original graph and color it blue if it was excluded in the original graph.)

This colorful interpretation suggests a generalization: let $k \in \mathbb{Z}^+$ and let $a_1, \ldots, a_k$ be positive integers. We define the Ramsey number $R(a_1, \ldots, a_k)$ to be the least positive integer n so that for any k-coloring of the edge set of the complete graph K_n, then for some $1 \leq i \leq k$ there is an i-clique in which all vertices are colored i. For instance, we can interpret $R(3, 4, 5)$ as the least number of vertices so that if we color every edge of the complete graph red, blue or yellow, then there is either a red 3-clique or a blue 4-clique or a yellow 5-clique...if there is such number.

In Exercise 9.15 you are asked to show that $R(a_1, \ldots, a_k)$ does not depend upon the ordering of $a_1, \ldots, a_k$. Because of this it is no loss of generality to consider only Ramsey numbers $R(a_1, \ldots, a_k)$ with $a_1 \leq \ldots \leq a_k$. In Exercise 9.16 you are asked to evaluate $R(a_1, \ldots, a_k)$ in two easy cases. In particular, Exercise 9.16b) explains why we may restrict to the case $a_1 \geq 3$.

One's first instinct might be to worry that proving the finiteness of the multicolor Ramsey numbers $R(a_1, \ldots, a_k)$ could be much harder than for $k = 2$. Happily, that is not the case:

PROPOSITION 9.52. *Let $k \geq 3$ and let $1 \leq a_1, \ldots, a_k \in \mathbb{Z}^+$. Then:*

a) *We have*

$$R(a_1, \ldots, a_k) \leq R(a_1, \ldots, a_{k-2}, R(a_{k-1}, a_k)). \tag{59}$$

b) *It follows that $R(a_1, \ldots, a_k) < \infty$.*

PROOF. a) Let $N := R(a_1, \ldots, a_{k-2}, R(a_{k-1}, a_k))$ and let $\mathfrak{c}$ be any k-coloring of the edge set of the complete graph K_N on N-vertices. To this coloring we attach a $(k-1)$-coloring $\mathfrak{c}' : V(K_N) \to [k-1]$ just by redefining $\mathfrak{c}$ to take the value $k-1$ whenever it takes the value k. By definition of $R(a_1, \ldots, a_{k-2}, R(a_{k-1}, a_k))$, with respect to the coloring $\mathfrak{c}'$, our graph either contains an i-colored a_i-clique for some

$1 \leq i \leq k-2$ (in which case we're done) or a $(k-1)$-colored $R(a_{k-1}, a_k)$ clique. (By Theorem 9.48b), we know that $R(a_{k_1}, a_k)$ is finite.) But if we know focus in on the complete graph G_{k-1} on the vertices that are $\mathfrak{c}'$-colored $k-1$, the original coloring $\mathfrak{c}$ colors each edge with either color $k-1$ or color k, i.e., is a 2-coloring of G_{k-1}. By definition of $R(a_{k-1}, a_k)$, we either have a $(k-1)$-colored a_{k-1}-clique in G_{k-1} (hence also in K_N) or a k-colored a_k-clique in G_k (hence also in K_N).
b) This follows from part a) and Theorem 9.48b). □

The state of knowledge of exact values of multi-color (i.e., $k \geq 3$) Ramsey numbers is yet more dire than the $k = 2$ case. By Exercise 9.16b) we may restrict to the case $3 \leq a_1 \leq \ldots \leq a_k$. The *only* known values are given in the following result.

THEOREM 9.53.

a) *(Greenwood-Gleason* [**GG55**]*)* $R(3,3,3) = 17$.
b) *(Codish-Frank-Itzkhakov-Miller* [**CFIM16**]*)* $R(3,3,4) = 30$.

Now we will use the finiteness of the multi-color Ramsey numbers to prove Schur's Theorem (Theorem 9.51). In fact this is implied by the following result:

THEOREM 9.54. *For all $k \in \mathbb{Z}^+$, we have*

$$S(k) \leq R(3, \ldots, 3 \text{ } (k \text{ } \textit{times})) - 1.$$

PROOF. It suffices to take

$$N := R(3, \ldots, 3 \text{ } (k \text{ times})) - 1,$$

let $\mathfrak{c}$ be any k-coloring of $[N] = \{1, 2, \ldots, N\}$ and show that there is a monochromatic Schur triple (a, b, c) with $c \leq N$.

We use $\mathfrak{c}$ to define a k-coloring $\mathcal{C}$ on the complete graph on $N+1$ vertices – we represent the vertex set as $[N+1]$ – as follows: for all $1 \leq i < j \leq N+1$, we color the edge (i, j) with the color $\mathfrak{c}(j-i)$. By definition of the Ramsey number $R(3, \ldots, 3$ (k times), are $1 \leq x < y < z \leq N+1$ such that (x, y), (y, z) and (x, z) all have the same color, which menas that $y - x$, $z - y$ and $z - x$ all have the same color. But

$$z - x = (z - y) + (y - x),$$

so $(y - x, z - y, z - x)$ is a monochromatic Schur triple. Moreover, since $1 \leq x, z \leq N+1$, we have $z - x \leq N + 1 - 1 = N$. □

Taking $k = 2$ in Theorem 9.54 we find

$$S(2) \leq R(3,3) - 1 = 6 - 1 = 5.$$

This bound is sharp: by Proposition 6.15 we have $S(2) = 5$. Taking $k = 3$ in Theorem 9.54, we find

$$S(3) \leq R(3,3,3) - 1 = 17 - 1 = 16.$$

This bound is *not* sharp: according to Proposition 9.50 we have $S(3) = 14$.

The next result collects the other known Schur numbers:

THEOREM 9.55.

a) *(Baumert-Golomb 1965)* $S(4) = 45$.
b) *(Heule 2018)* $S(5) = 161$.

The method of proof of Theorem 9.54 extends almost verbatim to studying monochromatic solutions to the slightly more general equation $x_1 + \ldots + x_{r-1} = x_r$. You are asked to give a corresponding bound on the "generalized Schur number" $S_r(k)$ in Exercise 9.54. In this regard we mention one recent success story:

THEOREM 9.56 (Boza-Marín-Revuelta-Sanz [**BMRS19**]). *For all $r \geq 3$ we have*

$$S_r(3) = r^3 - r^2 - r - 1.$$

Taking $r = 3$ in Theorem 9.56 we get $S(3) = S_3(3) = 14$: this is Proposition 9.50.

3.5. Matchings. A **matching** on a graph $G = (V, E)$ is a subset $M \subseteq E$ of edges that are pairwise disjoint: i.e., no two distinct $e_1, e_2 \in M$ share a vertex. A matching M is **perfect** if every vertex of G lies in exactly one edge in M. If V is finite and M is a perfect matching, then there are precisely twice as many vertices as edges, so $\#V$ must be even. However, having $\#V$ be even is not sufficient for the existence of a perfect matching: as an extreme example, an edgeless graph on an even number of vertices admits no perfect matching! Slightly more generally, for a finite graph to have a perfect matching, we clearly need $\#E \geq \frac{\#V}{2}$.

EXAMPLE 9.57.

a) *Let n be an even positive integer, and let P_n be the path on $[n]$ with edges $e_i = \{i, i+1\}$ for all $1 \leq i \leq n-1$. Let M be a perfect matching on P_n. Then M must contain e_1, because every vertex must lie in some edge in M and e_1 is the only edge of P_n containing 1. Since $2 \in e_1$, the perfect matching M cannot contain $e_2 = \{2, 3\}$, but it must contain some edge that contains 3, so it must contain the only other such edge of P_n, namely $e_3 = \{3, 4\}$. Continuing in this way, we see that we must have*

$$M = \{e_1, e_3, \ldots, e_{n-1}\}$$

and that M is indeed a perfect matching. Thus P_n admits a unique perfect matching.

b) *Let n be an even positive integer, and let C_n be the n-cycle: specifically, we take vertex set n and edges $e_i = \{i, i+1\}$ for all $1 \leq i \leq n-1$ and also $e_n = \{n, 1\}$. A similar analysis to the above shows that C_n admits exactly two perfect matchings:*

$$M_1 = \{e_1, e_3, \ldots, e_{n-1}\},\ M_2 = \{e_2, e_4, \ldots, e_n\}.$$

c) *Comparing parts a) and b), we observe that the perfect matching M_1 on C_n is "the same as" the perfect matching M on P_n. To formalize this, if $G_1 = (V, E_1)$ and $G_2 = (V, E_2)$ are two graphs on the same vertex set V, we say that G_1 is an **edge subgraph** of G_2 if $E_1 \subseteq E_2$: that is, G_2 is obtained from G_1 by adding some edges (or, technically, adding no edges: we can have $G_2 = G_1$). For all $n \in \mathbb{Z}^+$, the path P_n is an edge subgraph of C_n. Now we observe that if M is a matching on G_1 and G_1 is an edge subgraph of G_2 then M is also a matching on G_2, and moreover M is a perfect matching on G_1 if and only if M is a perfect matching on G_2. The upshot of this is that adding edges to a graph only makes it easier to have a perfect matching. In particular, for all even $n \in \mathbb{Z}^+$ the complete graph K_n admits a perfect matching.*

Understanding when perfect matchings exist is easier if we change the rules a bit. A **bipartitioned graph** is a triple $G = (V_1, V_2, E)$ in which V_1 and V_2 are disjoint sets and E is a set each element of which is of the form $\{v_1, v_2\}$ with $v_1 \in V_1$ and $v_2 \in V_2$. That is, assuming that V_1 and V_2 are nonempty a bipartioned graph is a partition of the vertex set into two parts such that each edge connects a vertex in the first part to a vertex in the second part. A graph is called **bipartite** if it admits a bipartition. In Exercise 9.20c) you are asked to show that a graph is bipartite if and only if it admits no cycle of odd length. In particular, every tree is bipartite.

On a bipartitioned graph (V_1, V_2, E) a **semiperfect matching** is a matching $M \subseteq E$ such that $V_1 \subseteq \bigcup_{e \in M} e$: i.e., every vertex in V_1 appears in some edge of M. Equivalently, a semiperfect matching is determined by an injection $\iota : V_1 \to V_2$ such that for all $v \in V_1$ we have $v \sim \iota(v)$. We say tha the matching is "semiperfect" because there is no condition that all vertices in V_2 appear in the matching. A semiperfect matching is perfect if and only if ι is a bijection.

If V_2 is finite, the existence of a semiperfect matching implies that V_1 is finite and $\#V_1 \leq \#V_2$. Moreover, if V_1 and V_2 are finite, then a perfect matching can only exist if $\#V_1 = \#V_2$, and when this holds every semiperfect matching is perfect. In particular, there is no essential difference between a perfect matching and a semiperfect matching when V_1 and V_2 are finite of the same size.

EXAMPLE 9.58. a) *Let S_n the the graph on $[n+1]$ in which $1 \sim i$ for all $2 \leq i \leq n+1$. This graph is a tree, and its Prüfer code is $(1, \ldots, 1)$ $(n-1)$ times. Taking $v_\bullet = 1$, the associated bipartition is*

$$V_1 = \{2, \ldots, n+1\},\ V_2 = \{1\}.$$

For no $n \geq 2$ does S_n admit a semiperfect matching. Indeed, every edge in this graph contains 1, so there is no matching in S_n that consists of more than a single edge, hence that covers more than two vertices.

On the other hand, if we took the bipartition $\{\{1\}, \{2, \ldots, n+1\}\}$ obtained by interchanging V_1 and V_2, now taking any one edge $\{1, i\}$ gives a semiperfect matching.

b) *Let $n \in \mathbb{Z}^+$, and let T be the tree with vertex set $[2n]$ given by the Prüfer code $(1, 2, 2, 3, 3, \ldots, n-1, n-1, n)$. This tree consists of the path P_n on $[n]$ together with additional edges $\{1, n+1\}, \{2, n+2\}, \ldots, \{n, 2n\}$. Then*

$$M = \{\{1, n+1\}, \{2, n+2\}, \ldots, \{n, 2n\}\}$$

is a perfect matching.

It is worth thinking about what "goes wrong" with the star in Example 9.58a) and not with the tree in Example 9.58b) that prevents the former bipartitioned graph from having a perfect matching. The problem is that all the vertices $2, \ldots, n+1$ can only be matched with the "central" vertex 1, which can in turn only be matched with one of them. This motivates the following definition.

For a graph $G = (V, E)$ and a subset $X \subseteq V$, we define the **neighborhood**

$$N(X) \coloneqq \{y \in V \mid \{x, y\} \in E \text{ for some } x \in X\}.$$

Elements of $N(X)$ are called **neighbors** of X. But beware: although a vertex cannot be adjacent to itself, when X has more than one element, we can have

elements of $X \cap N(X)$. (E.g. $N(V) \cap V$ consists of all vertices in V of positive degree.)

THEOREM 9.59 (Hall's Marriage Theorem). *Let $G = (V_1, V_2, E)$ be a bipartitioned graph in which every vertex in V_1 has finite degree. The following are equivalent:*

(i) *There is a semiperfect matching $\iota : V_1 \hookrightarrow V_2$.*
(ii) *For all finite subsets $X \subseteq V_1$ we have the **Hall condition** $\#X \leq \#N(X)$.*

PROOF. (i) $\implies$ (ii): As above, a semiperfect matching is given by an injection $\iota : V_1 \to V_2$ such that $x \sim \iota(x)$ for all $x \in V_1$. Suppose we have such an injection. Then for any subet $X \subseteq V_1$ the restriction $\iota|_X$ of ι to X remains an injection, and since for all $x \in X$ we have $\iota(x) \in N(X)$, we have

$$\iota|_X : X \hookrightarrow N(X).$$

Therefore $\#X \leq \#N(X)$.
(ii) $\implies$ (i): Here we will prove the result in the case that V_1 is finite. Exercise 9.24 treats a proof of the infinite case due to Halmos-Vaughan [**HV50**] that uses Tychonoff's Theorem from general topology.

We go by strong induction on $\#V_1$. The case $\#V_1$ is absolutely trivial: the empty set is a semiperfect matching. If $\#V_1 = 1$, then V_1 consists of a single vertex v, and applying the condition with $X = \{v\}$, that vertex is adjacent to some vertex in V_2, so we get our semiperfect matching. Now suppose that $\#V_1 = n > 1$ and that the implication holds for all bipartitioned graphs in which the first vertex set has fewer than n vertices. Without loss of generality, we may assume $V_1 = [n]$.
Case 1: Suppose that for all $1 \leq k < n$, each k-element subset of V_1 has at least $k+1$ neighbors. Above we saw Hall's condition implies that every vertex in V_1 is adjacent to at least one element in V_2, so choose some $y \in V_2$ that is adjacent to n and put $\iota(n) \coloneqq y$. Now remove y and all edges containing it. By our Case 1 assumption, in this new graph, every k-element subset of $\{1, \ldots, n-1\}$ still has at least k neighbors, so by induction there is an injective function $\iota : \{1, \ldots, n-1\} \to V_2 \setminus \{y\}$ such that $k \sim \iota(k)$ for all $1 \leq k \leq n-1$, and are done.
Case 2: Otherwise, there is some $1 \leq k < n$ and a k-element subset $X \subseteq [n]$ with $\#N(X) = k$. By induction there is a semiperfect matching $\iota_1 : X \hookrightarrow V_2$, so it suffices to show that the Hall condition still holds on the bipartitioned graph

$$G' \coloneqq ([n] \setminus X, V_2 \setminus \iota(X), E \setminus \{\{x, \iota_1(x)\} \mid x \in X\}).$$

If this were not the case, then for some $1 \leq h \leq n-k$ there would be an h-element subset $Y \subseteq V_1 \setminus X$ such that

$$\#(N(Y) \setminus \iota_1(X)) < h.$$

But then we would have

$$\#N(X \cup Y) = \#\left(N(X) \cup N(Y \setminus \iota_1(X))\right)$$
$$\leq \#N(X) + \#(N(Y) \setminus \iota_1(X)) < k + h = \#(X \cup Y),$$

contradicting the Hall condition. □

In any bipartitioned graph $G = (V_1, V_2, E)$, if $\iota : V_1 \hookrightarrow V_2$ is a semiperfect matching then for *all* subsets $X \subseteq V_1$, the restriction of ι to X induces an injection $X \hookrightarrow N(X)$. Exercise 9.25 exhibits a graph (that is not locally finite!) for which this condition holds but for which there is no semiperfect matching.

In the setting of Theorem 9.59, it is easier to check the Hall Condition than to construct a semiperfect matching, but even so the Hall Condition can be difficult or time-consuming to check. There is however one useful case in which we can check the Hall Condition once and for all: recall that for $d \in \mathbb{N}$ a graph is d-regular if each vertex has degree d. If $G = (V_1, V_2, E)$ is a bipartitioned graph and d_1, d_2 are positive integers, we say that G is (d_1, d_2)**-biregular** if every vertex in V_1 has degree d_1 and every vertex in V_2 has degree d_2.

COROLLARY 9.60. *Let $G = (V_1, V_2, E)$ be a finite bipartitioned graph that is (d_1, d_2)-biregular for some $d_1, d_2 \in \mathbb{Z}$.*

a) *If $d_1 \geq d_2$, then there is a semiperfect matching $\iota : V_1 \hookrightarrow V_2$.*
b) *If $d_2 \geq d_1$, then there is a semiperfect matching $\iota : V_2 \hookrightarrow V_1$.*
c) *If $d_1 = d_2$, there is a perfect matching.*

PROOF. a) Let $X \subseteq V_1$. Then the number of edges between X and $N(X)$ is $d_1 \cdot \#X$, which is also, of course, the number of edges between $N(X)$ and X. Every edge that runs between $N(X)$ and X is in particular an edge that runs between $N(X)$ and V_1 (but not necessarily conversely; vertices in $N(X)$ may be adjacent to vertices lying outside of X), so $d_1 \cdot \#X$ is less than or equal to the number of edges between $N(X)$ and V_1, which is $d_2 \cdot \#N(X)$. So we get:

$$d_1 \cdot \#X \leq d_2 \cdot \#N(X), \tag{60}$$

and thus

$$\frac{\#N(X)}{\#X} \geq \frac{d_1}{d_2} \geq 1,$$

so $\#N(X) \geq \#X$. Thus Hall's Criterion applies, so by Theorem 9.59 we have a semiperfect matching $\iota : V_1 \to V_2$.
b) This follows by applying part a) to the bipartitioned graph $G^T := (V_2, V_1, E)$.
c) Let us write $d = d_1 = d_2$. This is actually a special case of an upcoming result, Theorem 9.63, but since G is finite the proof is much easier. Indeed, applying (60) to $X = V_1$ we get $d\#V_1 \leq d\#V_2$, so $\#V_1 \leq \#V_2$. This same bound applied to $G^T = (V_2, V_1, E)$ gives $\#V_2 \leq \#V_1$, so $\#V_1 = \#V_2$. Thus any semiperfect matching $\iota : V_1 \hookrightarrow V_2$ is an injection between two finite sets of the same cardinality hence is a bijection, hence a perfect matching. □

THEOREM 9.61 (König-Hall Theorem). *Let V be a set, and let $\mathcal{S} = \{S_i\}_{i \in I}$ be a family of finite subsets of V. The following are equivalent:*

(i) *(Hall Condition) For every subset $J \subseteq I$, we have $\#J \leq \#\bigcup_{i \in J} S_i$.*
(ii) *The pair (V, I) admits a **transversal**: that is, there is a subset $X \subseteq V$ and a bijection $f : X \to I$ such that for all $x \in X$ we have $x \in S_{f(x)}$.*

PROOF. (i) $\implies$ (ii): We put $V_1 := I$, $V_2 := V$, and we take

$$E := \{\{i, x\} \mid i \in I \text{ and } x \in S_i\}.$$

Then $G = (V_1, V_2, E)$ is a bipartitioned graph in which each vertex in V_1 has finite degree. For any finite subset $J \subseteq V_1 = I$, the set $N(J)$ consists of all $x \in V$ such that x lies in S_i for some $i \in J$. In other words, we have

$$N(J) = \bigcup_{i \in J} S_i$$

and therefore our assumption (i) that $\#J \leq N(J)$ is indeed the Hall Condition (ii) of Theorem 9.59. So by that result there is a semiperfect matching $\iota : I \to V$, i.e., an injection such that for all $i \in I$ we have $\iota(i) \in S_i$. Put $X \coloneqq \iota(I)$. Then $\iota : I \to X$ is a bijection; let $f : X \to I$ be the inverse function. Then for all $x \in X$ we have $x = \iota(i)$ for a unique $i \in I$, and $i = f(x)$, so

$$x = \iota(i) \in S_i = S_{f(x)}.$$

(ii) $\implies$ (i): Let $J \subseteq I$. For all $j \in J$, put $x_j \coloneqq f^{-1}(j)$. We have

$$x_j \in S_{f(x_j)} = S_j,$$

so $\{x_j \mid j \in J\} \subseteq \bigcup_{i \in J} S_i$, and since the map $j \mapsto x_j$ is the bijection f^{-1}, we have

$$\# \bigcup_{i \in J} S_i \geq \#\{x_j \mid j \in J\} = \#J. \qquad \square$$

REMARK 9.62. *An equivalent statement of condition (ii) of Theorem 9.61 is: there is an injection $g : I \to V$ such that for all $i \in I$ we have $g(i) \in S_i$. (Just take g to be the inverse function of f.) This is arguably more natural: for each element of I we can choose an element of S_i such that all these elements are distinct.*

THEOREM 9.63 (König-König). *Let $G = (V_1, V_2, E)$ be a bipartitioned graph. Suppose that we have a semiperfect matching $\iota_1 : V_1 \hookrightarrow V_2$ and also a semiperfect matching $\iota_2 : V_2 \to V_1$. Then G admits a perfect matching.*

PROOF. The proof is very close to that of Theorem 9.6. Namely, with $V = V_1 \coprod V_2$, we define a function $\iota : V \to V$ by $\iota|_{V_1} = \iota_1$ and $\iota|_{V_2} = \iota_2$. Since ι_1 and ι_2 are injections, so is ι. Therefore V gets partitioned into cycles under iteration of ι. Our goal is to use the cycle structure of ι to see how to modify ι_1 to get a map $\alpha : V_1 \to V_2$ that has the propery $x \sim \alpha(x)$ for all $x \in V_1$ and is a bijection: then α gives a perfect matching $M = \{\{v, \alpha(v)\} v \in V_1\}$ of G.

There are three types of cycles: finite cycles, singly infinite cycles and doubly infinite cycles, and ι induces a surjection from each finite cycle to itself and from each doubly infinite cycle to itself. An element $x_1 \in V_2$ does not lie in the image of ι_1 if and only it does not lie in the image of ι_2 if and only if the equivalence class $\mathfrak{c}_\iota(x_1)$ is a singly infinite cycle starting at x_1: that is, we have an infinite sequence

$$x_1, x_2, \ldots, x_n \ldots$$

of distinct elements with $x_{n+1} = \iota(x_n)$ for all $n \in \mathbb{Z}^+$. In terms of the graph G, we know that $x_n \sim x_{n+1}$ for all $n \in \mathbb{Z}^+$ and therefore for all $n \geq 2$ we have $x_n \sim x_{n-1}$ and $x_n \sim x_{n+1}$. So whereas the portion of the singly infinite cycle that lies in V_1 is $x_2, x_4, x_6, \ldots$ and we have $\iota_1(x_{2n}) = x_{2n+1}$ for all $n \in \mathbb{Z}^+$, we need only redefine ι on these values by

$$\forall n \in \mathbb{Z}^+, \alpha(x_{2n}) \coloneqq x_{2n-1}.$$

For every other $x \in V_1$ we define $\alpha(x) = \iota(x)$. This has the effect of replacing each single infinite ι-cycle starting at a vertex in V_2 with infinitely many pairs of 2-cycles: for all $n \in \mathbb{Z}^+$, $x_{2n-1} \mapsto x_{2n} \mapsto x_{2n-1}$. Thus α is a perfect matching.[5] $\square$

[5]We note that unlike in the proof of Theorem 9.6, we did not make the corresponding adjustment to ι_2. We could have done so, but it is not necessary to do so, either here or before, so we wanted to present both forms of the argument.

Not only is the proof of Theorem 9.63 very close to that of Theorem 9.6, actually the König-König Theorem is a generalization of the Dedekind-Schröder-Bernstein Theorem. Namely, let V_1 and V_2 be sets, and let $\iota_1 : V_1 \hookrightarrow V_2$ and $\iota_2 : V_2 \hookrightarrow V_1$ be injections. If we put

$$E \coloneqq \{\{v, \iota(v_1)\}\}_{v \in V_1} \cup \{\{w, \iota_2(w)\}\}_{w \in V_2}$$

then $G = (V_1, V_2, E)$ is a bipartitioned graph and $\iota_1 : V_1 \to V_2$ and $\iota_2 : V_2 \to V_1$ are semiperfect matchings. By Theorem 9.63 there is a perfect matching on G, which determines a bijection $\alpha : V_1 \to V_2$.

Finally, remember that we changed the rules a bit by looking at bipartitioned graphs. The conditions for an arbitrary finite graph to admit a perfect matching are known but lie a bit deeper. In order to state the following result, we need the notion of an **induced subgraph** of a graph $G = (V, E)$. Namely, for any subset $W \subseteq V$, we define the subraph induced by W to be

$$G_W \coloneqq (W, E \cap 2^W).$$

In other words, the vertices are the elements of W and the edges are the edges $e = \{w, w'\} \in E$ that run between vertices w, w' of W.

THEOREM 9.64 (Tutte). *For a finite graph G, the following are equivalent:*

(i) *The graph G admits a perfect matching.*
(ii) *For every subset $X \subseteq V$, the number of connected components of the induced subgraph $G_{V \setminus X}$ with an odd number of vertices is at most $\#X$.*

PROOF. By an **odd component** of a finite graph, we mean a connected component with an odd number of vertices. For a subset $X \subseteq V$, we write $\mathrm{odd}(G_X)$ for the set of odd components of the induced graph G_X.

(i) $\implies$ (ii): Let $\mathcal{M} \subseteq E$ be a perfect matching on G, and let $X \subseteq V$. Let C be an odd component of $G_{V \setminus X}$. Since $\#C$ is odd, there is at least one vertex v_C in C that is matched under $\mathcal{M}$ to a vertex $w_C \in X$. This defines a function

$$w : \mathrm{odd}(G_{V \setminus X}) \to X.$$

Since the edges of a matching are pairwise disjoint, w is an injection. Thus $\#\,\mathrm{odd}(G_{V \setminus X}) \leq \#X$.

(ii) $\implies$ (i): We show the contrapositive: suppose $G = (V, E)$ is a finite graph without a perfect matching. We must find a **Tutte violator**: that is, a subset $X \subseteq V$ with $\#X < \#\,\mathrm{odd}(G_{V \setminus X})$. We observe that if $\#V$ is odd, then G must have at least one odd component, so $\varnothing$ is a Tutte violator. Henceforth we suppose that $\#V$ is even. Since $\#V$ is even, for any $X \subseteq V$, we have $\#X \equiv \#\,\mathrm{odd}(G_{V \setminus X}) \pmod 2$.

Step 1: Suppose X is a Tutte violator in $G = (V, E)$. Then for every subset $E' \subseteq E$, X is also a Tutte violator in $G' \coloneqq (V, E')$: since $G'_{V \setminus X}$ is a subgraph of $G_{V \setminus X}$ obtained by removing a set of edges, every component of $G_{V \setminus X}$ is a union of components of $G'_{V \setminus X}$, and if any partition of an odd finite set has at least one element of odd order, so

$$\#X < \mathrm{odd}(G_{V \setminus X}) \leq \#\,\mathrm{odd}(G'_{V \setminus X}).$$

It follows that we may assume that G is *edge-maximal* in the sense if we adjoin any edge to G, then we do get a perfect matching.

Step 2: Let $S \subseteq V$ be the set of vertices of degree $\#V - 1$, i.e., the vertices that are adjacent to every vertex of v other than themselves. First we suppose that every component of $G_{V\setminus S}$ is a complete graph. In this case we claim that S is a Tutte violator: assuming it is not, we will find a perfect matching in G. Indeed, if S is not a Tutte violator, then $\#\operatorname{odd}(G_{V\setminus S}) \leq \#S$. Then we get a perfect matching by: (i) choosing a perfect matching in each component of $G_{V\setminus S}$ that is complete on an even number of vertices, (ii) on each component C of $G_{V\setminus S}$ that is a complete on an odd number of vertices, we may perfectly match all vertices except one vertex v_C, (iii) matching each v_C with some vertex of S, and (iv) matching the remaining vertices of S (a non-negative even number) with each other.
Step 3: Now suppose that some component C of $G_{V\setminus S}$ is not a complete graph: there are $x \neq y$ in C such that $e \coloneqq \{x, y\}$ is not an edge of G. Let

$$\ell : a_0 = x, a_1, a_2, \ldots, a_n = y$$

be a path from x to y in $G_{X\setminus S}$ of minimal length $n \geq 2$. The minimality implies

$$e_1 \coloneqq \{x, a_2\} \notin E.$$

Since $a_1 \notin S$, there is $c \in V$

$$e_2 \coloneqq \{a_1, c\} \notin E.$$

By the edge-maximality of G, we must have a perfect matching $\mathcal{M}_1$ in

$$\tilde{G}_1 \coloneqq (V, E \cup \{e_1\})$$

and a perfect matching $\mathcal{M}_2$ in

$$\tilde{G}_2 \coloneqq (V, E \cup \{e_2\}).$$

Moreover we must have $e_1 \in \mathcal{M}_1$, for otherwise $\mathcal{M}_1$ would be a perfect matching in G, and similarly we must have $e_2 \in \mathcal{M}_2$.

Now consider the path of maximal length

$$P : c = y_0, y_1, \ldots, y_N$$

starting from c and with first edge in $\mathcal{M}_1$, second edge in $\mathcal{M}_2$ next edge from $\mathcal{M}_1$, and so forth. The vertex y_1 to which $\mathcal{M}_1$ matches $c = y_0$ cannot be a_1, since $e_2 \notin \mathcal{M}_1$. Since $\mathcal{M}_\in$ matches c to a_1 and $y_1 \notin \{a_1, c\}$, the vertex to which $\mathcal{M}_2$ matches y_1 is not c, so this vertex is y_2 and the path has length at least 2. Since $\mathcal{M}_\infty$ has already matched y_0 to y_1, it must match y_2 to a new vertex $y_3 \notin \{y_0, y_1, y_2\}$, so the path has length at least 3. Now $\mathcal{M}_\in$ matches y_1 and y_2 to each each other so must match y_3 to something else; if this something else is $c = y_0$, then the path ends at length 3. Otherwise the path continues to a vertex $y_4 \neq y_0$.

Continuing in this manner we see that for any $k \geq 1$, if the path begins $y_0, \ldots, y_{2k-1}, y_{2k}$, then $\mathcal{M}_1$ matches y_0 to y_1, matches y_2 to y_3, and so forth, finally matching y_{2k-2} to y_{2k-1}, so $\mathcal{M}_\infty$ mut match y_{2k} to some new vertex $y_{2k+1} \notin \{y_0, \ldots, y_{2k}\}$. So the path must end at y_{2k+1} for some $k \geq 1$, when $\mathcal{M}_2$ matches y_{2k+1} to $y_0 = c$, which mean that $y_{2k+1} = a_1$.

Now let $\mathcal{M}$ consist of all edges that lie either in exactly one of $\mathcal{M}_\in$ and the cycle

$$\mathcal{C} \coloneqq \{\{y_0, y_1\}, \ldots, \{y_{2k}, y_{2k+1}\}, \{y_{2k+1}, y_0\}\}.$$

Since $e_2 = \{a_1, c\} = \{y_{2k+1}, y_0\}$ lies in both $\mathcal{M}_2$ and in $\mathcal{C}$, it does not lie in $\mathcal{M}$, so every element of $\mathcal{M}$ is an edge of G. We claim that $\mathcal{M}$ is a perfect matching in G, which will be a contradiction that ends the proof. Indeed, $\mathcal{M}_2$

is a perfect matching, and the only edges that lie in $\mathcal{M}_2$ and not in $\mathcal{M}$ are the edges $\{y_1, y_2\}, \ldots, \{y_{2k+1}, y_0\}$. All of these vertices are matched via the edge $\{y_0, y_1\}, \ldots, \{y_{2k}, y_{2k+1}\}$, which lie in $\mathcal{C}$ but not in $\mathcal{M}_2$ hence lie in $\mathcal{M}$. Moreover all the edges of $\mathcal{M}$ are pairwise disjoint: all the edges of $\mathcal{M}_2$ are pairwise disjoint, the edges $\{y_0, y_1\}, \ldots, \{y_{2k}, y_{2k+1}\}$ are pairwise disjoint, and we removed from $\mathcal{M}$ all the edges $\{y_1, y_2\}, \ldots, \{y_{2k+1}, y_0\}$, which were all the edges of $\mathcal{M}_2$ containing any vertex in $\mathcal{C}$. □

Although the criterion given in Tutte's Theorem for the existence of a perfect matching looks elegant, it is perhaps not immediately clear how useful it is. In fact this is a very useful result, and we give one classic application.

A graph is **cubic** if every vertex has degree 3.

THEOREM 9.65 (Petersen 1891). *Let G be a finite graph that is cubic and bridgeless. Then G admits a perfect matching.*

PROOF. Let $G = (V, E)$ be a finite, cubic, bridgeless graph. By Theorem 9.64, it suffices to show that for all $X \subseteq V$, the number $\#\operatorname{odd}(G_{V \setminus X})$ of odd components of the induced graph on $V \setminus X$ is at most $\#X$.

So let $X \subseteq V$, and let $C_i = (V_i, E_i)$ be an odd component of $G_{V \setminus X}$, and let be the number of edges of G with one vertex in V_i and the other vertex in X. Then

$$3\#V_i = \sum_{v \in V_i} \deg_G(v) = 2\#E_i + m_i.$$

Since $3\#V_i$ is odd, so is $2\#E_i + m_i$ and thus so is m_i. Since G is bridgeless, we must have $m_i \geq 3$. Let $\mathcal{E}$ be the number of edges in G with one vertex in X and one vertex in $V \setminus X$. Above we saw that each odd component contributes $m_i \geq 3$ edges to $\mathcal{E}$, and these sets of edges are pairwise disjoint, so

$$\#\mathcal{E} \geq 3\#\operatorname{odd}(G_{V \setminus X}).$$

Since every $e \in \mathcal{E}$ has a vertex in X and the graph G is cubic, we have

$$\#\mathcal{E} \leq 3\#X.$$

Combining these inequalities, we get

$$\#X \geq \frac{\#\mathcal{E}}{3} \geq \#\operatorname{odd}(G_{V \setminus X}),$$

confirming that X is not a Tutte violator. □

In particular Petersen's Theorem implies that a finite, bridgeless cubic graph has an even number of vertices. This is not really a surprise. More generally, for $d \in \mathbb{N}$ we say that a graph is **d-regular** if every vertex has degree d. A graph is **regular** if it is d-regular for some d; that is, if it is locally finite and all vertices have the same degree. Then it follows from Proposition 9.31b) that if d is odd, every finite d-regular graph has an even number of vertices.

EXAMPLE 9.66. *The **Petersen graph** is a famous cubic graph with* 10 *vertices. It can be described as follows: first consider the vertices of a regular pentagon on the unit circle. Label these vertices in cyclic counterclockwise order as* $(1, 1)$, $(2, 1)$,

$(3,1)$, $(4,1)$, $(5,1)$. *(These are* not *the x and y coordinates of these vertices; those are $(\cos\frac{2\pi k}{5}, \sin\frac{2\pi k}{5})$ for $k \in [5]$. They are just labels.) We include edges:*

$$(1,1) \sim (3,1) \sim (5,1) \sim (2,1) \sim (4,1) \sim (1,1).$$

Now we consider the vertices of a second regular pentagon on the circle of radius 2, labelling them in cyclic counterclockwise order as $(1,2)$, $(2,2)$, $(3,2)$, $(4,2)$, $(5,2)$. We include edges:

$$(1,2) \sim (2,2) \sim (3,2) \sim (4,2) \sim (5,2) \sim (1,2).$$

So far we have two disjoint 5-cycles described in a slightly complicated way. Finally though we add five more edges, connecting each inner vertex to the corresponding outer vertex:

$$\forall 1 \leq i \leq 5,\ (i,1) \sim (i,2).$$

Or, if you like, consult `https://en.wikipedia.org/wiki/Petersen_graph`*: this is a case where a picture is better than a bunch of points and $\sim$'s.*

The Petersen graph is a bridgeless cubic graph. It has a visible perfect matching:

$$\mathcal{M} = \{\{(i,1),\ (i,2)\} \mid 1 \leq i \leq 5\}.$$

You are asked to prove a generalization of Petersen's Theorem giving a sufficient condition for a finite d-regular graph to admit a perfect matching in Exercise 9.30.

3.6. Graph Derangements. Let $G = (V, E)$ be a graph. A **graph injection** is an injection $\iota : V \to V$ such that for all $v \in V$, either $\iota(v)$ is adjacent to v or $\iota(v) = v$. A **graph derangement** is an injection $\iota : V \to V$ such that for all $v \in V$ we have that $\iota(v)$ is adjacent to v.

On any nonempty set X, let us say that $f : X \to X$ is a **derangement on X** if it is a fixed-point free injection: that is, f is an injection such that for all $x \in X$ we have $f(x) \neq x$. Since no vertex in a graph is adjacent to itself, every graph derangement $\iota : V \to V$ is indeed a derangement on V.

If the graph is complete on its vertex set – i.e., the edge set consists of all two-element subsets of V – then any derangement is a graph derangement. On any finite set, a graph injection on X (or any injection on X!) is necessarily surjective (Theorem 8.58). On the singly infinite path: i.e., the graph with vertex set $\mathbb{Z}^+$ and $x \sim y$ if $|x - y| = 1$, we have that $\iota(n) = n + 1$ is a graph derangement that is not surjective. Nevertheless, in Exercise 9.26 you are asked to show that if a graph admits a graph derangment then it also admits a bijective graph derangement. (In fact treating the case of a singly infinite path is the crux of the matter.)

EXAMPLE 9.67. *Let $G = (V, E)$ be a graph.*

a) *Let $M \subseteq E$ be a perfect matching. Then M determines a surjective graph derangement: for each $v \in V$, there is a unique $e \in M$ such that $e = \{v, v'\}$ and we put $f(v) \coloneqq v'$.*

b) *Suppose that G is finite, and let $f : V \to V$ be a graph derangement such that there is just one $\sim_f$-equivalence class in the sense of §9.1: that is, if we fix $v_0 \in V$, then every $v \in V$ is of the form $f^{\circ k}(v_0)$ for some $k \in \mathbb{N}$. Letting n be the least positive integer such that $f^{\circ n}(v_0) = v_0$, we have that*

$$v_0, f(v_0), f^{\circ 2}(v_0), \ldots, f^{\circ n}(v_0)$$

*is a cycle in G in which each vertex of G appears: a **Hamiltonian cycle**.*

Thus the concept of a graph derangement is a generalization/interpolation of the graph-theoretic concepts of perfect matchings and Hamiltonian cycles.

In (14) above we determined the number of derangements on a finite set with n elements. We can now see that this is also the number of graph derangements on the complete graph K_n on $[n]$. It follows that as $n \to \infty$ the probability that a graph injection on K_n is a derangement approaches $\frac{1}{e}$.

EXAMPLE 9.68. *For $m, n \in \mathbb{Z}^+$ we consider the* ***checkerboard graph*** *$C(m,n)$. It has vertex set $[m] \times [n]$ and we have $(x_1, y_1) \sim (x_2, y_2)$ if and only if*

$$|x_1 - x_2| + |y_1 - y_2| = 1.$$

This graph models the vertices on an $m \times n$ checkerboard, and two vertices are adjacent if the corresponding squares are "orthogonally adjacent." I claim that $C(m,n)$ admits a graph derangement if and only if m and n are not *both odd.*

Indeed, if either m or n is even, then $C(m,n)$ admits a perfect matching, which can here be visualized as a domino tiling*. If the number of rows m is even, then we can tile every row with $\frac{m}{2}$ dominos; similarly if the number of columns n is even.*

Next we observe that this graph has a natural bipartition, which corresponds to the usual checkerboard coloring, say into red and black vertices. If m and n are both odd, then the number of vertices mn is also odd, so the number of red and black vertices cannot be the same (indeed they differ by one). Without loss of generality we suppose that there are more red vertices than black vertices. Then there cannot be a graph derangement on $C(m,n)$, because every red vertex should get mapped to a different black vertex, and there are not enough black vertices to do this.

Let us reflect a bit on Example 9.68. First of all it gives the solution to Exercise 3.15, which is the case $m = n = 5$. Second of all it is an instance of the Hall Condition being violated in 9.59: e.g. when $m = n = 5$, the 13 red vertices have only 12 black neighbors.

This suggests focusing on the Hall Condition in general. If a graph $G = (V, E)$ admits a graph derangement $f : V \to V$, then for any finite subset $X \subseteq V$, then since $f(X) \subseteq N(X)$ we must have $\#X \leq \#N(X)$. It is rather remarkable that this simple necessary condition turns out to be sufficient for all locally finite graphs. In fact even an apparently weaker condition suffices.

THEOREM 9.69 (Tutte [**Tu53**]). *For a locally finite graph $G = (V, E)$, the following are equivalent:*

(i) *There is a graph derangement $f : V \to V$.*
(ii) *For all finite independent subsets $X \subseteq V$, we have $\#X \leq \#N(X)$.*

PROOF. (i) $\implies$ (ii): Above we explained why the existence of a graph derangement on any graph implies $\#X \leq \#N(X)$ for all finite subsets $X \subseteq V$.
(ii) $\implies$ (i): Step 1: We show that (ii) implies the apparently stronger condition
(ii′) For all finite subsets $X \subseteq V$, we have $\#X \leq \#N(X)$.
We show this by contraposition: suppose that there is a finite subset $X \subseteq V$ such that $\#X > \#N(X)$. Put

$$Y := X \setminus N(X),$$

so Y is an independent set. Put

$$m_1 := \#Y,\ m_2 := \#(X \setminus Y),\ n := m_1 + m_2 = \#X.$$

By assumption we have $\#N(X) < n = m_1 + m_2$; since

$$N(Y) \subseteq N(X)$$

and

$$X \setminus Y \subseteq N(X) \setminus N(Y),$$

we have

$$\#N(Y) \leq \#N(X) - (\#X \setminus Y) < n - m_2 = m_1 = \#Y.$$

Step 2: Now we assume (ii′). For $x \in V$, let S_x be the set of neighbors of x; the local finiteness condition precisely means that each S_x is a finite set. The indexed family $\{S_x\}_{x \in V}$ of finite subsets of V satisfies the Hall Condition: for every finite subset $J \subseteq V$ we have

$$\#J \leq \#N(J) = \bigcup_{v \in J} S_v.$$

By Theorem 9.61 (cf. Remark 9.62) there is an injection $f : V \hookrightarrow V$ such that for all $v \in V$ we have $f(v) \in S_v$, i.e., we have that $f(v)$ is adjacent to v. So f is a graph derangement. □

In Exercise 9.27 you are asked to show that a locally finite bipartitioned graph admits a graph derangement if and only if it admits a perfect matching.

We end with a discussion of an open problem concerning graph derangements. For simplicity, we restrict to finite graphs $G = (V, E)$. Then any graph injection $f : V \to V$ is a bijection, so has a **cycle type** $(k_1, \ldots, k_r)$. Notice that f is a graph derangement if and only if $k_1 \geq \ldots \geq k_r \geq 2$. This suggests a refinement of the question of asking whether a given finite graph G admits a graph derangement: namely, of determining all cycle types of graph derangements (or even of all graph injections) on G.

If G is moreover bipartite, then by Exercise 9.20 it admits no cycle of odd length $\ell \geq 3$, so the only possible cycle types $(k_1, \ldots, k_r)$ are those with k_i even for all $1 \leq i \leq r$. (Let us point out a 2-cycle in the cycle type decomposition corresponds to an edge that matches 2 vertices. This is a cycle in the sense of §9.1 but not in the graph-theoretic sense.) Let us say that a finite bipartite graph $G = ([n], E)$ is **even universal** if for all even positive integers $k_1 \geq \ldots \geq k_r$ with $k_1 + \ldots + k_r = n$, there is a graph derangement $f : [n] \to [n]$ with cycle type $(k_1, \ldots, k_r)$.

EXAMPLE 9.70. a) *Consider the checkerboard graph $C(2,2)$. It has 4 vertices, so the even cycle types are (4) and $(2,2)$.*
$f_1 : [2] \times [2] \to [2] \times [2]$ by

$$(1,1) \mapsto (1,2) \mapsto (2,2) \mapsto (2,1) \mapsto (1,1)$$

is a graph derangement of cycle type (4), while $f_2 : [2] \times [2] \to [2] \times [2]$ by

$$(1,1) \mapsto (1,2) \mapsto (1,1),\ (2,1) \mapsto (2,2) \mapsto (2,1)$$

is a graph derangement of cycle type $(2,2)$. Thus the graph $C(2,2)$ is even universal.

b) *Consider the checkerboard graph $C(2,3)$. It has 6 vertices, so the even cycle types are (6), $(4,2)$ and $(2,2,2)$. By drawing pictures on a 2×3 rectangle divided into squares, it is easy to see that there are graph*

derangements with all three of these cycle types. Thus the graph $C(2,3)$ is even universal.

c) *For $n \geq 2$, consider the checkerboard graph $C(2,n)$. Again it is not hard to see that there are graph derangements of all possible even cycle types: that is, $C(2,n)$ is even universal.*

d) *Consider the checkerboard graph $C(3,4)$. It turns out that there is no graph derangement on $C(3,4)$ with cycle type $(8,4)$. Any four-cycle on a checkerboard graph determines a 2×2 square inside the rectangle. Up to symmetries of the rectangle there are only two possible ways to place this square: it is either $S_1 = \{(1,1),\ (1,2),\ (2,1),\ (2,2)\}$ or $S_2 = \{(1,2),\ (1,3)\ (2,2),\ (2,3)\}$. In the graph $C(3,4) \setminus S_1$, the vertex $(3,1)$ is pendant, hence not part of any cycle of length 8. In the graph $C(3,4) \setminus S_2$, the vertex $(1,1)$ is pendant, hence not part of any cycle of length 8. There is also no graph derangement of cycle type $(4,4,4)$, since this would involve writing $C(3,4)$ as a disjoint union of three 2×2 squares, which a little thought shows is not possible. It is not hard see that all the other even cycle types*

$$(12),\ (10,2),\ (8,2,2),\ (6,6),\ (6,4,2),\ (6,2,2,2),\ (4,4,2,2),\ (4,2,2,2,2),\ (2,2,2,2,2,2)$$

can be realized by graph derangements. In particular $C(3,4)$ is not *even universal.*

Some further instances of pairs m and n (with mn even) such that $C(m,n)$ is or is not not even universal were given in [**Cl13**, §4.2]. For instance, $C(4,5)$ is not even universal, while $C(4,6)$ is even universal.

QUESTION 9.71. *Is it true that for all even $m,n \geq 6$ the checkerboard graph $C(m,n)$ is even universal?*

In 2013 I filled up several notepads showing an affirmative answer to Question 9.71 for a few values of m and n. It was an agreeable way to kill time, although after spending hours at my local coffee shop intently poring over doodles on a notepad, I eventually became self-conscious that others might interpret my pastime as evidence of a mental disorder. In any case, although the purpose of this text was to introduce basic mathematical structures and concepts and show how quickly mathematical edifices can be built on these firm foundations, I was not able to find any theory that could usefully be brought to bear on this problem, so instead just worked out as many cases as time allowed. Perhaps *you* can solve this problem, or even make some progress on it. I would love to hear about it if you do.

4. Theorems of Sperner, Dilworth and Mirsky

4.1. Sperner's Theorem. Recall the notion of a **Sperner family** $\mathcal{F}$ of subsets of $[n]$ for some positive integer n: this is a set of subsets of $[n]$ no one of which is properly contained in any other. We mentioned that determining the number D_n of Sperner families is open for all $n \geq 9$. But another question about Sperner families was answered by Sperner himself in 1928 [**Sp28**], a result that initiated an entire branch of mathematics, **extremal combinatorics**. Is is the following: rather than trying to count Sperner families, we can ask how large a Sperner family of subsets of $[n]$ can be: i.e., what is the largest number k of subsets $A_1, \ldots, A_k$ of $[n]$ such that for no $1 \leq i,j \leq n$ do we have $A_i \subsetneq A_j$?

EXAMPLE 9.72.

a) *Let $n = 1$. The only two subsets of $[1]$ are $\varnothing$ and $\{1\}$ and the former properly contains the latter, so the Sperner families are $\mathcal{F}_1 = \varnothing$, $\mathcal{F}_2 = \{\varnothing\}$ and $\mathcal{F}_3 = \{\{1\}\}$. So the largest size of a Sperner family is 1.*

b) *Let $n = 2$. There are 6 Sperner families of subsets of $[2]$: the empty family, the $2^2 = 4$ one-element families, and one two-element family, $\mathcal{F} = \{\{1\}, \{2\}\}$.*

c) *Let $n = 3$. As seen in Example 2.39, there are 20 Sperner families of subsets of $[3]$. The largest such families are*

$$\mathcal{F}_1 = \{\{1\}, \{2\}, \{3\}\} \text{ and } \mathcal{F}_2 = \{\{1,2\}, \{1,3\}, \{2,3\}\},$$

each with three elements.

Looking over Example 9.72 we see that for all $1 \leq n \leq 3$, the family of 1-element subsets of $[n]$ is a Sperner family of largest possible size. Moreover for $1 \leq n \leq 2$, this is the only Sperner family of maximal size, whereas for $n = 3$ the family of al 2-element subsets of [3] is another Sperner family of maximal size. For any $n \in \mathbb{Z}^+$, the family of all 1-element subsets of $[n]$ is a Sperner family of size n. Is it reasonable to guess that this is always a Sperner family of maximal size?

Well, it may be a reasonable *guess* but it is easily seen not to be the case for any $n \geq 4$. For a set X and $k \in \mathbb{N}$, let us write $\binom{X}{k}$ for the set of all k-element subsets of X. The point of this notation is that

$$\forall n \in \mathbb{N},\ \#\binom{[n]}{k} = \binom{n}{k}.$$

Now for any set X and any $k \in \mathbb{N}$, the set $\binom{X}{k}$ is a Sperner family of subsets of X, just because any two elements are finite sets with k elements, and we cannot have a proper containment between two finite sets with the same number of elements. Coming back to the case of $n = 4$, we see that in particular $\binom{[4]}{2}$ is a Sperner family of subsets of [4], with size $\binom{4}{2} = 6$, so it is larger than the Sperner family of all 1-element subsets of [4], which by the way is $\binom{[4]}{1}$. So is $\binom{[4]}{2}$ a Sperner family of subsets of [4] of maximal size?

This is already an interesting question. It is certainly the largest Sperner family we can take by taking all k-element subsets of [4], because $\binom{4}{0} = 1$, $\binom{4}{1} = 4$, $\binom{4}{3} = 4$, $\binom{4}{4} = 1$ are all smaller than $\binom{4}{2}$. So far we have checked five Sperner families of [4]. According to a 1940 result of Church (Theorem 2.40a)) there are altogether $D_4 = 168$ Sperner families of [4]. If we happen to have access to a list of them, we could examine them all and see whether any of them has more than six subsets. Well, someone has access to this list, but we don't. Another thing we could do is simply check all 7-element families of subsets of [4] and see whether any of them are Sperner families. (This works because any subfamily of a Sperner family is a Sperner family, so if there is a Sperner family with more than six elements, there must be one with exactly 7 elements.) There are $\binom{2^4}{7} = 11440$ such families. For each such family $\mathcal{F} = \{A_1, \ldots, A_7\}$, we could check all $7 \cdot 6 = 42$ ordered pairs (A_i, A_j) of sets in the family; if for each $1 \leq i \neq j \leq 7$ there is at least one element of A_i that is not an element of A_j, then A_i is not a subset of A_j, and we have found a Sperner family, and conversely if for some $i \neq j$ every element of A_i is also an element of A_j then A_i is a subset of A_j (necessarily proper, since $A_i \neq A_j$), and we don't have a Sperner family. This is the sort of thing that it would be easy to write

code for a computer to check, and I think that any computer you have in the year $N \geq 2023$ would succeed in this calculation. Let's pretend we did that, and I'll tell you the answer: in fact $\binom{4}{2}$ is the largest Sperner family of [4]. Are we satisfied?

No! We want to provide proofs that (i) provide insight and (ii) can be made to work to prove more general results. Let me show you a proof of this same result that satisfies these criterion.

PROPOSITION 9.73. *The largest size of a Sperner family of subsets of* $[4]$ *is* 6.

PROOF. As mentioned above, the family $\binom{[4]}{2}$ of 2-element subsets of [4] is a Sperner family of size 6, so the matter of it is to show that if $\mathcal{F}$ is a Sperner family of subsetes of [4] then $\#\mathcal{F} \leq 6$. We will show this by contemplating a certain partition of the set $2^{[4]}$ of subsets of [4]. If we put

$$\mathcal{C}_1 = \{\varnothing,\ \{1\},\ \{1,2\},\ \{1,2,3\},\ \{1,2,3,4\}\},$$
$$\mathcal{C}_2 = \{\{2\},\ \{2,3\},\ \{2,3,4\}\},$$
$$\mathcal{C}_3 = \{\{3\},\ \{3,4\}\},$$
$$\mathcal{C}_4 = \{\{4\},\ \{2,4\}\},$$
$$\mathcal{C}_5 = \{\{1,3\},\ \{1,3,4\},$$
$$\mathcal{C}_6 = \{\{1,4\},\ \{1,2,4\},$$

then our partition is

$$\mathcal{P} = \{\mathcal{C}_1, \mathcal{C}_2, \mathcal{C}_3, \mathcal{C}_4, \mathcal{C}_5, \mathcal{C}_6\}.$$

Note that each $\mathcal{C}_i$ is *not* a Sperner family (or "antichain"): on the contrary, each subset $\mathcal{C}_i$ has the property that, in the order we have written the elements, each subset is properly contained in the next. In other words, given any two elements of $\mathcal{C}_i$, one of them is contained in the other. Such a family of subsets of any set X is called a **chain** in X. So we have written $2^{[4]}$ as the disjoint union of six chains.

Remarkably, this implies the desired result! Here is why: let $\mathcal{F}$ be a Sperner family. Then there is a function $\iota : \mathcal{F} \to \mathcal{P}$ in which we send each element $A \in \mathcal{F}$ to the unique element $\mathcal{C}_i$ containing A. Because $\mathcal{F}$ is a Sperner family, the map f must be injective: otherwise we have distinct elements A and B in $\mathcal{F}$ mapping into the same set $\mathcal{C}_i$: but since A and B lie in the Sperner family $\mathcal{F}$ neither contains the other; moreover A and B both lie in the chain $\mathcal{C}_i$, so one must contain the other: contradiction! But now the Pigeonhole Principle tells us that $\#\mathcal{F} \leq \#\mathcal{P}$, otherwise there are no injective functions. So $\#\mathcal{F} \leq \#\mathcal{P} = 6$, as we wanted to show. □

Looking back over the proof of Proposition 9.73 in terms of the two criteria mentioned just before its statement, I think it does a better job on both (i) and (ii) than the brute force argument we pretended to give. It does an especially good job at "can be made to work to prove more general results," since the Pigeonhole Principle argument at the end actually shows the following:

THEOREM 9.74.

a) *Let X be a set. Let $\mathcal{F}$ be a Sperner family of subsets of X (i.e., for no two distinct $A, B \in \mathcal{F}$ do we have $A \subseteq B$), and let $\mathcal{P}$ be a partition of the power set 2^X for which each element $\mathcal{C} \in \mathcal{P}$ is a chain (i.e., for all $A, B \in \mathcal{C}$ we have either $A \subseteq B$ or $B \subseteq A$). Then mapping each element $A \in \mathcal{F}$ to the unique element of $\mathcal{P}$ containing A gives an injection $\iota : \mathcal{F} \hookrightarrow \mathcal{P}$.*

b) *If X is finite, then the size of the largest Sperner family in X is at most the smallest size of a partition of 2^X into chains.*

You are asked to prove Theorem 9.74 in Exercise 9.31.

Theorem 9.74 provides a powerful tool for giving *upper bounds* on the size of the largest Sperner family of subsets of $[n]$: an upper bound is given by any partition of $2^{[n]}$ into chains. Certainly $2^{[n]}$ *can* always be partitioned into chains: indeed, we can just take the discrete partition $\mathcal{P} = \{\{A\} \mid A \subseteq [n]\}$, which has size 2^n. Doing this, the upper bound we get on the size of any Sperner family of $[n]$ is 2^n. This is not really news: a Sperner family is in particular a subset of $2^{[n]}$, so its size is at most the number of elements of $2^{[n]}$, which is 2^n.

So the game is to partition $2^{[n]}$ into as few chains as possible. Let us look at the case of $n = 5$. In this case, we already know that for all $0 \leq k \leq 5$, the family $\binom{[5]}{k}$ of all k-element subsets of $[5]$ is a Sperner family. The sizes of these Sperner families are

$$\binom{5}{0} = 1,\ \binom{5}{1} = 5,\ \binom{5}{2} = 10,\ \binom{5}{3} = 10,\ \binom{5}{4} = 5,\ \binom{5}{5} = 1,$$

so among these the largest are $\binom{[5]}{2}$ and $\binom{[5]}{3}$, each with 10 elements. So using the same ideas as above, if we wanted to show that the largest size of a Sperner family of $[5]$ is 10, it suffices to partition $2^{[5]}$ into 10 chains. To this by hand is actually not so bad; you are asked to do it in Exercise 9.32.

We have by now just about zeroed in on what the result should be. One more observation: let $n \in \mathbb{Z}^+$. If n is even, then by Exercise 3.5 the finite sequence $\binom{n}{k}$ for $0 \leq k \leq n$ is strictly increasing on $0 \leq k \leq \frac{n}{2}$ and then strictly decreasing on $\frac{n}{2} \leq k \leq n$, so the unique largest binomial coefficient is the middle one $\binom{n}{\frac{n}{2}}$. If n is odd, then by the same exercise shows that the sequence $\binom{n}{k}$ is strictly increasing on $0 \leq k \leq \frac{n-1}{2}$, then $\binom{n}{\frac{n-1}{2}} = \binom{n}{\frac{n+1}{2}}$, then it is strictly decreasing on $\frac{n+1}{2} \leq k \leq n$, so that the largest value occurs at the two middlemost coefficients, $\binom{n}{\frac{n-1}{2}}$ and $\binom{n}{\frac{n+1}{2}}$. Thus the largest size of a Sperner family of $[n]$ is at least $\binom{n}{\frac{n}{2}}$ when n is even and at least $\binom{n}{\frac{n-1}{2}}$ when n is odd. And now:

THEOREM 9.75 (Sperner [**Sp28**]). *Let $n \in \mathbb{Z}^+$.*

a) *If n is even, then $\binom{[n]}{\frac{n}{2}}$ is a Sperner family of $[n]$ of maximum size.*

b) *If n is odd, then both $\binom{[n]}{\frac{n-1}{2}}$ and $\binom{[n]}{\frac{n+1}{2}}$ are Sperner families of $[n]$ of maximum size.*

PROOF. By Theorem 9.74, it is enough to partition $2^{[n]}$ into $\binom{n}{\frac{n}{2}}$ chains if n is even and into $\binom{n}{\frac{n-1}{2}}$ chains if n is odd. To do this, when n is even it is enough to find a partition of $2^{[n]}$ into chains such that each chain contains exactly one subset of $[n]$ of size $\frac{n}{2}$; and when n is odd it is enough to find a partition of $2^{[n]}$ into chains such that each chain contains exactly one subset of $[n]$ of size $\frac{n-1}{2}$ and exactly one subset of size $\frac{n+1}{2}$. The rest of the proof builds these chains in the two cases.

a) Suppose that n is even. We CLAIM that for all $0 \leq k \leq \frac{n}{2}$ there is an injection

$$\alpha_k : \binom{[n]}{k} \hookrightarrow \binom{[n]}{k+1}$$

such that for all $A \in \binom{[n]}{k}$ we have $A \subsetneq \alpha_k(A)$. In plainer terms, we are claiming that for each $0 \leq k \leq \frac{n}{2}$ there is a way to add one element to each k-element subset of $[n]$ to obtain a $k+1$-element subset in such a way that we get a different $k+1$-element subset for each k-element subset that we started with. Similarly, we claim that for all $\frac{n}{2}+1 \leq k \leq n$ we have an injection

$$\beta_k : \binom{[n]}{k} \hookrightarrow \binom{[n]}{k-1}$$

such that for all $B \in \binom{[n]}{k}$, we have $\beta_k(B) \subsetneq B$. In plainer terms, we are claiming that for each $\frac{n}{2}+1 \leq k \leq n$, there is a way to remove one element from each k-element subset of $[n]$ to obtain a $k-1$-element subset in such a way that we get a different $k-1$-element subset for each k-element subset that we started with. Let us first assume that these maps α_k and β_k exist and show how to use them to build the desired partition of $2^{[n]}$.

Suppose that Y is a k-element subset of $[n]$ for some $0 \leq k < \frac{n}{2}$. Then applying α_k we get a $k+1$-element subset $\alpha_k(Y)$. Applying α_{k+1} we get a $k+2$-element subset $\alpha_{k+1}(\alpha_k(Y))$, and so forth: by applying these maps enough times we will get an $\frac{n}{2}$-element subset of $[n]$. Similarly, if Z is a k-element subset of $[n]$ for some $\frac{n}{2} < k \leq n$, then by applying β_k, then β_{k-1}, and so forth, we will eventually get an $\frac{n}{2}$-element subset of $[n]$. Now we have one element $\mathcal{C}_X$ of the partition for each $\frac{n}{2}$-element subset X of $[n]$. The set $\mathcal{C}_X$ has as elements:
• X; • Every k-element subset Y with $k < \frac{n}{2}$ such that repeatedly applying the α maps as above yields X; and
• Every k-element subset Z with $k > \frac{n}{2}$ such that repeatedly applying the β maps as above yields X.

Every subset of $[n]$ lies in exactly one $\mathcal{C}_X$ and each $\mathcal{C}_X$ contains X as an element, so

$$\mathcal{P} \coloneqq \{\mathcal{C}_X\}_{X \in \binom{[n]}{\frac{n}{2}}}$$

is a partition of $2^{[n]}$. To see why each $\mathcal{C}_X$ is a chain, it helps to think of reversing the α and β processes. Starting at the set X, we try to "go down," i..e., find an $\frac{n}{2}-1$ element subset Y such that $\alpha_{\frac{n}{2}-1}(Y) = X$. This may or may not be possible – if not, X is the bottom element of the chain $\mathcal{C}_X$ – but if it is, there is a *unique* such Y because $\alpha_{\frac{n}{2}-1}$ is an injection. Now we try to "go down" again starting with $\alpha_{\frac{n}{2}-1}(Y)$. This process must terminate after at most $\frac{n}{2}$ steps, and we get a chain of sets descending from X. Now we start again from X and try to "go up," i.e., find an $(\frac{n}{2}+1)$-element subset Z such that $\beta_{\frac{n}{2}+1}(Z) = X$. This may or may not be possible, but if so it is possible in exactly one way. So we can also go up at most $\frac{n}{2}$ times from X. This shows that for any two distinct elements of $\mathcal{C}_X$, one of them has fewer elements than the other, and the one with fewer elements is contained in the one with more elements: so $\mathcal{C}_X$ is a chain.

To complete the proof of part a) we need to construct the maps α_k and β_k. This seems like it should be the hardest part – and perhaps it is, but by now we have some friends in high places. Let $0 \leq k < \frac{n}{2}$. There is a finite partitioned graph

$G_k := (\binom{[n]}{k}, \binom{[n]}{k+1}, E)$, where an edge connects a k-element subset X_1 to a $k+1$-element subset X_2 if and only if $X_1 \subsetneq X_2$. Then the desired map $\alpha_k : \binom{[n]}{k} \hookrightarrow \binom{[n]}{k+1}$ is precisely a semiperfect matching in G_k, so we can show its existence using Hall's Marriage Theorem. Even better, every vertex X_1 in $\binom{[n]}{k}$ has degree $n-k$ – this is the number of remaining elements of $[n]$ that we can use to add an element to X_1 – while every vertex X_2 in $\binom{[n]}{k+1}$ has degree $k+1$: we can remove any one of the $k+1$-element subset X_2 to get a k-element subset. Because $k < \frac{n}{2}$ we have $2k \leq n-1$ and thus $n-k \geq k+1$. (In fact, because n is even and thus $n-1$ is odd, we must have $2k < n-1$ and thus $n-k > k+1$.) By Corollary 60a), a semiperfect matching exists. The existence of the injections β_k for $\frac{n}{2} < k \leq n$ can be proved in a very similar way; we leave it to the reader to show this in Exercise 9.33a). This completes the proof of part a).

b) Suppose now that n is odd. The overall strategy is the same as in part a). In this case, we want to construct:

• For all $0 \leq k < \frac{n-1}{2}$, an injection $\alpha_k : \binom{[n]}{k} \to \binom{[n]}{k+1}$ such that for all $Y \in \binom{[n]}{k}$, we have $Y \subsetneq \alpha_k(Y)$;

• For all $\frac{n+1}{2} < k \leq n$, an injection $\beta_k : \binom{[n]}{k} \to \binom{[n]}{k-1}$ such that for all $Z \in \binom{[n]}{k}$, we have $\beta_k(Z) \subsetneq Z$; and

• A bijection $\gamma : \binom{[n]}{\frac{n-1}{2}} \to \binom{[n]}{\frac{n+1}{2}}$ such that for all $X \in \binom{[n]}{\frac{n-1}{2}}$, we have $X \subsetneq \gamma(X)$.

These maps can similarly be constructed via Corollary 60; you are asked to fill in the details in Exercise 9.33b). Then for each $X \in \binom{[n]}{\frac{n-1}{2}}$ we define a subset $\mathcal{C}_X$ of $2^{[n]}$ which contains:

• X and $\gamma(X)$; • Every subset obtained by repeatedly going down from X via the α maps; and

• Every subset obtained by repreatedly going up from $\gamma(X)$ via the β maps.

The same arguments as in part a) then show that

$$\mathcal{P} := \{\mathcal{C}_X\}_{X \in \binom{[n]}{\frac{n-1}{2}}}$$

is a partition of $2^{[n]}$ in which each element is a chain. This completes the proof of Sperner's Theorem. □

Sperner's Theorem is as of this writing almost 100 years old, and it has several proofs. We like this proof, first, because it is a nice application of Hall's Marriage Theorem, and second because it establishes several other results of interest:

COROLLARY 9.76. *Let $n \in \mathbb{Z}^+$.*

a) *The minimum number of elements in a partition of $2^{[n]}$ into chains is the largest size of a Sperner family in $[n]$.*

b) *If $0 \leq k < \frac{n}{2}$, it is possible to injectively map $\binom{[n]}{k}$ into $\binom{[n]}{k+1}$ by adding one element to each k-element subset.*

c) *If $\frac{n}{2} < k \leq n$, it is possible to injectively map $\binom{[n]}{k}$ into $\binom{[n]}{k-1}$ by removing one element from each k-element subset.*

d) *If n is odd, it is possible to bijectively map $\binom{[n]}{\frac{n-1}{2}}$ to $\binom{[n]}{\frac{n+1}{2}}$ by adding one element to each $\frac{n-1}{2}$-element subset.*

Another nice example of a result that follows easily from this proof of Sperner's Theorem is given in Exercise 9.34.

On the other hand, in fact not only is it always the case that the collection of k-element subsets for suitable k gives a maximal size Sperner family in $[n]$ but that these are the only maximal size Sperner families: i.e., if n is even then $\binom{[n]}{\frac{n}{2}}$ is the unique maximal size Sperner family while if n is odd then $\binom{[n]}{\frac{n-1}{2}}$ and $\binom{[n]}{\frac{n-1}{2}}$ are the only maximal size Sperner families. An approach due (independently) to Yamamoto [**Ya54**], Meshalkin [**Me63**], Bollobás [**Bo65**] and Lubell [**Lu66**] naturally yields this more precise conclusion.

4.2. Dilworth's Theorem. As mentioned, our proof of Sperner's Theorem also establishes that the largest size of a Sperner family in $[n]$ is equal to the minimum number of elements in a partition of $2^{[n]}$ into chains. In Theorem 9.74 we gave a simple, transparent argument for why the former quantity must be less than or equal to the latter quantity, but that they turn out to be equal for all $n \in \mathbb{Z}^+$ looks rather lucky: we took a particular Sperner family $\mathcal{F}$ that we didn't know was of maximal size and managed to construct a partition of $2^{[n]}$ into $\#\mathcal{F}$ chains.

In fact the equality of these two quantities can be shown independently of the computation of either of them and in much more generality. We work in the context of partially ordered sets: a set X endowed with a relation $\leq$ that is reflexive, anto-symmetric and transitive. If X is any set, then the inclusion relation $\subseteq$ is a partial ordering on the power set 2^X: that is, if we have subsets A, B, C of X, then $A \subseteq A$, if $A \subseteq B$ and $B \subseteq A$ then $A = B$, and if $A \subseteq B$ and $B \subseteq C$ then $A \subseteq C$. These are all very familiar set-theoretic properties.

In any partially ordered set $(X, \leq)$ a **chain** is a subset $\mathcal{C} \subseteq X$ that is totally ordered under the restricted relation $\leq$: that is, for any $x, y \in \mathcal{C}$ we have that $x \leq y$ or $y \leq x$. Note that if X itself is totally ordered then every subset is a chain, but if it is only partially ordered then it is helpful to consider the chains inside it. An **antichain** is a subset $\mathcal{A} \subseteq X$ such that for no elements $a_1 \neq a_2$ in $\mathcal{A}$ do we have $a_1 \leq a_2$. Thus when $X = 2^{[n]}$ partially ordered under inclusion, an antichain is precisely a Sperner family (and now recall that we mentioned that "antichain" is another word for Sperner family.)

If $(X, \leq)$ is finite, we define the **chain number** $c(X)$ to be the smallest size of a partition of X into chains. (If X is infinite we could define $c(X)$ as a cardinal number in the sense of Chapter 11, but in this section we will only work with finite partially ordered sets.) Further, if X is a finite partially ordered set and $x \in X$, we define the **height of** x $h(x)$ to be the largest size of a chain $\mathcal{C}$ of X in which x is the largest element. Similarly we define the **height of** X $h(X)$ to be the largest size of a chain of X. In Exercise 9.35 you are asked to show that

$$h(X) = \max_{x \in X} h(x).$$

Still when $(X, \leq)$ is finite, we define the **width** $w(X)$ to be the largest size of an antichain in X: i.e., a subset $\mathcal{F} \subset X$ such that for no $x, y \in \mathcal{F}$ do we have $x < y$. Finally, we define the **antichain number** $\mathrm{ac}(X)$ of a finite partially ordered set $(X, \leq)$ as the smallest size of a partition of X into antichains.

EXAMPLE 9.77. *Let $n \in \mathbb{Z}^+$, and let $X = 2^{[n]}$, partially ordered under inclusion. For $A \subseteq [n]$, the height $h(A)$ is $\#A + 1$: indeed, we can make a chain from the empty set to A, adding one element at any stage; and conversely, any set in a chain of sets has at least one more element than the last element of the chain, so these elements are maximal. The height $h(2^{[n]})$ is therefore $h([n]) = n+1$.*[6] *By Sperner's Theorem, the width $w(2^{[n]})$ and the chain number $c(2^{[n]})$ are both equal to $\binom{[n]}{\lfloor \frac{n}{2} \rfloor}$.*

We claim that the antichain number $\mathrm{ac}(2^{[n]})$ is equal to $h(2^{[n]}) = n+1$. First, for all $0 \leq k \leq n$, as we know the set $\binom{[n]}{k}$ of k-element subsets of $2^{[n]}$ is an antichain, which gives a partition of $2^{[n]}$ into $n+1$ antichains, so $\mathrm{ac}(2^{[n]}) \leq n+1$. Moreover, the argument of Theorem 9.74 can be adapted to show that $h(2^{[n]}) \leq \mathrm{ac}(2^{[n]})$: indeed, given any chain $\mathcal{C}$ and any partition $\mathcal{P}$ of $2^{[n]}$ into antichains, the natural map $\iota : \mathcal{C} \hookrightarrow \mathcal{P}$ that sends each set A in $\mathcal{C}$ into the unique antichain containing it must be an injection, since otherwise two subsets of $[n]$ belong to both a chain and to antichain, which is impossible. Thus

$$n + 1 = h(2^{[n]}) \leq \mathrm{ac}(2^{[n]}),$$

and it follows that $\mathrm{ac}(2^{[n]}) = n+1$.

Now we have the following result, a vast generalization of Corollary 9.76.

THEOREM 9.78 (Dilworth [**Di50**]). *Let $(X, \leq)$ be a finite partially ordered set. Then $w(X) = c(X)$.*

PROOF. Step 1: If $\mathcal{F}$ is an antichain in X and $\mathcal{P}$ is a partition of X into chains, then the map $\iota : \mathcal{F} \to \mathcal{P}$ that maps each $x \in \mathcal{F}$ to the unique element of $\mathcal{P}$ that contains it must be an injection, so $\#\mathcal{F} \leq \#\mathcal{P}$. It follows that $w(X) \leq c(X)$.
Step 2: It remains to find a partition of X into $w(X)$ chains, since this will show that $c(X) \leq w(X)$ and thus after Step 1 that $w(X) = c(X)$. We will do this by strong induction on the size $\#X$ of X. The base case is $\#X = 0$, i.e., $X = \varnothing$, in which case $a = 0$ so we may take the empty partition.

So suppose that X is nonempty and that the conclusion of the theorem holds for all partially ordered sets of size smaller than $\#X$. The basic idea of the proof is to remove a single chain $\mathcal{C}$ from X in such a way as to make the width drop: then, inductively, the chain number of $X \setminus \mathcal{C}$ is equal to the width of $X \setminus \mathcal{C}$ which is smaller than the width of X, so $X \setminus \mathcal{C}$ is a disjoint union of fewer than $w(X)$ chains and thus adding back $\mathcal{C}$ we get that X is a disjoint union of at most $w(X)$ chains.

To implement this, fix a maximal element x_M of X: i.e., an element that is not strictly less than any element of X. At least one such element must exist because X is finite and nonempty. Put

$$X' \coloneqq X \setminus \{x_M\}$$

Then X' is a partially ordered set with the restricted relation $\leq$. As mentioned above, the favorable case is if $w(X') < w(X)$, because then by induction we can cover X' with fewer than $w(X)$ chains and then adding back $\{x_M\}$ we can cover X with $w(X)$ chains. So we may suppose that $w(X') = w(X)$. Since $\#X' < \#X$, there is a decomposition of X' into $w(X)$ chains, say $\mathcal{C}_1, \ldots, \mathcal{C}_{w(X)}$. Here is a key

[6]The $+1$'s appearing in these formulas explain why a common alternate convention is to define the height of an element and of a partially ordered set by subtracting 1 from our definition. If we write a finite chain as $x_0 < x_1 < \ldots < x_h$ then this alternate definition of the height gives the number of $<$'s, or the number of "links in the chain."

observation: if $\mathcal{F}$ is any antichain in X' of size $w(X)$, then it must have nonempty intersection with every chain $\mathcal{C}_i$: indeed, the map $\iota : \mathcal{F} \to \{\mathcal{C}_1, \ldots, \mathcal{C}_{w(X)}\}$ is an injective function between two finite sets of size $w(X)$, so it must also be surjective. Now for each $1 \leq i \leq w(X)$, let y_i be the largest element of the chain $\mathcal{C}_i$ that lies in some antichain of maximal length $w(X)$, and let

$$\mathcal{F} \coloneqq \{y_1, \ldots, y_{w(X)}\}.$$

We claim that $\mathcal{F}$ is an antichain in X'. Indeed, if not there are distinct i and j such that $y_i < y_j$. Let $\mathcal{F}_j$ be an antichain of length $w(X)$ that includes the element y_j. As mentioned above, the antichain $\mathcal{F}_j$ must contain some element z_i of $\mathcal{C}_i$, and by definition of y_i we must have $z_i \leq y_i$. So then z_i and z_j both lie in $\mathcal{F}_j$ and $z_i \leq y_i < z_j$, a contradiction.

We claim that we must have $y_i \leq x_M$ for at least one i: if not, then $\{y_1, \ldots, y_{w(X)}, x_M\}$ would be an antichain in X of size $w(X)+1$, a contradiction. So fix an i such that $y_i \leq x_M$, and let $\mathcal{K}$ be the chain consisting of x_M and all elements of $\mathcal{C}_i$ that are less than or equal to x_M. Then $X \setminus \mathcal{K}$ does *not* have an antichain of size $w(X)$: such an antichain would be an antichain of size $w(X)$ in X' that does not contain any element of $\mathcal{C}_i$ that is less than or equal to y_i, and by definition of y_i it cannot contain any element of $\mathcal{C}_i$ that is greater than y_i. So we have succeeded in removing one chain $\mathcal{K}$ from X so as to decrease the width of X, which as mentioned above implies inductively that X can be partitioned into $w(X)$ chains. □

4.3. Mirsky's Theorem. Comparing Example 9.77 with Dilworth's Theorem, we quickly find that in any finite partially ordered set $(X, \leq)$ we have $h(X) \leq \mathrm{ac}(X)$. This suggests investigating when equality occurs. Somewhat curiously, the answer is "always" and the argument for this is significantly *easier* than for Sperner's Theorem, but the result came more than twenty years later.

THEOREM 9.79 (Mirsky [**Mi71**]). *Let $(X, \leq)$ be a finite partially ordered set. Then $h(X) = \mathrm{ac}(X)$.*

It is remarkable how similar the proof of Theorem 9.79 is to the special case $X = 2^{[n]}$ of Example 9.77. You are asked to carry this argument over in Exercise 9.36.

5. Exercises

EXERCISE 9.1. *Let X be a finite nonempty set, and let $f : X \to X$ be a map. If the esential image $f^\infty(X)$ consists of a single point x, show that $\mathfrak{c}_f(x) = X$ and $f(x) = x$.*

EXERCISE 9.2. *Let $N \in \mathbb{Z}^+$.*

a) *Show that multiplication in $\mathbb{Z}/N\mathbb{Z}$ is commutative.*
b) *Show that multiplication in $\mathbb{Z}/N\mathbb{Z}$ is associative.*
c) *Show that $1 \pmod N$ is an identity element for multiplication in $\mathbb{Z}/N\mathbb{Z}$: that is, for all $X \in \mathbb{Z}/N\mathbb{Z}$ we have $(1 \pmod N) \cdot X = X$.*
(d) *Check the distributive property in $\mathbb{Z}/N\mathbb{Z}$: for all $X, Y, Z \in \mathbb{Z}/N\mathbb{Z}$ we have $(X + Y) \cdot Z = (X \cdot Z) + (Y \cdot Z)$.*

EXERCISE 9.3. *Show that for any integers $0 < u < v$, we have that $(v^2 - u^2, 2uv, v^2 + u^2)$ is a Pythagorean triple.*

EXERCISE 9.4. *Let $a, b \in \mathbb{Z}$ and $N \in \mathbb{Z}^+$. Show: if $a \equiv b \pmod N$ then for all $n \in \mathbb{Z}^+$ we have $a^n \equiv b^n \pmod N$.*
(Suggestion: use induction.)

EXERCISE 9.5. *In this section we explore versions of the Chinese Remainder Theorem for positive integers N_1, N_2 that need not be coprime.*

a) *Consider the map*

$$\Phi : \mathbb{Z}/N_1N_2\mathbb{Z} \to \mathbb{Z}/N_1\mathbb{Z}\times\mathbb{Z}/N_2\mathbb{Z},\ x \pmod{N_1N_2} \mapsto (x \pmod{N_1}, x \pmod{N_2}).$$

(i) *Show that for $x, y \in \mathbb{Z}$, we have $\Phi(x \pmod{N_1N_2}) = \Phi(y \pmod{N_1N_2})$ if and only if $x \equiv y \pmod{\operatorname{lcm}(N_1, N_2)}$.*
(ii) *Deduce that every nonempty fiber of Φ has size $\gcd(N_1, N_2)$.*
(iii) *Show that $\Phi(\mathbb{Z}/N_1N_2\mathbb{Z})$ consists of pairs $(a \pmod{N_1}, b \pmod{N_2})$ such that $a \pmod{\gcd(N_1, N_2)} = b \pmod{\gcd(N_1, N_2)}$.*
(iv) *Deduce that $\#\Phi(\mathbb{Z}/N_1N_2\mathbb{Z}) = \operatorname{lcm}(N_1, N_2)$.*

b) *In view of part a), it may be the case that a cleaner generaliation is obtained as follows: consider the map*

$$\Psi : \mathbb{Z}/\operatorname{lcm}(N_1, N_2)\mathbb{Z} \to \mathbb{Z}/N_1\mathbb{Z}\times\mathbb{Z}/N_2\mathbb{Z},\ x \pmod{\operatorname{lcm}(N_1, N_2)} \mapsto (x \pmod{N_1}, x \pmod{N_2}).$$

(i) *Show: Ψ is injective.*
(ii) *Show $\Psi(\mathbb{Z}/\operatorname{lcm}(N_1N_2)\mathbb{Z}) = \Phi(\mathbb{Z}/N_1N_2\mathbb{Z})$.*

EXERCISE 9.6. *Let $N_1, N_2 \in \mathbb{Z}^+$, and let $a, b \in \mathbb{Z}$. Show that the following are equivalent:*

(i) *There is $c \in \mathbb{Z}$ such that $c \equiv a \pmod{N_1}$ and $c \equiv b \pmod{N_2}$.*
(ii) *We have $a \pmod{\gcd(N_1, N_2)} = b \pmod{\gcd(N_1, N_2)}$.*

EXERCISE 9.7. *Prove Theorem 9.17.*
(Suggestion for part a): use induction on k. Suggestion for part b): use the fact that since $N_1, \ldots, N_k$ are pairwise coprime, we have $\operatorname{lcm}(N_1, \ldots, N_k) = N_1 \cdots N_k$.)

EXERCISE 9.8. *Let $N_1, \ldots, N_r$ be pairwise coprime positive integers. Show:*

$$\varphi(N_1 \cdots N_r) = \varphi(N_1) \cdots \varphi(N_r).$$

EXERCISE 9.9. *Prove Euler's generalization of Fermat's Little Theorem: let $N \in \mathbb{Z}^+$, and let $x \in \mathbb{Z}$ be coprime to N. Show:*

$$x^{\varphi(N)} \equiv 1 \pmod N.$$

(Suggestion: adapt the proof we have given of Fermat's Little Theorem.)

EXERCISE 9.10. *Let $p \equiv 3 \pmod 4$ be a prime number.*

a) *Show: $(\frac{p-1}{2})! \equiv \pm 1 \pmod p$.*
b) *For each prime $p \equiv 3 \pmod 4$ with $p < 100$, determine whether $\frac{p-1}{2}!$ is 1 or -1 modulo p. Note in particular that both can occur.*
c) *Do you have a conjecture as to for which primes $p \equiv 3 \pmod 4$ we have $\frac{p-1}{2}! \equiv 1 \pmod p$?*
(This is known, but this time the answer not easy to state: see [**Mo61**]*.)*

EXERCISE 9.11. *Show: for all $a, b \in \mathbb{Z}^{\geq 2}$ we have*

$$R(a, b) \leq \binom{a+b-2}{a-1}.$$

(Suggestion: Go by Strong Induction on $a + b$, using (58) and Proposition 3.9.)

EXERCISE 9.12. *Let $a, b \in \mathbb{Z}^{\geq 3}$. Suppose that $R(a-1,b)$ and $R(a,b-1)$ are both even. Show:*

$$R(a,b) \leq R(a-1,b) + R(a,b-1) - 1.$$

EXERCISE 9.13.

a) *Show that $R(3,4) \leq 9$.*
b) *Construct a graph with 8 vertices that has neither a clique of order 3 nor an independent set of order 4.*
c) *Deduce: $R(3,4) = 9$.*

EXERCISE 9.14.

a) *Show that $R(4,4) \leq 18$.*
b) *The **Paley graph** of order 17, say P_{17} is a graph with vertex set $\{0,1,\ldots,16\}$ and for which vertex i is adjacent to vertex j if and only if ($i \neq j$ and there is $x \in \mathbb{Z}$ such that $(i-j) - x^2$ is a multiple of 17). Show[7] that P_{17} has neither a clique of order 4 nor an independent set of order 4.*
c) *Deduce: $R(4,4) = 18$.*

EXERCISE 9.15. *Suppose that $R(a_1,\ldots,a_k)$ is finite. Show that it does not depend upon the ordering of $a_1,\ldots,a_k$: that is, if $\sigma : [k] \to [k]$ is a bijection, then $R(a_1,\ldots,a_k) = R(\sigma(a_1),\ldots,\sigma(a_k))$.*

EXERCISE 9.16. *Let $1 \leq a_1 \ldots \leq a_k$ be integers.*

a *) Show: if $a_1 = 1$, then $R(a_1,\ldots,a_k) = 1$.*
b) *Show: if $a_1 = 2$, then $R(a_1,\ldots,a_k) = R(a_2,\ldots,a_k)$.*

EXERCISE 9.17. *Let $r \in \mathbb{Z}^{\geq 3}$. A **Schur r-tuple** is an r-tuple of positive integers $(x_1,\ldots,x_{r-1},x_r)$ such that $x_1 + \ldots + x_{r-1} = x_r$.[8] We define the **generalized Schur number** $S_r(k)$ to be the least positive integer N such that any k-coloring $\mathfrak{c}$ of $[N] = \{1,\ldots,N\}$ admits a monochromatic Schur r-tuple: i.e., there are integers $x_1,\ldots,x_r \in [N]$ such that $x_1 + \ldots + x_{r-1} = x_r$ and $\mathfrak{c}(x_1) = \ldots = \mathfrak{c}(x_r)$. Show:*

$$S_r(k) \geq R(r,\ldots,r(k \text{ times})) - 1.$$

(Suggestion: adapt the proof of Theorem 9.54.)

EXERCISE 9.18. *Let $n \in \mathbb{Z}^{\geq 2}$, and let G be a graph with vertex set $[n]$. Show: at least one of G and its complement $\overline{G}$ are connected.*

EXERCISE 9.19.

a) *Show: if a graph admits a circuit of odd length, then it also admits a cycle of odd length.*
(Suggestion: proceed by induction on the length.)
b) *Show: if a graph has an edge, then it has a circuit of length n for all even $n \in \mathbb{N}$.*

EXERCISE 9.20. *Let $G = (V,E)$ be a graph. A **bipartition** of G is a partition $\mathcal{P} = \{V_1, V_2\}$ of the vertex set V of G into two parts, such that every edge $e \in E$ is of the form $e = \{v_1, v_2\}$ with $v_1 \in V_1$ and $v_2 \in V_2$. (In other words, a bipartition writes the vertex set as a disjoint union of two nonempty subsets V_1 and V_2 such*

[7]This is a result of Greenwood and Gleason [**GG55**]. Paley graphs can be defined for any prime $p = 4k+1$, and 4-element cliques in such graphs were studied by Evans, Pulham and Sheehan [**EPS81**].

[8]Thus a Schur 3-tuple is nothing else than a Schur triple...fortunately.

that no edge runs between two vertices of V_1 or between two vertices of V_2.) A graph is ***bipartite*** *if it admits a bipartition.*

a) *Show: a graph is bipartite if and only if each of its connected components is bipartite.*

b) *Let G be a graph with no cycles of odd length, let $v, w \in V$, and let ℓ_1, ℓ_2 be two walks from v to w in G. Show: the length of ℓ_1 has the same parity as the length of ℓ_2.*
(Suggestion: use Exercise 9.19a).)

c) *Show: a graph is bipartite if and only if it has no cycle of odd length.*
(For one direction, it suffices to show that a cycle of odd length is not bipartite, which is straightforward. For the other direction, suppose that G admits no cycle of odd length. By part a), we may assume that G is connected. Fix a vertex v_0 of G. Let V_0 be the set of all vertices v for which there is a path of even length from v_0 to v, and let V_1 be the set of all vertices v for which there is a path of odd length from v_0 to v. Show: $\{V_0, V_1\}$ is a bipartition of G.)

EXERCISE 9.21. *Show: a finite tree with exactly two pendant vertices is a path.*

EXERCISE 9.22. *A* ***forest*** *is a graph that has no cycles.*

a) *Show: a graph is a forest if and only if each of its connected components is a tree.*

b) *Show: the Euler characteristic of a finite forest is the number of connected components.*

EXERCISE 9.23. *Let $G = (V, E)$ be a finite graph, Let $V_0 \subseteq V$ be the subset of isolated vertices.*

a) *Show: G admits an Eulerian walk if and only if $G^\circ := (V \setminus V_0, E)$ admits an Eulerian walk.*

b) *Show: if a graph without isolated vertices admits an Eulerian walk, it is connected.*

c) *Suppose G admits an Eulerian walk $W : x_0, \ldots, x_n$ that is* not *an Eulerian circuit. Show: the initial and final vertices x_0 and x_n each have odd degree, and every other vertex has even degree.*

d) *Let G be a finite connected graph for which there are vertices $v \neq w$ of odd degree, whereas every other vertex has even degree. Show: G admits an Eulerian walk starting at v and ending at w.*
(Suggestion: consider the graph $\tilde{G}$ obtained from G by adjoining a vertex v_∞ not in V and two edges $e_v = \{v_\infty, v\}$ and $e_w = \{v_\infty, w\}$. Show that $\tilde{G}$ admits an Eulerian circuit and deduce that G admits an Eulerian walk from v to w.)

EXERCISE 9.24. *We sketch a* topological *derivation of the infinite case of Theorem 9.59 from the finite case, following Halmos and Vaughan* [**HV50**].

a) *For each $x \in V_1$, let $S_x = N(\{x\})$ be the set of all vertices adjacent to x, a subset of V_2. By hypothesis, each S_x is finite. Endow each S_x with the discrete topology, and put the product topology on*

$$S := \prod_{x \in V_1} S_x.$$

Show: Tychonoff's Theorem [**Cl-GT**, Thm. 5.24] *implies S is compact.*

b) *For a finite subset $X \subseteq V_1$, put*

$$H_X \coloneqq \{s = \{s_x\} \in S \mid s_x \neq s_y \forall x \neq y \in X\}.$$

Show that the already proved finite case of Theorem 9.59 implies that H_X is nonempty.

c) *Show: H_X is closed in S.*

d) *Deduce: since S is compact, there is $s \in \bigcap_X H_X$.*

e) *Show: every element $s \in \bigcap_X H_X$ defines a semiperfect matching on G.*

EXERCISE 9.25. *Consider the following bipartitioned graph $G = (V_1, V_2, E)$: $V_1 = \mathbb{N} \times \{1\}$, $V_2 = \mathbb{Z}^+ \times \{2\}$. For each $n \in \mathbb{Z}^+$ there is an edge*

$$e_n = \{(n,1),(n,2)\},$$

and for all $m \in \mathbb{Z}^+$ there are edges

$$e_m \coloneqq \{(0,m)\}.$$

One can think of this graph as follows: for each positive integer n there is a man M_n and a woman W_n who want to marry each other. The man M_n only wants to marry the woman W_n, and the woman W_n does not want to marry any man M_m with $m \in \mathbb{Z}^+ \setminus \{n\}$. However, there is also another man M_0, whom every single woman W_n would like to marry.[9]

a) *Show that for every subset $X \subseteq V_1$, there is an injection $\iota : X \hookrightarrow N(X)$.*

b) *Show that nevertheless G admits no semiperfect matching.*

EXERCISE 9.26. *Let $G = (V, E)$ be a graph, and let $f : V \to V$ be a graph derangement. Show: there is a surjective graph derangement $g : V \to V$.*

EXERCISE 9.27. *Let $G = (V_1, V_2, E)$ be a locally finite bipartitioned graph. Show that the following are equivalent:*

(i) *The graph G admits a perfect matching.*

(ii) *The graph G admits a graph derangement.*

(iii) *For every finite subset $X \subseteq V$, we have $\#X \leq \#N(X)$.*

(iv) *For every finite subset X of either V_1 or V_2, we have $\#X \leq \#N(X)$.*

EXERCISE 9.28. *Show: there is a connected cubic graph $G = ([16], E)$ without a perfect matching.*

EXERCISE 9.29.

a) *Let G be a 2-regular graph. Show: every connected component of G is a cycle.*

b) *Deduce: a finite 2-regular graph $G = (V, E)$ admits a perfect matching if and only if $\#V$ is even.*

EXERCISE 9.30. *Prove the **Generalized Petersen Theorem:** Let $d \in \mathbb{Z}^{\geq 2}$, and let $G = (V, E)$ be a finite d-regular graph. Suppose that G is **(d-1)-edge-connected**: that is, the removal of fewer than $d-1$ edges does not disconnect G. If d is even, suppose moreover that $\#V$ is even. Show: G admits a perfect matching. (Suggestion: adapt the proof of Theorem 9.65.)*

EXERCISE 9.31. *Prove Theorem 9.74.*

EXERCISE 9.32. *Find a partition of $2^{[5]}$ into 10 chains.*[10]

[9] Ryan Gosling?

[10] The intent is for this exercise to be done *before* covering the proof of Sperner's Theorem.

EXERCISE 9.33.

a) *Complete the proof of Theorem 9.75a) by constructing the maps β_k. (Comment: it is possible to do this* either *by adapting the proof that constructs the α_k or using the α_k's to construct the β_k's.)*

b) *Complete the proof of Theorem 9.75B) by constructing the maps α_k, β_k and γ.*

EXERCISE 9.34. *Let $n, k \in \mathbb{Z}^+$ with $k \leq \frac{n}{2}$. Let $\mathcal{F}$ be a Sperner family of subsets of $[n]$ such that every element of $\mathcal{F}$ has size at most k. Show:*

$$\#\mathcal{F} \leq \binom{n}{k}.$$

EXERCISE 9.35. *Let $(X, \leq)$ be a finite, nonempty partially ordered set.*

a) *Show: $h(X) = \max_{x \in X} h(x)$.*

b) *Suppose that $x \in X$ is such that $h(x) = h(X)$. Show: x is a maximal element of X: i.e., there is no $y \in X$ such that $x <$.*

c) *Give an example of a finite partially ordered set $(X, \leq)$ with a maximal element x such that $h(x) < h(X)$.*

EXERCISE 9.36. *Let $(X, \leq)$ be a finite partially ordered set. For $x \in X$, we define the* ***height*** *$h(x)$ of x to be the maximal size of a chain in X for which x is the largest element. We define the* ***height*** *$h(X)$ of X to be the largest size of a chain in X. We define the* ***antichain number*** *$\mathrm{ac}(X)$ to be the smallest size of a partition of X into antichains.*

a) *Show that $h(X) \leq \mathrm{ac}(X)$. (Hint: the argument of Example 9.77 goes through verbatim.)*

b) *For $1 \leq i \leq h(X)$, let*

$$\mathcal{A}_i := \{x \in X \mid h(x) = i\}.$$

Show that each $\mathcal{A}_i$ is antichain.

c) *Prove Mirsky's Theorem: $h(X) = \mathrm{ac}(X)$.*

EXERCISE 9.37. *Let $(X, \leq)$ be a nonempty, finite partially ordered set.*

a) *Let $a, b \in \mathbb{Z}^+$. Show: if $\#X \geq ab + 1$, then X has either a chain of size $a + 1$ or an antichain of size $b + 1$. (This follows* either *from Dilworth's Theorem or from Mirsky's Theorem.)*

b) *Show the* ***Rectangular Law****: $\max(h(X), w(X)) \geq \sqrt{\#X}$.*

c) *Prove the* ***Erdős-Szekeres Theorem****: let $a, b \in \mathbb{Z}^+$. A finite sequence $(x_1, \ldots, x_{ab+1})$ of real numbers has either an increasing subsequence of size $a + 1$ or a decreasing subsequence of size $b + 1$. (Hint: define a partial ordering $\preceq$ on $X := [ab + 1]$ by: for $i, j \in [ab + 1]$, $i \preceq j$ if $i \leq j$ and $x_i \leq x_j$. Apply part a) to $(X, \preceq)$.)*

d) *Show that the result of part c) is* sharp *in the sense that for all $a, b \in \mathbb{Z}^+$, there is a finite real sequence of length ab with neither an increasing subsequence of length $a + 1$ nor an increasing subsequence of length $b + 1$.*

e) *Deduce that the result of part a) is* sharp *in the sense that for all $a, b \in \mathbb{Z}^+$, there is a partially ordered set X of size ab with neither a chain of size $a + 1$ nor an antichain of size $b + 1$.*

CHAPTER 10

Countable and Uncountable Sets

1. Introducing equivalence of sets, countable and uncountable sets

We assume known the set $\mathbb{Z}^+$ of positive integers, and the set $\mathbb{N} = \mathbb{Z}^+ \cup \{0\}$ of natural numbers. For any $n \in \mathbb{Z}^+$, we denote by $[n]$ the set $\{1, \ldots, n\}$. We take it as obvious that $[n]$ has n elements, and also that the empty set $\emptyset$ has 0 elements. Just out of mathematical fastidiousness,[1] let's define $[0] = \emptyset$ (why not?).

It is pretty clear what it means for an arbitrary set S to have 0 elements: it must be the empty set. That is – and this is a somewhat curious property of the empty set – $\emptyset$ as a set is uniquely characterized by the fact that it has 0 elements.

What does it mean for an arbitrary set S to have n elements? By definition, it means that there exists a bijection $\iota : S \to [n]$, i.e., a function which is both injective and surjective; or, equivalently, a function for which there exists an inverse function $\iota' : [n] \to S$.[2]

Let us call a set *finite* if it has n elements for some $n \in \mathbb{N}$, and a set *infinite* if it is not finite.

Certainly there are some basic facts that we feel should be satisfied by these definitions. For instance:

FACT 10.1. *The set $\mathbb{Z}^+$ is infinite.*

PROOF. The set $\mathbb{Z}^+$ certainly nonempty, so we would like to show that for no $n \in \mathbb{Z}^+$ is there a bijection $\iota : [n] \to \mathbb{Z}^+$. This seems obvious. Unfortunately, sometimes in mathematics we must struggle to show that the obvious is true (and sometimes what seems obvious is not true!). Here we face the additional problem of not having formally axiomatized things, so it's not completely clear what's "fair game" to use in a proof. But consider the following: does $\mathbb{Z}^+$ have one element? Absolutely not: for any function $\iota : [1] = \{1\} \to \mathbb{Z}^+$, ι is not surjective because it does not hit $\iota(1) + 1$. Does $\mathbb{Z}^+$ have two elements? Still, no: if ι is not injective, the same argument as before works; if ι is injective, its image is a 2 element subset of $\mathbb{Z}^+$. Since $\mathbb{Z}^+$ is totally ordered (indeed well-ordered), one of the two elements in the image is larger than the other, and then that element plus one is not in the image of our map. We could prove it for 3 as well, which makes us think we should probably work by induction on n. How to set it up properly? Let us try to show

[1] Well, not really: this will turn out to be quite sensible.

[2] I am assuming a good working knowledge of functions, injections, surjections, bijections and inverse functions. This asserts at the same time (i) a certain amount of mathematical sophistication, and (ii) a certain amount of metamathematical informality.

that for all n and all $\iota : [n] \to \mathbb{Z}^+$, there exists $N = N(\iota)$ such that $\iota([n]) \subseteq [N]$. If we can do this, then since $[N]$ is clearly a proper subset of $\mathbb{Z}^+$ (it does not contain $N+1$, and so on) we will have shown that for no n is there a surjection $[n] \to \mathbb{Z}^+$ (which is in fact stronger than what we claimed). But carrying through the proof by induction is now not obvious but (much better!) easy, so is left to the reader. □

What did we use about $\mathbb{Z}^+$ in the proof? Some of the Peano axioms for $\mathbb{Z}^+$, most importantly that it satisfies the principle of mathematical induction (PMI). Since it is hard to imagine a rigorous proof of a nontrivial statement about $\mathbb{Z}^+$ that does not use PMI, this is a good sign: things are proceeding well so far.

What about $\mathbb{Z}$: is it too infinite? It should be, since it contains an infinite subset. This is logically equivalent to the following fact:

FACT 10.2. *A subset of a finite set is finite.*

PROOF. More concretely, it suffices to show that for any $n \in \mathbb{N}$ and and subset $S \subseteq [n]$, then for some $m \in \mathbb{N}$ there exists a bijection $\iota : S \to [m]$. As above, for any specific value of n, it straightforward to show this, so again we should induct on n. Let's do it this time: assume the statement for n, and let $S \subseteq [n+1]$. Put $S' = S \cap [n]$, so by induction there exists a bijection $\iota' : [m] \to S'$ for some $m' \in \mathbb{N}$. Composing with the inclusion $S' \subseteq S$ we get an injection $\iota : [m] \to S$. If $n+1$ is not an element of S, then $S' = S$ and ι is a bijection. If $n+1 \in S$, then extending ι' to a map from $[m+1]$ to S by sending $m+1$ to $n+1$ gives a bijection. □

Again, by contraposition this shows that many of our most familiar sets of numbers – e.g. $\mathbb{Z}$, $\mathbb{Q}$, $\mathbb{R}$, $\mathbb{C}$ – are infinite.

Let us press on to study the properties of *infinite* sets.

Basic Definition (Cantor): We say that S and T as *equivalent*, and write $S \cong T$ if there exists a bijection $\iota : S \to T$.

Historical Remark: When there exists a bijection between S and T, Cantor first said that S and T have the same *power*.[3] As is often the case in mathematics, this forces us to play a linguistic-grammatical game – given that a definition has been made to have a certain part of speech, write down the cognate words in other parts of speech.[4] Thus a faithful rendition of Cantor's definition in adjectival form would be something like *equipotent*. The reader should be warned that it would be more common to use the term *equinumerous* at this point.

However, we have our reasons for choosing to use "equivalent." The term "equinumerous," for instance, suggests that the two sets have the same number of elements, or in other words that there is some numerical invariant we are attaching to a single set with the property that two sets can be put in bijection exactly when both have the same value of this numerical invariant. But we would like to view things in exactly the opposite way. Let us dilate a bit on this point.

It was Cantor's idea that we should regard two sets as "having the same size"

[3]Or rather, he said something in German that gets translated to this. Such pedantic remarks will be omitted from now on!

[4]This is a game that some play better than others, viz.: generization, sobrification, unicity.

if and only if they are equivalent, i.e., if and only if their elements can be paired off via a one-to-one correspondence. Certainly this is consistent with our experience from finite sets. There is, however, a brilliant and subtle twist: colloquially one thinks of counting or measuring something as a process which takes as input one collection of objects and outputs a "number." We therefore have to have names for all of the "numbers" which measure the sizes of things: if you like, we need to count arbitrarily high. Not every civilization has worked out such a general counting scheme: I have heard tell that in a certain "primitive tribe" they only have words for numbers up to 4 and anything above this is just referred to as "many." Indeed we do not have proper names for arbitrarily large numbers in the English language (except by recourse to iteration, e.g., million million for a trillion).

But notice that we do not have to have such an elaborate "number knowledge" to say whether two things have the same size or not. For instance, one may presume that shepherding predates verbal sophistication, so the proto-linguistic shepherd needs some other means of making sure that when he takes his sheep out to graze in the countryside he returns with as many as he started with. The shepherd can do this as follows: on his first day on the job, as the sheep come in, he has ready some sort of sack and places stones in the sack, one for each sheep. Then in the future he counts his sheep, not in some absolute sense, but in relation to these stones. If one day he runs out of sheep before stones, he knows that he is missing some sheep (at least if he has only finitely many sheep!).

Even today there are some situations where we test for equivalence rather than count in an absolute sense. For instance, if you come into an auditorium and everyone is sitting in a (unique!) seat then you know that there are at least as many seats as people in the room without counting both quantities.

What is interesting about infinite sets is that these sorts of arguments break down: the business of taking away from an infinite set becomes much more complicated than in the finite case, in which, given a set S of n elements and any element $x \in S$, then $S \setminus x$ has $n-1$ elements. On the other hand, $\mathbb{Z}^+$ and $\mathbb{N}$ are equivalent, since the map $n \mapsto n-1$ gives a bijection between them. Similarly $\mathbb{Z}^+$ is equivalent to the set of even integers ($n \mapsto 2n$). Indeed, we soon see that much more is true:

FACT 10.3. *For any infinite subset $S \subseteq \mathbb{Z}^+$, S and $\mathbb{Z}^+$ are equivalent.*

PROOF. Using the fact that $\mathbb{Z}^+$ is well-ordered, we can define a function from S to $\mathbb{Z}^+$ by mapping the least element s_1 of S to 1, the least element s_2 of $S \setminus \{s_1\}$ to 2, and so on. If this process terminates after n steps then S has n elements, so is finite, a contradiction. Thus it goes on forever and clearly gives a bijection. □

It is now natural to wonder which other familiar infinite sets are equivalent to $\mathbb{Z}^+$ (or $\mathbb{N}$). For this, let's call a set equivalent to $\mathbb{Z}^+$ *countable*.[5] A slight variation of the above argument gives

FACT 10.4. *Every infinite set has a countable subset.*

PROOF. For an infinite set S, just keep picking elements to define a bijection from $\mathbb{Z}^+$ to some subset of S; we can't run out of elements since S is infinite! □

[5]Perhaps more standard is to say "countably infinite and reserve "countable" to mean countably infinite or finite. Here we suggest simplifying the terminology.

As a first example:

FACT 10.5. *The two sets $\mathbb{Z}$ and $\mathbb{Z}^+$ are equivalent.*

PROOF. We define an explicit bijection $\mathbb{Z} \to \mathbb{Z}^+$ as follows: we map $0 \mapsto 1$, then $1 \mapsto 2$, $-1 \mapsto 3$, $2 \mapsto 4$, $-2 \mapsto 5$ and so on. □

The method proves something more general, a "splicing" result.

FACT 10.6. *Suppose that S_1 and S_2 are two countable sets. Then $S_1 \bigcup S_2$ is countable.*

Indeed, we can make a more general splicing construction:

FACT 10.7. *Let $\{S_i\}_{i \in I}$ be an indexed family of pairwise disjoint nonempty sets; assume that I and each S_i is at most countable (countable or finite). Then $S := \bigcup_{i \in I} S_i$ is at most countable. Moreover, S is finite if and only if I and all the S_i are finite.*

PROOF. We sketch the construction: since each S_i is at most countable, we can order the elements as s_{ij} where either $1 \leq j \leq \infty$ or $1 \leq j \leq N_j$. If everything in sight is finite, then S will be finite (a finite union of finite sets is finite). Otherwise, we define a bijection from $\mathbb{Z}^+$ to S as follows: $1 \mapsto s_{11}$, $2 \mapsto s_{12}$, $3 \mapsto s_{22}$, $4 \mapsto s_{13}$, $5 \mapsto s_{23}$, $6 \mapsto s_{33}$, and so on. Here we need the convention that when s_{ij} does not exist, we omit that term and go on to the next element in the codomain. □

Fact 10.7 is used very often in mathematics. As one immediate application:

FACT 10.8. *The set of rational numbers $\mathbb{Q}$ is countable.*

PROOF. Each nonzero rational number α can be written uniquely as $\pm\frac{a}{b}$, where $a,\ b \in \mathbb{Z}^+$. We define the height $h(\alpha)$ of α to be $\max a, b$ and also $h(0) = 0$. It is clear that for any height $n > 0$, there are at most $2n^2$ rational numbers of height n,[6] and also that for every $n \in \mathbb{Z}^+$ there is at least one rational number of height n, namely the integer $n = \frac{n}{1}$. Therefore taking $I = \mathbb{N}$ and putting some arbitrary ordering on the finite set of rational numbers of height n, Fact 10.7 gives us a bijection $\mathbb{Z}^+ \to \mathbb{Q}$. □

In a similar way, one can prove that the set $\overline{\mathbb{Q}}$ of algebraic numbers is countable.

FACT 10.9. *If A and B are countable, then the Cartesian product $A \times B$ is countable.*

The buck stops with $\mathbb{R}$. Let's first prove the following theorem of Cantor, which is arguably the single most important result in set theory. Recall that for a set S, its power set 2^S is the set of all subsets of S.

THEOREM 10.10. *(First Fundamental Theorem of Set Theory)*
There is no surjection from a set S to its power set 2^S.

Remark: When S is finite, this is just saying that for all $n \in \mathbb{N}$, $2^n > n$, which is, albeit true, not terribly exciting. On the other hand, taking $S = \mathbb{Z}^+$ Cantor's Theorem provides us with an uncountable set $2^{\mathbb{Z}^+}$. In fact it tells us much more than this, as we shall see shortly.

[6] I will resist the temptation to discuss how to replace the 2 with an asymptotically correct constant.

PROOF. Suppose that $f : S \to 2^S$ is any function. We will produce an element of 2^S which is not in the image of f. Namely, let T be the set of all $x \in S$ such that x is not an element of $f(x)$, so T is some element of 2^S. Could it be $f(s)$ for some $s \in S$? Well, suppose $T = f(s)$ for some $s \in S$. We ask the innocent question, "Is $s \in T$?" Suppose first that it is: $s \in T$; by definition of T this means that s is not an element of $f(s)$. But $f(s) = T$, so in other words s is not an element of T, a contradiction. Okay, what if s is not in T? Then $s \in f(s)$, but again, since $f(s) = T$, we conclude that s *is* in T. In other words, we have managed to define, in terms of f, a subset T of S for which the notion that T is in the image of f is logically contradictory. So f is not surjective! □

What does this have to do with $\mathbb{R}$? Let us try to show that the interval $(0, 1]$ is uncountable. By Fact 10.3 this implies that $\mathbb{R}$ is uncountable. Now using binary expansions, we can identify $(0, 1]$ with the power set of $\mathbb{Z}^+$. Well, almost: there is the standard slightly annoying ambiguity in the binary expansion, that

$$.a_1a_2a_3 \cdots a_n01111111111 \ldots = .a_1a_2a_2 \cdots a_n1000000000 \ldots .$$

There are various ways around this: for instance, suppose we agree to represent every element of $(0, 1]$ by an element which does not terminate in an infinite string of zeros. Thus we have identified $(0, 1]$ with a certain subset T of the power set of $\mathbb{Z}^+$, the set of *infinite* subsets of $\mathbb{Z}^+$. But the set of finite subsets of $\mathbb{Z}^+$ is countable (Fact 10.7 again), and since the union of two countable sets would be countable (and again!), it must be that T is uncountable. Hence so is $(0, 1]$, and so is $\mathbb{R}$.

There are many other proofs of the uncountability of $\mathbb{R}$. For instance, we could contemplate a function $f : \mathbb{Z}^+ \to \mathbb{R}$ and, imitating the proof of Cantor's theorem, show that it cannot be surjective by finding an explicit element of $\mathbb{R}$ not in its image. We can write out each real number $f(n)$ in its decimal expansion, and then construct a real number $\alpha \in [0, 1]$ whose nth decimal digit α_n is different from the nth decimal digit of $f(n)$. Again the ambiguity in decimal representations needs somehow to be addressed: here we can just stay away from 9's and 0's. Details are left to the reader.

The above was just one example of the importance of distinguishing between countable and uncountable sets. Let me briefly mention some other examples:

EXAMPLE 10.11. *(Measure theory) A measure is a $[0, \infty]$-valued function defined on a certain family of subsets of a given set; it is required to be countably additive but not uncountably additive. For instance, this gives us a natural notion of size on the unit circle, so that the total area is π and the area of any single point is 0. The whole can have greater measure than the sum of the measures of the parts if there are uncountably many parts!*

EXAMPLE 10.12. *Given a differentiable manifold M of dimension n, then any submanifold of dimension $n-1$ has, in a sense which is well-defined independent of any particular measure on M, measure zero. In particular, one gets from this that a countable family of submanifolds of dimension at most $n-1$ cannot "fill out" an n-dimensional manifold. In complex algebraic geometry, such stratifications occur naturally, and one can make reference to a "very general" point on a variety as a point lying on the complement of a (given) countable family of lower-dimensional subvarieties, and be confident that such points exist!*

EXAMPLE 10.13. *Model theory is a branch of mathematics that often exploits the distinction between countable and uncountable in rather sneaky ways. Namely, there is the Löwenheim-Skolem theorem, which states in particular that any theory (with a countable language) that admits an infinite model admits a countable model. Moreover, given any uncountable model of a theory, there is a countable submodel which shares all the same "first order" properties, and conversely the countable/uncountable dichotomy is a good way to get an intuition on the difference between first-order and second-order properties.*

2. Some further basic results

2.1. Dedekind's characterization of infinite sets.

FACT 10.14. *A set S is infinite if and only if it is equivalent to a proper subset of itself.*

PROOF. One direction expresses an obvious fact about finite sets. Conversely, let S be an infinite set; as above, there is a countable subset $T \subseteq S$. Choose some bijection ι between T and $\mathbb{N}$. Then there is a bijection ι' between $T' := T \setminus \iota^{-1}(0)$ and T (just because there is a bijection between $\mathbb{N}$ and $\mathbb{Z}^+$. We therefore get a bijection between $S' := S \setminus \iota^{-1}(0$ and S by applying ι' from T' to T and the identity on $S \setminus T$. □

This characterization of infinite sets is due to Dedekind. What is ironic is that in some sense it is cleaner and more intrinsic than our characterization of finite sets, in which we had to compare against a distinguished family of sets $\{[n] \mid n \in \mathbb{N}\}$. Thus perhaps we should define a set to be finite if it cannot be put in bijection with a proper subset of itself! (On the other hand, this is not a "first order" property, so is not in reality that convenient to work with.)

2.2. An uncountable set not of continuum type. Notice that in making the definition "uncountable," i.e., an infinite set which is not equivalent to $\mathbb{Z}^+$, we have essentially done what we earlier made fun of the "primitive tribes" for doing: giving up distinguishing between very large sets. In some sense, set theory begins when we attempt to classify uncountable sets up to equivalence. This turns out to be quite an ambitious project – we will present the most basic results of this project in the next installment – but there are a few further facts that one should keep in mind throughout one's mathematical life.

Let us define a set S to be *of continuum type* (or, more briefly, a continuum[7]) if there is a bijection $\iota : S \to \mathbb{R}$. One deserves to know the following:

FACT 10.15. *There exists an uncountable set not of continuum type, namely $2^{\mathbb{R}}$.*

PROOF. By Theorem 10.10 there is no surjection from $\mathbb{R}$ to $2^{\mathbb{R}}$, so $2^{\mathbb{R}}$ is certainly not of continuum type. We must however confirm what seems intuitively plausible: that $2^{\mathbb{R}}$ is indeed uncountable. It is certainly infinite, since via the natural injection $\iota : \mathbb{R} \to 2^{\mathbb{R}}$, $r \mapsto \{r\}$, it contains an infinite subset. But indeed, this also shows that $2^{\mathbb{R}}$ is uncountable, since if it were countable, its subset $\iota(\mathbb{R}) \cong \mathbb{R}$ would be countable, which it isn't. □

[7]This has a different meaning in general topology, but no confusion should arise.

2.3. Some sets of continuum type. For any two sets S and T, we define T^S as the set of all functions $f : S \to T$. When $T = [2]$, the set of all functions $f : S \to [2]$ is naturally identified with the power set 2^S of S (so the notation is *almost* consistent: for full consistency we should be denoting the power set of S by $[2]^S$, which we will not trouble ourselves to do).

FACT 10.16. *The sets* $(0,1]$, $2^{\mathbb{Z}^+}$ *and* $\mathbb{R}^{\mathbb{Z}^+}$ *are of continuum type.*

PROOF. Earlier we identified the unit interval $(0,1]$ in $\mathbb{R}$ with the infinite subsets of $\mathbb{Z}^+$ and remarked that, since the finite subsets of $\mathbb{Z}^+$ form a countable set, this implies that $(0,1]$ hence $\mathbb{R}$ itself is uncountable. □

Let us refine this latter observation slightly:

LEMMA 10.17. *Let S be an uncountable set and $C \subseteq S$ an at most countable subset. Then $S \setminus C \cong S$.*

PROOF. Suppose first that C is finite, say $C \cong [n]$. Then there exists an injection $\iota : \mathbb{Z}^+ \to S$ such that $\iota([n]) = C$ (as follows immediately from Fact 6). Let $C_\infty = \iota(\mathbb{Z}^+)$. Now we can define an explicit bijection β from $S \setminus C$ to S: namely, we take β to be the identity on the complement of C_∞ and on C_∞ we define $\beta(\iota(k)) = \iota(k-n)$.

Now suppose C is countable. We do something similar: taking $C_1 = C$, since $S \setminus C_1$ is uncountable, we can find a countably infinite subset $C_2 \subseteq S \setminus C_1$. Proceeding in this way we can find a family $\{C_i\}_{i \in \mathbb{Z}^+}$ of pairwise disjoint countable subsets of S. Let us identify each of these subsets with $\mathbb{Z}^+$, getting a doubly indexed countable subset $C_\infty := \bigcup_i C_i = \{c_{ij}\}$ – here c_{ij} is the jth element of C_i. Now we define a bijection β from $S \setminus C_1$ to S by taking β to be the identity on the complement of C_∞ and by putting $\beta(c_{ij}) = c_{(i-1)j}$. This completes the proof of the lemma. □

Thus the collection of infinite subsets of $\mathbb{Z}^+$ – being a subset of $2^{\mathbb{Z}^+}$ with countable complement – is equivalent to $2^{\mathbb{Z}^+}$, and hence $(0,1] \cong 2^{\mathbb{Z}^+}$. So let us see that $(0,1]$ is of continuum type. One way is as follows: again by the above lemma, $[0,1] \cong (0,1)$, and $\mathbb{R}$ is even homeomorphic to $(0,1)$: for instance, the function

$$\arctan(\pi(x - \frac{1}{2})) : (0,1) \xrightarrow{\sim} \mathbb{R}.$$

For the case of $(\mathbb{Z}^+)^{\mathbb{R}}$: since $\mathbb{R} \cong 2^{\mathbb{Z}^+}$, it is enough to find a bijection from $(\mathbb{Z}^+)^{2^{\mathbb{Z}^+}}$ to $2^{\mathbb{Z}^+}$. This is in fact quite easy: we are given a sequence a_{ij} of binary sequences and want to make a single binary sequence. But we can do this just by choosing a bijection $\mathbb{Z}^+ \times \mathbb{Z}^+ \to \mathbb{Z}^+$.

A little more abstraction will make this argument seem much more reasonable:

LEMMA 10.18. *Suppose A, B and C are sets. Then there is a natural bijection*

$$(A^B)^C \cong A^{C \times B}.$$

PROOF. Indeed, given a function F from C to A^B and an ordered pair $(c,b) \in C \times B$, $F(c)$ is a function from B to A and so $F(c)(b)$ is an element of a. Conversely, every function from $C \times B$ to A can be viewed as a function from C to the set A^B of functions from B to A, and these correspondences are mutually inverse.[8] □

[8]This is canonical bijection is sometimes called "adjunction."

So what we said above amounts to

$$2^{\mathbb{Z}^+} \cong 2^{\mathbb{Z}^+ \times \mathbb{Z}^+} \cong (2^{\mathbb{Z}^+})^{\mathbb{Z}^+}.$$

It is also the case that $(\mathbb{Z}^+)^{\mathbb{Z}^+}$ is of continuum type. At the moment I do not see a proof of this within the framework we have developed. What we can show is that there exists an injection $\mathbb{R} \hookrightarrow (\mathbb{Z}^+)^{\mathbb{Z}^+}$ – indeed, since $\mathbb{R} \cong 2^{\mathbb{Z}^+}$, this is obvious – and also that there exists an injection $(\mathbb{Z}^+)^{\mathbb{Z}^+} \hookrightarrow 2^{\mathbb{Z}^+} \cong \mathbb{R}$.

To see this latter statement: given any sequence of positive integers, we want to return a binary sequence – which it seems helpful to think of as "encoding" our original sequence – in such a way that the decoding process is unambiguous: we can always reconstruct our original sequence from its coded binary sequence. The first thought here is to just encode each positive integer a_i in binary and concatenate them. Of course this doesn't quite work: the sequence 2, 3, 1, 1, 1 ... gets coded as 1011 followed by an infinite string of ones, as does the sequence 11, 1, 1, 1 ..... But this can be remedied in many ways. One obvious way is to retreat from binary notation to *unary* notation: we encode a_i as a string of i ones, and in between each string of a_i ones we put a zero to separate them. This clearly works (it seems almost cruelly inefficient from the perspective of information theory, but no matter).

Roughly speaking, we have shown that $(\mathbb{Z}^+)^{\mathbb{Z}+}$ is "at least of continuum type" and "at most of continuum type," so if equivalences of sets do measure some reasonable notion of their size, we ought to be able to conclude from this that $(\mathbb{Z}^+)^{\mathbb{Z}+}$ is itself of continuum type. This is true, a special case of the important Dedekind-Schröder-Bernstein Theorem.

2.4. Lots of inequivalent uncountable sets. From the fundamental Theorem 10.10 we first deduced that not all infinite sets are equivalent to each other, because the set $2^{\mathbb{Z}^+}$ is not equivalent to the countable infinite set $\mathbb{Z}^+$. We also saw that $2^{\mathbb{Z}^+} \cong \mathbb{R}$ so called it a set of continuum type. Then we noticed that Cantor's theorem implies that there are sets not of continuum type, namely $2^{\mathbb{R}} \cong 2^{2^{\mathbb{Z}^+}}$. By now one of the most startling mathematical discoveries of all time must have occurred to the reader: we can keep going!

To simplify things, let us use (and even slightly abuse) an obscure[9] but colorful notation due to Cantor: instead of writing $\mathbb{Z}^+$ we shall write $\beth_0$. For $2^{\mathbb{Z}+}$ we shall write $\beth_1$, and in general, for $n \in \mathbb{N}$, having defined $\beth_n$ (informally, as the n-fold iterated power set of $\mathbb{Z}^+$), we will define $\beth_{n+1}$ as $2^{\beth_n}$. Now hold on to your hat:

FACT 10.19. *The infinite sets $\{\beth_n\}_{n \in \mathbb{N}}$ are pairwise inequivalent.*

PROOF. Let us first make the preliminary observation that for any nonempty set S, there is a surjection $2^S \to S$. Indeed, pick your favorite element of S, say x; for every $s \in S$ we map $\{s\}$ to s, which is "already" a surjection; we extend the mapping to all of 2^S by mapping every other subset to x.

Now we argue by contradiction: suppose that for some $n > m$ there exists even a surjection $s : \beth_m \to \beth_n$. We may write $n = m+k$. By the above, by concatenating (finitely many) surjections we get a surjection $\beta : \beth_{m+k} \to \beth_{m+1}$. But then $\beta \circ s : \beth_m \to \beth_{m+1} = 2^{\beth_m}$ is a surjection, contradicting Cantor's theorem. □

[9]At least, *I* didn't know about it until recently; perhaps this is not your favorite criterion for obscurity.

Thus there are rather a lot of inequivalent infinite sets. Is it possible that the $\beth_n$'s are all the infinite sets? In fact it is *not*: define $\beth_\omega := \bigcup_{n \in \mathbb{N}} \beth_n$. This last set $\beth_\omega$ is certainly not equivalent to $\beth_n$ for any n, because it visibly surjects onto $\beth_{n+1}$. Are we done yet? No, we can keep going, defining $\beth_{\omega+1} := 2^{\beth_\omega}$.

To sum up (!!), we have a two-step process for generating a mind-boggling array of equivalence classes of sets. The first step is to pass from a set to its power set, and the second stage is to take the union over the set of all equivalence classes of sets we have thus far considered. Inductively, it seems that each of these processes generates a set which is not surjected onto by any of the sets we have thus far considered, so it gives a new equivalence class. Does the process ever end?!?

Well, the above sentence is an example of the paucity of the English language to describe the current state of affairs, since even the sequence $\beth_0$, $\beth_1$, $\beth_2 \ldots$ does not end in the conventional sense of the term. Better is to ask whether or not we can reckon the equivalence classes of sets even in terms of infinite sets. At least we have only seen countably many equivalence classes of sets[10] thus far: is it possible that the collection of all equivalence classes of sets is countable?

No again, and in fact that's easy to see. Suppose $\{S_i\}_{i \in \mathbb{N}}$ is any countable collection of pairwise inequivalent sets. Then – playing both of our cards at once! – one checks immediately that there is no surjection from any S_i onto $2^{\bigcup_{i \in \mathbb{N}} S_i}$. In fact it's even stranger than this:

FACT 10.20. *For no set I does there exists a family of sets $\{S_i\}_{i \in I}$ such that every set S is equivalent to S_i for at least one i.*

PROOF. Again, take $S_{\text{bigger}} = 2^{\bigcup_{i \in I} S_i}$. There is no surjection from $\bigcup_{i \in I} S_i$ onto S_{bigger}, so for sure there is no surjection from any S_i onto S_{bigger}. □

3. Some final remarks

Fact 20 is a truly amazing result. Once you notice that it follows readily from Cantor's Theorem 10.10, you may believe, as I do, that this theorem is the single most amazing result in all of mathematics.

There is also the question of whether this result is disturbing, or paradoxical. Can we then not speak of the set of all equivalence classes of sets (let alone, the set of all sets)? Evidently we cannot. There are too many sets to wrap all of them up into a single set. Some people have referred to this as **Cantor's Paradox**, although I do not favor this terminology: as far as I am aware, Cantor did not regard his results as paradoxical, nor do I. It does destroy the "ultranaive" notion of a set, namely, that given any "property" P, there is a set $S_P = \{x \mid P(x)\}$: according to Cantor's result, we cannot take P to be the property $x = x$. This was surprising in the late 19th century. But now we know of such things as Russell's paradox, which shows that the property $P(x)$ given by $x \notin x$ does not give rise to a set: the set of all sets which are not members of itself is a logical contradiction.

But in truth it is hard to find anyone in the 21st century who has thought for more than a few hours about sets and is this naive, i.e., who thinks that every

[10]The day you ever "see" uncountably many things, let me know.

"property" of "objects" should give rise to a set. Indeed, as you can see from the quotation marks in the previous sentence, the idea that "all mathematical objects" is well-defined and meaningful has itself come to be regarded as problematic: what is the definition of a "mathematical object"? In some sense our idea of what sets are has come to be more dynamic and iterative following Cantor's work: we start with some simple sets and some operations (like union, subsets, and power sets), and by applying various procedures these operations allow us to create new and more complicated sets.

It is certainly true that deciding what "procedures" are legal is a difficult point: none of these procedures are of the sort that the truly finitistic mind need admit to as meaningful or possible. One can only say that in order to do mathematics the vast majority of us are willing to admit (indeed, unwilling to deny) the existence of certain infinite structures and processes: note that we began by saying "[w]e assume known the set $\mathbb{Z}^+$," i.e., we assumed the existence of an infinite set. If you decide to press on to read about a more explicit examination of what properties we think sets should satisfy, you will see that one of them baldly asserts the existence of infinite sets (of a certain kind). If we remove this axiom from the list, then the collection of sets $\{[n] \mid n \in \mathbb{N}\}$ becomes a model (in the sense of mathematical logic) for all the remaining axioms: that is, it is entirely consistent and logical to believe that sets of n elements exist for every n and not to believe that the collection of *all* n's makes sense as a set. It just happens to be extraordinarily useful and interesting – and, apparently, noncontradictory – to believe in the existence of infinite sets. When contemplating the "legality" of certain abstruse-looking set-theoretic constructions, it seems wise to keep in mind the leap of faith we make even to entertain $\mathbb{Z}^+$.

4. Exercises

EXERCISE 10.1. *Show: for nonempty sets S and T, the following are equivalent:*

a) *There is a surjection $S \to T$.*
b) *There is an injection $T \to S$.*

EXERCISE 10.2. *Prove Fact 11.*
(Strategy 1: Reduce to the case of $\mathbb{Z}^+ \times \mathbb{Z}^+$ and use the diagonal path from the proof of Fact 10.7. Strategy 2: Observe that $A \times B \cong \bigcup_{a \in A} B$ and apply Fact 10.7 directly.)

EXERCISE 10.3. *Show: a subinterval of $\mathbb{R}$ containing more than one point is of continuum type.*

CHAPTER 11

Order and Arithmetic of Cardinalities

Here we pursue Cantor's theory of cardinalities of infinite sets a bit more deeply. We also begin to take a more sophisticated approach in that we identify which results depend upon the Axiom of Choice and strive to give proofs which avoid it when possible. However, we defer a formal discussion of the Axiom of Choice and its equivalents to a later installment, so the reader who has not encountered it before can ignore these comments and/or skip ahead to the next installment.

We warn the reader that the main theorem in this installment – Theorem 11.4 (which we take the liberty of christening "The Second Fundamental Theorem of Set Theory") – will not be proved until the next installment, in which we give a systematic discussion of well-ordered sets.

For More Advanced Readers: We will mostly be content to use the Axiom of Choice (AC) in this handout, despite the fact that we will not discuss this axiom until Handout 3. However, whereas in the previous chapter we blithely used AC without any comment whatsoever, here for a theorem whose statement requires AC we indicate that by calling it **AC-Theorem**. (If a theorem holds without AC, we sometimes still gives proofs which use AC, if they are easier or more enlightening.)

1. The fundamental relation $\leq$

Let's look back at what we did in the last section. We introduced a notion of equivalence on sets: namely $S_1 \equiv S_2$ if there is a bijection $f : S_1 \to S_2$. This sets up a project of classifying sets up to equivalence. Looking at finite sets, we found that each equivalence class contained a representative of the form $[n]$ for a unique natural number n. Thus the set of equivalence classes of finite sets is $\mathbb{N}$. Then we considered whether all infinite sets were equivalent to each other, and found that they are not.

If we look back at finite sets (it is remarkable, and perhaps comforting, how much of the inspiration for some rather recondite-looking set-theoretic constructions comes from the case of finite sets) we can't help but notice that $\mathbb{N}$ has so much more structure than just a set. First, it is a semiring: this means that we have operations of $+$ and $\cdot$, but in general we do not have $-$ or $/$. Second it has a natural ordering $\leq$ which is indeed a *well-ordering*: that is, $\leq$ is a linear ordering on x in which every non-empty subset has a least element. (The well-ordering property is easily seen to be equivalent to the principle of mathematical induction.)

Remarkably, *all* of these structures generalize fruitfully to equivalence classes of sets! What does this mean? For a set S, let $\#S$ stand for its equivalence class.

(This construction is commonplace in mathematics but has problematic aspects in set theory since the collection of sets equivalent with any nonempty set S does not form a set. Let us run with this notion for the moment, playing an important mathematician's trick: rather than worrying about what $\#S$ *is*, let us see how it *behaves*, and then later we can attempt to define it in terms of its behavior.)

DEFINITION 11.1. *We write $S_1 \leq S_2$ if there exists an injection $\iota : S_1 \hookrightarrow S_2$.*

PROPOSITION 11.2. *Let S_1 be a nonempty set and S_2 a set. If there is an injection from S_1 to S_2, then there is a surjection from S_2 to S_1.*

PROOF. Let $\iota : S_1 \to S_2$ be an injection. We define $s : S_2 \to S_1$ as follows. Let $x_1 \in S_2$. If $y \in \iota(S_1)$, then since ι is injective there is exactly one $x \in S_1$ with $\iota(x) = y$, and we set $s(y) = x$. If $y \notin \iota(S_1)$, we set $s(y) = x_1$. This is a surjection. □

AC-THEOREM 11.3. *Let S_1 be a nonempty set and S_2 a set. If there is a surjection from S_2 to S_1, then there is an injection from S_1 to S_2.*

PROOF. Let $s : S_2 \to S_1$ be a surjection. We define $\iota : S_1 \to S_2$ as follows. For each $x \in S_1$, we **choose** $y \in S_2$ with $s(y) = x$ and define $\iota(x) = y$. If for $x_1, x_2 \in S_1$ we have $\iota(x_1) = \iota(x_2)$, then $x_1 = s(\iota(x_1)) = s(\iota(x_2)) = x_2$, so ι is an injection. □

Let $\mathcal{F}$ be any family (i.e., set!) of sets S_α. Then our $\leq$ gives a relation on $\mathcal{F}$; what properties does it have? It is of course reflexive and transitive, which means it is (by definition) a *quasi-ordering*. On the other hand, it is generally not a partial ordering, because $S_{\alpha_1} \leq S_{\alpha_2}$ and $S_{\alpha_2} \leq S_{\alpha_1}$ does not in general imply that $S_{\alpha_1} = S_{\alpha_2}$: indeed, suppose have two distinct, but equivalent sets (say, two sets with three elements apiece). However, given a quasi-ordering we can formally associate a partial ordering, just by taking the quotient by the equivalence relation $x \leq y, y \leq x$. However, exactly how the associated partial ordering relates to the given quasi-ordering is in general unclear.

Therefore we can try to do something less drastic. Namely, let us write $\#S_1 \leq \#S_2$ if $S_1 \leq S_2$. We must check that this is well-defined, but no problem: indeed, if $S_i \equiv T_i$ then choosing bijections $\beta_i : S_i \to T_i$, we get an injection

$$\beta_2 \circ \iota \circ \beta_1^{-1} : T_1 \to T_2.$$

Thus we can pass from the quasi-ordered set $(\mathcal{F}, \leq)$ to the quasi-ordered set of equivalence classes $(\#\mathcal{F}, \leq)$. Since we removed an obvious obstruction to the quasi-ordering being a partial ordering, it is natural to wonder whether or not this partial ordering on equivalence classes is better behaved. If $\mathcal{F}$ is a family of finite sets, then $\#\mathcal{F}$ is a subset of $\mathbb{N}$ so we have a well-ordering. The following stunning result asserts that this remains true for infinite sets:

AC-THEOREM 11.4. *(Second Fundamental Theorem of Set Theory) For any family $\mathcal{F}$ of sets, the relation $\leq$ descends to $\#\mathcal{F}$ and induces a well-ordering.*

In its full generality, Theorem 11.4 is best derived in the course of a systematic development of the theory of well-ordered sets, and we shall present this theory later on. However, the following special case can be proved now:

THEOREM 11.5. *(Dedekind-Schröder-Bernstein) If $M \leq N$ and $N \leq M$, then $M \equiv N$.*

PROOF. Certainly we may assume that M and N are disjoint. Let $f : M \hookrightarrow N$ and $g : N \hookrightarrow M$. Consider the following function B on $M \cup N$: if $x \in M$, $B(x) = f(x) \in N$; if $x \in N$, $B(x) = g(x) \in M$. Now we consider the B orbits on $M \cup N$. Put $B^m = B \circ \ldots \circ B$ (m times). There are three cases:
Case 1: The forward B-orbit of x is finite. Equivalently, for some m, $B^m(x) = x$. Then the backwards B-orbit is equal to the B-orbit, so the full B-orbit is finite. Otherwise the B-orbit is infinite, and we consider the backwards B-orbit.
Case 2: The backwards B-orbit also continues indefinitely, so for all $m \in \mathbb{Z}$ we have pairwise disjoint elements $B^m(x) \in M \cup N$.
Case 3: For some $m \in \mathbb{Z}^+$, $B^{-m}(x)$ is not in the image of f or g.
As these possibilities are exhaustive, we get a partition of $M \cup N$ into three types of orbits: (i) finite orbits, (ii) $\{B^m \mid m \geq m_0\}$, and (iii) $\{B^m \mid m \in \mathbb{Z}\}$. We can use this information to define a bijection from M to N. Namely, f itself is necessarily a bijection from the Case 1 elements of M to the Case 1 elements of N, and the same holds for Case 3. f need not surject onto every Case 2 element of N, but the Case 2 element of $M \cup N$ have been partitioned into sets isomorphic to $\mathbb{Z}^+$, and pairing up the first element occurring in M with the first element occurring in N, and so forth, we have defined a bijection from M to N! □

Theorem 11.4 asserts that $\#S$ is measuring, in a reasonable sense, the *size* of the set S: if two sets are inequivalent, it is because one of them is larger than the other. This motivates a small change of perspective: we will say that $\#S$ is the *cardinality* of the set S. Note well that we have not made any mathematical change: we have not defined cardinalities in an absolute sense – i.e., we have not said what sort of object $\#\mathbb{N}$ is – but only in a relational sense: i.e., as an invariant of a set that measures whether a set is bigger or smaller than another set.

Notation: For brevity we will write

$$\aleph_0 := \#\mathbb{N}$$

and

$$\mathfrak{c} := \#\mathbb{R}.$$

Here $\aleph$ is the Hebrew letter "aleph", and $\aleph_0$ is usually pronounced "aleph naught" or "aleph null" rather than "aleph zero". Exactly why we are choosing such a strange name for $|\mathbb{N}|$ will not be explained until the third handout. In contrast, we write $\mathfrak{c}$ for $\#\mathbb{R}$ simply because "c" stands for *continuum*, and in Handout 1 we said that a set S if **of continuum type** if $S \equiv \mathbb{R}$. In our new notation, [**?**, Fact 16] is reexpressed as

$$2^{\aleph_0} = \mathfrak{c}. \tag{61}$$

2. Addition of cardinalities

For two sets S_1 and S_2, define the disjoint union $S_1 \coprod S_2$ to be $S_1' \cup S_2'$, where $S_i' = \{(s, 1) \mid s \in S_i\}$. Note that there is an obvious bijection $S_i \to S_i'$; the point of this little artifice is that even if S_1 and S_2 are not disjoint, S_1' and S_2' will be.[1] Now consider the set $S_1 \coprod S_2$.

FACT 11.6. *The equivalence class $\#(S_1 \coprod S_2)$ depends only on the equivalence classes $\#S_1$ and $\#S_2$.*

[1]This in turn raises canonicity issues, which we will return to later.

PROOF. : All this means is that if we have bijections $\beta_i : S_i \to T_i$, then there is a bijection from $S_1 \coprod S_2$ to $T_1 \coprod T_2$, which is clear: there is indeed a canonical bijection, namely $\beta_1 \coprod \beta_2$: by definition, this maps an element $(s,1)$ to $(\beta_1(s),1)$ and an element $(s,2)$ to $(\beta_2(s),2)$. □

The upshot is that it makes formal sense to define $\#S_1 + \#S_2$ as $\#(S_1 \coprod S_2)$: our addition operation on sets descends to equivalence classes. On finite sets this amounts to

$$m + n = \#[m] + \#[n] = \#([m] \coprod [n]) = \#[m+n]| = m + n.$$

THEOREM 11.7. *Let $S \leq T$ be sets, with T infinite. Then $\#S + \#T = \#T$.*

There is a fairly quick and proof of Theorem 11.7, which however uses Zorn's Lemma (which is equivalent to the Axiom of Choice). At this stage in the development of the theory the reader might like to see such a proof, so we will present it now (certainly Zorn's Lemma is well known and used in "mainstream mathematics"). We begin with the following preliminary result which is of interest in its own right.

AC-THEOREM 11.8. *Any infinite set S is a disjoint union of countable subsets.*

PROOF. Consider the partially ordered set each of whose elements is a pairwise disjoint family of countable subsets of S, and with $\leq$ being set-theoretic inclusion. Any chain $\mathcal{F}_i$ in this poset has an upper bound: just take the union of all the elements in the chain: this is certainly a family of countable subsets of S, and if any element of $\mathcal{F}_i$ intersects any element of $\mathcal{F}_j$, then $\mathcal{F}_{\max(i,j)}$ contains both of these elements so is not a pairwise disjoint family, contradiction. By Zorn's Lemma we are entitled to a maximal such family $\mathcal{F}$. Then $S \setminus \bigcup_{i \in \mathcal{F}} S_i$ must be finite, so the remaining elements can be added to any one of the elements of the family. □

AC-THEOREM 11.9. *For any infinite set A, there are disjoint subsets B and C with $A = B \cup C$ and $\#A = \#B = \#C$.*

PROOF. Express $A = \bigcup_{i \in \mathcal{F}} A_i$, where each $A_i \cong \mathbb{Z}^+$. So partition S_i into $B_i \cup C_i$ where B_i and C_i are each countable, and take $B = \bigcup_{i \in \mathcal{F}} B_i,\ C = \bigcup_{i \in \mathcal{F}} C_i$. □

Proof of Theorem 11.7: Let S and T be sets; by Theorem 11.4 we may assume $\#S \leq \#T$. Then clearly $\#S + \#T \leq \#T + \#T$, but the preceding result avers $\#T + \#T = \#T$. So $\#S + \#T \leq \#T$. Clearly $\#T \leq \#S + \#T$, so by the Dedekind-Schröder-Bernstein Theorem we conclude $\#S + \#T = \#T$.

AC-THEOREM 11.10. *For all infinite sets S and T, we have*

$$\#S + \#T = \max(\#S, \#T).$$

3. Subtraction of cardinalities

It turns out that we cannot formally define a subtraction operation on infinite cardinalities, as one does for finite cardinalities using set-theoretic subtraction: given sets S_1 and S_2, to define $|S_1| - |S_2|$ we would like to find sets $T_i \equiv S_i$ such that $T_2 \subseteq T_1$, and then define $|S_1| - |S_2|$ to be $|T_1 \setminus T_2|$. Even for finite sets this only makes literal sense if $|S_2| \leq |S_1|$; in general, we are led to introduce negative numbers through a formal algebraic process, which we can recognize as the group completion of a monoid (or the ring completion of a commutative semiring).

However, here the analogy between infinite and finite breaks down: given $S_2 \subseteq$

S_1, $T_2 \subseteq T_1$ and bijections $\beta_i : S_i \to T_i$, we absolutely do not in general have a bijection $S_1 \setminus S_2 \to T_1 \setminus T_2$. For instance, take $S_1 = T_1 = \mathbb{Z}^+$ and $S_2 = 2\mathbb{Z}^+$, the even numbers. Then $|S_1 \setminus S_2| = |\mathbb{N}|$. However, we could take $T_2 = \mathbb{Z}^+$ and then $T_2 \setminus T_1 = \emptyset$. For that matter, given any $n \in \mathbb{Z}^+$, taking T_2 to be $\mathbb{Z}^+ \setminus [n]$, we get $T_1 \setminus T_2 = [n]$. Thus when attempting to define $|\mathbb{N}| - |\mathbb{N}|$ we find that we get all conceivable answers, namely all equivalence classes of at most countable sets. This phenomenon does generalize:

PROPOSITION 11.11. *(Subtraction theorem) For any sets $S_1 \subseteq S_2 \subseteq S_3$, there are bijections $\beta_1 : S_1 \to T_1$, $\beta_3 : S_3 \to T_3$ such that $T_1 \subseteq T_3$ and $|T_3 \setminus T_1| = |S_2|$.*

PROOF. If S_1 and S_2 are disjoint, we may take $T_1 = S_1$, $T_2 = S_2$ and $T_3 = S_1 \cup S_2$. Otherwise we may adjust by bijections to make them disjoint. □

4. Multiplication of cardinalities

DEFINITION 11.12. *Let S_1 and S_2 be sets. We define*

$$\#S_1 \times \#S_2 := \#(S_1 \times S_2).$$

In Exercise 11.6 you are asked to show that this multiplication operation is well-defined.

At this point, we have what appears to be a very rich structure on our cardinalities: suppose that $\mathcal{F}$ is a family of sets which is, up to bijection, closed under $\coprod$ and $\times$. Then the family $|\mathcal{F}|$ of cardinalities of these sets has the structure of a well-ordered semiring.

EXAMPLE 11.13. *Let $\mathcal{F}$ be any collection of finite sets containing, for all $n \in \mathbb{N}$, at least one set with n elements. Then $|\mathcal{F}| = \mathbb{N}$ and the semiring and (well)-ordering are the usual ones.*

EXAMPLE 11.14. *Let $\mathcal{F}$ be a family containing finite sets of all cardinalities together with $\mathbb{N}$. Then, since $\mathbb{N} \coprod \mathbb{N} \cong \mathbb{N}$ and $\mathbb{N} \times \mathbb{N} \cong \mathbb{N}$, the corresponding family of cardinals $|\mathcal{F}|$ is a well-ordered semiring. It contains $\mathbb{N}$ as a subring and one other element, $|\mathbb{N}|$; in other words, as a set of cardinalities it is $\mathbb{N} \cup \{\mathbb{N}\}$, a slightly confusing-looking construction which we will see much more of later on. As a well-ordered set we have just taken $\mathbb{N}$ and added a single element (the element $\mathbb{N}$!) which is is larger than every other element. It is clear that this gives a well-ordered set; indeed, given any well-ordered set $(S, \leq)$ there is another well-ordered set, say $s(S)$, obtained by adding an additional element which is strictly larger than every other element (check and see that this gives a well-ordering). The semiring structure is, however, not very interesting: every $x \in \mathbb{N} \cup \{\mathbb{N}\}$, $x + \mathbb{N} = x \cdot \mathbb{N} = \mathbb{N}$. In particular, the ring completion of this semiring is the 0 ring. (It suffices to see this on the underlying commutative monoid. Recall that the group completion of a commutative monoid M can be represented by pairs (p, m) of elements of M with $(p, m) \sim (p', m')$ iff there exists some $x \in M$ such that $x + p + m' = x + p' + m$. In our case, taking $x = \mathbb{N}$ we see that all elements are equivalent.)*

However multiplication of infinite cardinalities turns out not to be very interesting.

THEOREM 11.15. *Let T be infinite and S a nonempty subset of T. Then:*

$$\#S \times \#T = \#T.$$

The same remarks are in order here as for the addition theorem (Theorem 11.7): combining with cardinal trichotomy, we conclude that $\#S \times \#T = \max(\#S, \#T)$ for any infinite sets. This deduction uses the Axiom of Choice, whereas the theorem as stated does not. However, it is easier to give a proof using Zorn's Lemma, which is what we will do. Moreover, as for the additive case, it is convenient to first establish the case of $S = T$. Indeed, assuming that $T \times T \cong T$, we have

$$\#S \times \#T \leq \#T \times \#T = \#T \leq \#S \times \#T.$$

So let us prove that for any infinite set T, $T \times T \cong T$.

Consider the poset consisting of pairs (S_i, f_i), where $S_i \subseteq T$ and f_i is a bijection from S_i to $S_i \times S_i$. Again the order relation is the natural one: $(S_i, f_i) \leq (S_j, f_j)$ if $S_i \subseteq S_j$ and $f_j|_{S_i} = f_i$. Now we apply Zorn's Lemma, and the verification that every chain has an upper bound is immediate because we can just take the union over all elements of the chain. Therefore we get a maximal element (S, f).

Now, as for the case of the addition theorem, we need not have $S = T$; put $S' = T \setminus S$. What we *can* say is that $\#S' < \#S$. Indeed, otherwise we have $\#S' \geq \#S$, so that inside S' there is a subset S'' with $\#S'' = \#S$. But we can enlarge $S \times S$ to $(S \cup S'') \times (S \cup S'')$. The bijection $f : S \to S \times S$ gives us that

$$\#S'' = \#S = \#S \times \#S = \#S'' \times \#S''.$$

Thus using the addition theorem, there is a bijection $g : S \cup S'' \to (S \cup S'') \times (S \cup S'')$ which can be chosen to extend $f : S \to S \times S$, contradicting the maximality of (S, f).

Thus we have that $\#S' < \#S$ as claimed. But then we have

$$\#T = \#S \cup S' = \max(\#S, \#S') = \#S,$$

so

$$\#T \times \#T = \#S \times \#S = \#S = \#T,$$

completing the proof.

5. Cardinal Exponentiation

For two sets S and T, we define S^T to be the set of all functions $f : T \to S$. Why do we write S^T instead of T^S? Because the cardinality of the set of all functions from $[m]$ to $[n]$ is n^m: for each of the m elements of the domain, we must select one of the n elements of the codomain. As above, this extends immediately to infinite cardinalities:

DEFINITION 11.16. *For any sets S and T, we put $(\#S)^{\#T} := \#S^T$.*

Henceforth we may as well assume that X has at least two elements.

PROPOSITION 11.17. *For any sets X, Y, Z we have*

$$((\#X)^{\#Y})^{\#Z} = \#X^{\#Y \cdot \#Z}.$$

PROOF. By 10.18 we have $(X^Y)^Z \equiv X^{YZ}$. The result follows immediately. □

PROPOSITION 11.18. *For any sets X, Y, Z, we have*

$$(\#X)^{\#Y + \#Z} = (\#X)^{\#Y} \cdot (\#X)^{\#Z}$$

and

$$(\#X \cdot \#Y)^{\#Z} = (\#X)^{\#Y} \cdot (\#X)^{\#Z}.$$

You are asked to prove Proposition 11.18 in Exericse 11.1.

THEOREM 11.19. *Let X_1, X_2, Y_1, Y_2 be sets with $Y_1 \neq \varnothing$. If $\#X_1 \leq \#X_2$ and $\#Y_1 \leq \#Y_2$, then $(\#X_1)^{\#Y_1} \leq (\#X_2)^{\#Y_2}$.*

PROOF. Let $\iota_X : X_1 \to X_2$ be an injection. By Proposition 11.2, there is a surjection $s_Y : Y_2 \to Y_1$. There is an induced injection $X_1^{Y_1} \to X_2^{Y_1}$ given by

$$f : Y_1 \to X_1 \mapsto \iota_X \circ f : Y_1 \to X_2$$

and an induced injection $X_2^{Y_1} \to X_2^{Y_2}$ given by

$$f : Y_1 \to X_2 \mapsto f \circ s_Y : Y_2 \to X_2.$$

Composing these gives an injection from $X_1^{Y_1}$ to $X_2^{Y_2}$. □

If Y is finite, then $(\#X)^{\#Y} = \#X \cdot \ldots \cdot \#X$ so is nothing new. The next result evaluates, in a sense, $(\#X)^{\#Y}$ when $\#Y = \aleph_0$.

AC-THEOREM 11.20. *Let S be a set with $2 \leq \#S \leq \mathfrak{c}$. Then $(\#S)^{\aleph_0} = \mathfrak{c}$.*

PROOF. There is an evident bijection from the set of functions $\mathbb{N} \to \{1, 2\}$ to the power set $2^{\mathbb{N}}$, so $2^{\aleph_0} = \mathfrak{c}$. Combining this with Theorem 11.19 and Proposition 11.18 we get

$$\mathfrak{c} = 2^{\aleph_0} \leq (\#S)^{\aleph_0} \leq \mathfrak{c}^{\aleph_0} = (2^{\aleph_0})^{\aleph_0} = 2^{\aleph_0 \times \aleph_0} = 2^{\aleph_0} = \mathfrak{c}.$$ □

What about $(\#X)^{\#Y}$ when Y is uncountable? By Cantor's Theorem we have

$$(\#X)^{\#Y} \geq (\#\{0, 1\})^{\#Y} = 2^{\#Y} > \#Y.$$

Thus the first order of business seems to be the evaluation of $2^{\#Y}$ for uncountable Y. This turns out to be an extremely deep issue with a very surprising answer.

What might one expect $2^{\#S}$ to be? The most obvious guess seems to be the minimalist one: since any collection of cardinalities is well-ordered, for any cardinality κ, there exists a smallest cardinality which is greater than κ, traditionally called κ^+. Thus we might expect $2^{\#S} = (\#S)^+$ for all infinite S.

But comparing to finite sets we get a little nervous about our guess, since 2^n is very much larger than $n^+ = n + 1$. On the other hand, our simple formulas for addition and multiplication of infinite cardinalities do not hold for finite cardinalities either – in short, we have no real evidence so are simply guessing.

Notice that we did not even "compute" $2^{\aleph_0}$ in any absolute sense but only showed that it is equal to the cardinality $\mathfrak{c}$ of the real numbers. So already it makes sense to ask whether $\mathfrak{c}$ is the *least* cardinality greater than $\aleph_0$ or whether it is larger. The minimalist guess $\mathfrak{c} = \aleph_0^+$ was made by Cantor, who was famously unable to prove it, despite much effort: it is now called the **Continuum Hypothesis** (CH). Moreover, the guess that $2^{\#S} = (\#S)^+$ for all infinite sets is called the **Generalized Continuum Hypothesis** (GCH).

The Continuum Hypothesis and its generalization is a reasonable candidate for the most vexing problem in all of matheamtics. Starting with Cantor himself, some of the greatest mathematical minds have been brought to bear on this problem. For instance, in his old age David Hilbert claimed a proof of CH and published it in a prestigious journal, but the proof was flawed. Kurt Gödel proved in 1944 that

CH is relatively consistent with the ZFC axioms for set theory – in other words, assuming that the ZFC axioms are consistent (if not, all statements in the language can be formally derived from them!), it is not possible to deduce CH as a formal consequence of these axioms. In 1963, Paul Cohen showed that the negation of CH is also relatively consistent with ZFC, and for this he received the Fields Medal. Cohen's work undoubtedly revolutionized set theory, and his methods ("forcing") have since become an essential tool. But where does this leave the status of the Continuum Hypothesis?

The situation is most typically summarized by saying that Gödel and Cohen showed the undecidability of CH – i.e., that it is neither true nor false in the same way that Euclid's parallel postulate is neither true nor false. However, to accept this as the end of the story is to accept that what we know about sets and set theory is exactly what the ZFC axiom scheme tells us, but of course this is a position that would require (philosophical as well as mathematical) justification – as well as a position that seems to be severely undermined by the very issue at hand! Thus, a more honest admission of the status of CH would be: we are not even sure whether or not the problem is open. From a suitably Platonistic mathematical perspective – i.e., a belief that what is true in mathematics is different from what we are able (in practice, or even in principle) to prove – one feels that either there exists some infinite subset of $\mathbb{R}$ which is equivalent to neither $\mathbb{Z}^+$ nor $\mathbb{R}$, or there doesn't, and the fact that none of the ZFC axioms allow us to decide this simply means that the ZFC axioms are not really adequate. It is worth noting that this position was advocated by both Gödel and Cohen.

In recent years this position has begun to shift from a philosophical to a mathematical one: the additional axioms that will decide CH one way or another are no longer hypothetical. The only trouble is that they are themselves very complicated, and "intuitive" mostly to the set theorists that invent them. Remarkably – considering that the Axiom of Choice and GCH are to some extent cognate (and indeed GCH implies AC) – the consensus among experts seems to be settling towards *rejecting* CH in mathematics. Among notable proponents, we mention the leading set theorist Hugh Woodin. His and other arguments are vastly beyond the scope of these notes.

To a certain extent, cardinal exponentation reduces to the problem of computing the cardinality of 2^S. Indeed, one can show the following result.

AC-Theorem 11.21. *If X has at least 2 elements and Y has at least one element,*

$$\max(\#X, 2^{\#Y}) \leq (\#X)^{\#Y} \leq \max(2^{\#X}, 2^{\#Y}).$$

We omit the proof for now.

6. Embedding Countable Ordered Sets

7. Exercises

Exercise 11.1. *Prove Proposition 11.18.*

EXERCISE 11.2. *Suppose $S_1 = \varnothing$. Under what conditions on S_2 does Proposition 11.2 remain valid? What about Theorem 11.3?*

EXERCISE 11.3. *Check that the definition of cardinal exponentiation is well-defined (i.e., does not depend upon the sets but only their cardinalities).*

EXERCISE 11.4. *Suppose X has at most one element. Compute $(\#X)^{\#Y}$ for any set Y.*

EXERCISE 11.5. *Prove the analogue of Proposition 11.11 for cardinal division.*

EXERCISE 11.6. *Check that the multiplication of cardinals is well-defined.*

EXERCISE 11.7. *Verify that $+$ and $\cdot$ are commutative and associative operations on cardinalities, and that multiplication distributes over addition. (There are two ways to do this. One is to use the fact that $\#S + \#T = \#S \cdot \#T = \max(\#S, \#T)$ unless S and T are both finite. On the other hand one can verify these identities directly in terms of identities on sets.)*

EXERCISE 11.8. *Prove Proposition 11.18.*

EXERCISE 11.9. *Let T be an infinite set, and let S be a nonempty subset of T. Show that S can be expressed as a disjoint union of subsets of cardinality $\#T$.*

EXERCISE 11.10. *Deduce Theorem 11.10 from Theorem 11.4 and Theorem 11.7.*

CHAPTER 12

Well-Ordered Sets, Ordinalities and the Axiom of Choice

1. The Calculus of Ordinalities

1.1. Well-ordered sets and ordinalities.

The discussion of cardinalities in Chapter 2 suggests that the most interesting thing about them is their order relation, namely that any set of cardinalities forms a well-ordered set. So in this section we shall embark upon a systematic study of well-ordered sets. Remarkably, we will see that the problem of classifying sets up to bijection is literally contained in the problem of classifying well-ordered sets up to order-isomorphism.

PROPOSITION 12.1. *For a linearly ordered set $(X, \leq)$, the following are equivalent:*

(i) *X satisfies the descending chain condition: there are no infinite strictly descending sequences $x_1 > x_2 > \dots$ in X.*
(ii) *X is well-ordered.*

You are asked to Prove Proposition 12.1 in Exercise 12.1. We need the notion of "equivalence" of of well-ordered sets. A mapping $f : S \to T$ between partially ordered sets is **order preserving** (or an **order homomorphism**) if $s_1 \leq s_2$ in S implies $f(s_1) \leq f(s_2)$ in T.

An **order isomorphism** between posets is a mapping f which is order preserving, bijective, and whose inverse f^{-1} is order preserving. (This is the general – i.e., categorical – definition of isomorphism of structures.)

In Exercise 12.3 you are asked to show that the inverse of an order-preserving bijection between *partially* ordered sets need not be order-preserving. However:

LEMMA 12.2. *Let $(X, \leq)$ be a totally ordered set and $(Y, \leq)$ a partially ordered set, and let $f : X \to Y$ be an order-preserving bijection. Then f is an order isomorphism (so Y is also totally ordered).*

You are asked to prove Lemma 12.2 in Exercise 12.4.

LEMMA 12.3. *(Rigidity Lemma) Let S and T be well-ordered sets and $f_1,\ f_2 : S \to T$ two order isomorphisms. Then $f_1 = f_2$.*

PROOF. Let f_1 and f_2 be two order isomorphisms between the well-ordered sets S and T, which we may certainly assume are nonempty. Consider S_2, the set of elements s of S such that $f_1(s) \neq f_2(s)$, and let $S_1 = S \setminus S_2$. Since the least

element of S must get mapped to the least element of T by any order-preserving map, S_1 is nonempty; put $T_1 = f_1(S_1) = f_2(S_1)$. Supposing that S_2 is nonempty, let s_2 be its least element. Then $f_1(s_2)$ and $f_2(s_2)$ are both characterized by being the least element of $T \setminus T_1$, so they must be equal, a contradiction. □

Let us define an **ordinality** to be an order-isomorphism class of well-ordered sets, and write $o(X)$ for the order-isomorphism class of X. The intentionally graceless terminology will be cleaned up later on. Since two-order isomorphic sets are equipotent, we can associate to every ordinality α an "underlying" cardinality $\#\alpha$: $\#o(X)| = \#X$. It is natural to expect that the classification of ordinalities will be somewhat richer than the classification of cardinalities – in general, endowing a set with extra structure leads to a richer classification – but the reader new to the subject may be (we hope, pleasantly) surprised at how much richer the theory becomes.

From the perspective of forming "isomorphism classes" (a notion the ontological details of which we have not found it profitable to investigate too closely) ordinalities have a distinct advantage over cardinalities: according to the Rigidity Lemma, any two representatives of the same ordinality are *uniquely* (hence canonically!) isomorphic. This in turn raises the hope that we can write down a *canonical* representative of each ordinality. This hope has indeed been realized, by von Neumann, as we shall see later on: the canonical representatives will be called "ordinals." While we are alluding to later developments, let us mention that just as we can associate a cardinality to each ordinality, we can also – and this is much more profound – associate an ordinality $o(\kappa)$ to each cardinality κ. This assignment is *one-to-one*, and this allows us to give a canonical representative to each cardinality, a "cardinal." At least at the current level of discussion, there is no purely mathematical advantage to the passage from cardinalities to cardinals, but it has a striking ontological consequence, namely that, up to isomorphism, we may develop all of set theory in the context of "pure sets", i.e., sets whose elements (and whose elements' elements, and ...) are themselves sets.

But first let us give some basic examples of ordinalities and ways to construct new ordinalities from preexisting ones.

2. Algebra of ordinalities

EXAMPLE 12.4. *Trivially the empty set is well-ordered, as is any set of cardinality one. These sets, and only these sets, have unique well-orderings.*

EXAMPLE 12.5. *Our "standard" example $[n]$ of the cardinality n comes with a well-ordering. Moreover, on a finite set, the concepts of well-ordering and linear ordering coincide, and it is clear that there is up to order isomorphism a unique linear ordering on $[n]$. Informally, given any two orderings on an n element set, we define an order-preserving bijection by pairing up the least elements, then the second-least elements, and so forth. (For a formal proof, use induction.)*

EXAMPLE 12.6. *As we know, the usual ordering on $\mathbb{N}$ is a well-ordering. Notice that for any $N \in \mathbb{Z}$ we have an order isomorphism from $\mathbb{N} = \mathbb{Z}^{\geq 0}$ to $\mathbb{Z}^{\geq N}$ just by $x \mapsto x + N$. As is traditional, we write ω for the ordinality of $\mathbb{N}$.*

For a partially ordered set X, we can define a new partially ordered set $X^+ := X \cup \{\infty\}$ by adjoining some new element ∞ and decreeing $x \leq \infty$ for all $x \in X$.

PROPOSITION 12.7. *If X is well-ordered, so is X^+.*

Proof: Let Y be a nonempty subset of X^+. Certainly there is a least element if $|Y| = 1$; otherwise, Y contains an element other than ∞, so that $Y \cap X$ is nonempty, and its least element will be the least element of Y.

If X and Y are order-isomorphic, so too are X^+ and Y^+, so the passage from X to X^+ may be viewed as an operation on ordinalities. We denote $o(X^+)$ by $o(X) + 1$, the **successor ordinality** of $o(X)$.

Note that all the finite ordinalities are formed from the empty ordinality 0 by iterated successorship. However, not every ordinality is of the form $o + 1$, e.g. ω is clearly not: it lacks a maximal element. (On the other hand, it is obtained from *all* the finite ordinalities in a way that we will come back to shortly.) We will say that an ordinality o is a **successor ordinality** if it is of the form $o' + 1$ for some ordinality o' and a **limit ordinality** otherwise. Thus 0 and ω are limit ordinals.

EXAMPLE 12.8. *The successor operation allows us to construct from ω the new ordinals $\omega + 1$, $\omega + 2 := (\omega + 1) + 1$, and for all $n \in \mathbb{Z}^+$, $\omega + n := (\omega + (n-1)) + 1$: these are all distinct ordinals.*

Definition: For partially ordered sets $(S_1, \leq_1)$ and $(S_2, \leq_2)$, we define $S_1 + S_2$ to be the set $S_1 \coprod S_2$ with $s \leq t$ if either of the following holds:
(i) For $i = 1$ or 2, s and t are both in S_i and $s \leq_i t$;
(ii) $s \in S_1$ and $s \in S_2$.

Informally, we may think of $S_1 + S_2$ as "S_1 followed by S_2."

PROPOSITION 12.9. *If S_1 and S_2 are linearly ordered (resp. well-ordered), so is $S_1 + S_2$.*

You are asked to prove Proposition 12.9 in Exercise 12.7.

Again the operation is well-defined on ordinalities, so we may speak of the **ordinal sum** $o + o'$. By taking $S_2 = [1]$, we recover the successor ordinality: $o + [1] = o + 1$.

EXAMPLE 12.10. *The ordinality $2\omega := \omega + \omega$ is the class of a well-ordered set which contains one copy of the natural numbers followed by another. Proceeding inductively, we have $n\omega := (n-1)\omega + \omega$, with a similar description.*

Warning: We can also form the ordinal sum $1 + \omega$, which amounts to adjoining to the natural numbers a smallest element. But this is still order-isomorphic to the natural numbers: $1 + \omega = \omega$. In fact the identity $1 + o = o$ holds for every infinite ordinality (as will be clear later on). In particular $1 + \omega \neq \omega + 1$, so beware: the ordinal sum is not commutative! (To my knowledge it is the only non-commutative operation in all of mathematics that is invariably denoted by "+".) It is however immediately seen to be associative.

The notation 2ω suggests that we should have an ordinal product, and indeed we do:

Definition: For posets $(S_1, \leq_1)$ and $(S_2, \leq_2)$ we define the **lexicographic product** to be the Cartesian product $S_1 \times S_2$ endowed with the relation $(s_1, s_2) \leq (t_1, t_2)$ if(f) either $s_1 \leq t_1$ or $s_1 = t_1$ and $s_2 \leq t_2$.

PROPOSITION 12.11. *If S_1 and S_2 are linearly ordered (resp. well-ordered), so is $S_1 \times S_2$.*

You are asked to prove Proposition 12.11 in Exercise 12.8. As usual this operation is well-defined on ordinalities so leads to the **ordinal product** $o \cdot o'$.

EXAMPLE 12.12. *For any well-ordered set X, $[2]\cdot X$ gives us one copy $\{(1,x) \mid x \in X\}$ followed by another copy $\{(2,x) \mid x \in X\}$, so we have a natural isomorphism of $[2] \cdot X$ with $X + X$ and hence $2 \cdot o = o + o$. (Similarly for $3o$ and so forth.) This time we should be prepared for the failure of commutativity: $\omega \cdot n$ is isomorphic to ω. This allows us to write down $\omega^2 := \omega \times \omega$, which we visualize by starting with the positive integers and then "blowing up" each positive integer to give a whole order isomorphic copy of the positive integers again. Repeating this operation gives $\omega^3 = \omega^2 \cdot \omega$, and so forth. Altogether this allows us to write down ordinalities of the form $P(\omega) = a_n\omega^n + \ldots + a_1\omega + a_0$ with $a_i \in \mathbb{N}$, i.e., polynomials in ω with natural number coefficients. It is in fact the case that (i) distinct polynomials $P \neq Q \in \mathbb{N}[T]$ give rise to distinct ordinalities $P(\omega) \neq Q(\omega)$; and (ii) any ordinality formed from $[n]$ and ω by finitely many sums and products is equal to some $P(\omega)$ – even when we add/multiply in "the wrong order", e.g. $\omega * 7 * \omega^2 * 4 + \omega * 3 + 11 = \omega^3 + \omega + 11$ – but we will wait until we know more about the ordering of ordinalities to try to establish these facts.*

We also have a way to exponentiate ordinalities: let $\alpha = o(X)$ and $\beta = o(Y)$. Then by α^β we mean the order isomorphism class of the set $Z = Z(Y,X)$ of all functions $f : Y \to X$ with $f(y) = 0_X$ (0_X denotes the minimal element of X) for all but finitely many $y \in Y$, ordered by $f_1 \leq f_2$ if $f_1 = f_2$ or, for the greatest element $y \in Y$ such that $f_1(y) \neq f_2(y)$ we have $f_1(y) < f_2(y)$.

Some helpful terminology: one has the zero function, which is 0 for all values. For every other $f \in Z(Y,X)$, we define its **degree** $y_{\deg}$ to be the largest $y \in Y$ such that $f(y) \neq 0$ and its **leading coefficient** $x_l := f(y_{\deg})$.

PROPOSITION 12.13. *For ordinalities α and β, α^β is an ordinality.*

PROOF. Let Z be the set of finitely nonzero functions $f : Y \to X$ as above, and let $W \subseteq Z$ be a nonempty subset. We may assume 0 is not in W, since the zero function is the minimal element of all of Z. Thus the set of degrees of all elements of W is nonempty, and we may choose an element of minimal degree y_1; moreover, among all elements of minimal degree we may choose one with minimal leading coefficient x_1, say f_1. Suppose f_1 is not the minimal element of W, i.e., there exists $f' \in W_2$ with $f' < f_1$. Any such f' has the same degree and leading coefficient as f_1, so the last value y' at which f' and f_1 differ must be less than y_1. Since f_1 is nonzero at all such y' and f_1 is finitely nonzero, the set of all such y' is finite and thus has a *maximal* element y_2. Among all f' with $f'(y_2) < f(y_2)$ and $f'(y) = f(y)$ for all $y > y_2$, choose one with $x_2 = f'(y_2)$ minimal and call it f_2. If f_2 is not minimal, we may continue in this way, and indeed get a sequence of elements $f_1 > f_2 > f_3 \ldots$ as well as a descending chain $y_1 > y_2 > \ldots$. Since Y is well-ordered, this descending chain must terminate at some point, meaning that at some point we find a minimal element f_n of W. □

EXAMPLE 12.14. *The ordinality ω^ω is the set of all finitely nonzero functions $f : \mathbb{N} \to \mathbb{N}$. At least formally, we can identify such functions as polynomials in ω*

with $\mathbb{N}$-coefficients: $P_f(\omega) = \sum_{n\in\mathbb{N}} f(n)\omega^n$. The well-ordering makes $P_f < P_g$ if the at the largest n for which $f(n) \neq g(n)$ we have $f(n) < g(n)$, e.g. $\omega^3+2\omega^2+1 > \omega^3+\omega^2+\omega+100$.

It is hard to ignore the following observation: ω^ω puts a natural well-ordering relation on all the ordinalities we had already defined. This makes us look back and see that the same seems to be the case for all ordinalities: e.g. ω itself is order isomorphic to the set of all the finite ordinalities $[n]$ with the obvious order relation between them. Now that we see the suggested order relation on the ordinalities of the form $P(\omega)$ one can check that this is the case for them as well: e.g. ω^2 can be realized as the set of all linear polynomials $\{a\omega + b \mid a, b \in \mathbb{N}\}$.

This suggests the following line of inquiry:

(i) Define a natural ordering on ordinalities (as we did for cardinalities).
(ii) Show that this ordering *well-orders* any set of ordinalities.

In Exercise 12.10 you are asked to show (among other things) that for ordinalities α and β with $\alpha > 0$ and β infinite, we have $\#\alpha^\beta = \max(\#\alpha, \#\beta)$. In particular we generally *do not* have that $\#\alpha^\beta = (\#\alpha)^{\#\beta}$. In particular we have not yet seen any uncountable well-ordered sets, and one cannot construct an uncountable ordinal from ω by any finite iteration of the ordinal operations we have described (nor by a countable iteration either, although we have not yet made formal sense of that). This leads us to wonder: are there any uncountable ordinalities?

2.1. Ordering ordinalities. Let S_1 and S_2 be two well-ordered sets. In analogy with our operation $\leq$ on sets, it would seem natural to define $S_1 \leq S_2$ if there exists an order-preserving injection $S_1 \to S_2$. This gives a relation $\leq$ on ordinalities which is clearly symmetric and transitive.

However, this is *not* the most useful definition of $\leq$ for well-ordered sets, since it gives up the rigidity property. In particular, recall Dedekind's characterization of infinite sets as those which are in bijection with a proper subset of themselves, or, equivalently, those which *inject* into a proper subset of themselves. With the above definition, this will still occur for infinite ordinalities: for instance, we can inject ω properly into itself just by taking $\mathbb{N} \to \mathbb{N}$, $n \mapsto n+1$. Even if we require the least elements to be preserved, then we can still inject $\mathbb{N}$ into any infinite subset of itself containing 0.

So as a sort of mild *deus ex machina* we will work with a more sophisticated order relation. First, for a linearly ordered set S and $s \in S$, we define

$$I(s) = \{t \in S \mid t < s\},$$

an **initial segment** of S. Note that every initial segment is a proper subset. Let us also define

$$I[s] = \{t \in S \mid t \leq s\}.$$

Now, given linearly ordered sets S and T, we define $S < T$ if there exists an order-preserving embedding $f : S \to T$ such that $f(S)$ is an initial segment of T (say, an **initial embedding**). We define $S \leq T$ if $S < T$ or $S \cong T$.

Definition: In a partially ordered set X, a subset Z is an **order ideal** if for all $z \in Z$ and $x \in X$, if $x < z$ then $x \in Z$. For example, the empty set and X itself are always order ideals. We say that X is an **improper** order ideal of itself, and all other order ideals are **proper**. For instance, $I[s]$ is an order ideal, which may or may not be an initial segment.

LEMMA 12.15. *("Principal ideal lemma") Any proper order ideal in a well-ordered set is an initial segment.*

PROOF. Let Z be a proper order ideal in X, and s the least element of $X \setminus Z$. Then we have $Z = I(s)$. □

The following is a key result:

THEOREM 12.16. *(Ordinal trichotomy) For any two ordinalities $\alpha = o(X)$ and $\beta = o(Y)$, exactly one of the following holds: $\alpha < \beta$, $\alpha = \beta$, $\beta < \alpha$.*

PROOF. Part of the assertion is that no well-ordered set X is order isomorphic to any initial segment $I(s)$ in X (we would then have both $o(I(s)) < o(X)$ and $o(I(s)) = o(X)$); let us establish this first. Suppose to the contrary that $\iota : X \to X$ is an order embedding whose image is an initial segment $I(s)$. Then the set of x for which $\iota(x) \neq x$ is nonempty (otherwise ι would be the identity map, and no linearly ordered set is *equal* to any of its initial segments), so let x be the least such element. Then, since ι restricted to $I(x)$ is the identity map, $\iota(I(x)) = I(x)$, so we cannot have $\iota(x) < x$ – that would contradict the injectivity of ι – so it must be the case that $\iota(x) > x$. Since $\iota(X)$ is an initial segment, this means that x is in the image of ι, which is seen to be impossible.

Now if $\alpha < \beta$ and $\beta < \alpha$ then we have initial embeddings $i : X \to Y$ and $j : Y \to X$. By Exercise 12.12 their composite $j \circ i : X \to X$ is an initial embedding, which we have just seen is impossible. It remains to show that if $\alpha \neq \beta$ there is either initial embedding from X to Y or vice versa. We may assume that X is nonempty. Let us try to build an initial embedding from X into Y. A little thought convinces us that we have no choices to make: suppose we have already defined an initial embedding f on a segment $I(s)$ of X. Then we *must* define $f(s)$ to be the least element of $Y \setminus f(I(s))$, and we *can* define it this way exactly when $f(I(s)) \neq Y$. If however $f(I(s)) = Y$, then we see that f^{-1} gives an initial embedding from Y to X. So assume Y is not isomorphic to an initial segment of X, and let Z be the set of x in X such that there exists an initial embedding from $I(z)$ to Y. It is immediate to see that Z is an order ideal, so by Lemma 12.15 we have either $Z = I(x)$ or $Z = X$. In the former case we have an initial embedding from $I(z)$ to Y, and as above, the only we could not extend it to x is if it is surjective, and then we are done as above. So we can extend the initial embedding to $I[x]$, which – again by Lemma 12.15 is either an initial segment (in which case we have a contradiction) or $I[x] = X$, in which case we are done. The last case is that $Z = X$ has no maximal element, but then we have $X = \bigcup_{x \in X} I(x)$ and a uniquely defined initial embedding ι on each $I(x)$. So altogether we have a map on all of X whose image $f(X)$, as a union of initial segments, is an order ideal. Applying Lemma 12.15 yet again, we either have $f(X) = Y$ – in which case f is an order isomorphism – or $f(X)$ is an initial segment of Y, in which case $X < Y$: done. □

We immediately deduce:

COROLLARY 12.17. *Any set of ordinalities is linearly ordered under* $\leq$.

COROLLARY 12.18. *Any set* $\mathcal{F}$ *of ordinalities is* well-ordered *with respect to* $\leq$.

PROOF. Using Proposition 12.1, it suffices to prove that there is no infinite descending chain in $\mathcal{F} = \{o_\alpha\}_{\alpha \in I}$. So, seeking a contradiction, suppose that we have a sequence of well-ordered sets S_1, $S_2 = I(s_1)$ for $s_1 \in S_1$, $S_3 = I(s_2)$,...,$S_{n+1} = I(s_n)$ for $s_n \in S_n$,.... But all the S_n's live inside S_1 and we have produced an infinite descending chain $s_1 > s_2 > s_3 > \ldots > s_n > \ldots$ inside the well-ordered set S_1, a contradiction. □

Thus any set $\mathcal{F}$ of ordinalities itself generates an ordinality $o(\mathcal{F})$, the ordinality of the well-ordering that we have just defined on $\mathcal{F}$!

Now: for any ordinality o, it makes sense to consider the set $I(o)$ of ordinalities $\{o' \mid o' < o\}$: indeed, these are well-orderings on a set of cardinality at most the cardinality of o, so there are at most $2^{\#o \times \#o}$ such well-orderings. Similarly, define

$$I[o] = \{o' \mid o' \leq o\}.$$

COROLLARY 12.19. $I(o)$ *is order-isomorphic to* o *itself.*

PROOF. We shall define an order-isomorphism $f : I(o) \to o$. Namely, each $o' \in I(o)$ is given by an initial segment $I(y)$ of o, so define $f(o') = y$. That this is an order isomorphism is essentially a tautology which we leave for the reader to unwind. □

2.2. The Burali-Forti "Paradox".

Do the ordinalities form a set? As we have so far managed to construct only countably many of them, it seems conceivable that they might. However, Burali-Forti famously observed that the assumption that there is a set of all ordinalities leads to a paradox. Namely, suppose $\mathbb{O}$ is a set whose elements are the ordinalities. Then by Corollary 12.18, we have that $\mathbb{O}$ is itself well-ordered under our initial embedding relation $\leq$, so that the ordinality $o = o(\mathbb{O})$ would itself be a member of $\mathbb{O}$.

This is already curious: it is tantamount to saying that $\mathbb{O}$ is an element of itself, but notice that we are not necessarily committed to this: $(\mathbb{O}, \leq)$ is order isomorphic to one of its members, but maybe it is not *the same* set. (Anyway, is $o \in o$ paradoxical, or just strange?) Thankfully the paradox does not depend upon these ontological questions, but is rather the following: if $o \in \mathbb{O}$, then consider the initial segment $I(o)$ of $\mathbb{O}$: we have $\mathbb{O} \cong o \cong I(o)$, but this means that $\mathbb{O}$ is order-isomorphic to one of its initial segments, in contradiction to the Ordinal Trichotomy Theorem (Theorem 12.16).

Just as the proof of Cantor's *paradox* (i.e., that the cardinalities do not form a set) can be immediately adapted to yield a profound and useful *theorem* – if S is a set, there is no surjection $S \to 2^S$, so that $2^{\#S} > \#S$ – in turn the proof of the Burali-Forti paradox immediately gives the following result, which we have so far been unable to establish:

THEOREM 12.20. *(Burali-Forti's Theorem) For any cardinal* κ*, the set* $\mathcal{O}_\kappa$ *of ordinalities* o *with* $|o| \leq \kappa$ *has cardinality greater than* κ.

PROOF. $\mathcal{O}_\kappa$ is, like any set of ordinalities, well-ordered under our relation $\leq$, so if it had cardinality at most κ it would contain its own ordinal isomorphism class o as a member and hence be isomorphic to its initial segment $I(o)$ as above. □

In particular there are uncountable ordinalities. There is thus a *least* uncountable ordinality, traditionally denoted ω_1. This least uncountable ordinality is a truly remarkable mathematical object: mere contemplation of it is fascinating and a little dizzying. For instance, the minimality property implies that all of its initial segments are countable, so it is not only very large as a set, but it is extremely difficult to traverse: for any point $x \in \omega_1$, the set of elements less than x is countable whereas the set of elements greater than x is uncountable! In particular it has no largest element, so is a limit ordinal.[1]

Its successor ω_1^+ is also of interest, as explored in Exercise 12.14.

3. Von Neumann ordinals

Here we wish to report on an idea of von Neumann, which uses the relation $I(o) \cong o$ to define a canonical well-ordered set with any given ordinality. The construction is often informally defined as follows: "we inductively define o to be the set of all ordinals less than o." Unfortunately this definition is circular, and not for reasons relating to the induction process: step back and see that it is circular in the most obvious sense of using the quantity it purports to define!

However, it is quite corrigible: rather than building ordinals out of nothing, we consider the construction as taking as input a well-ordered set S and returning an order-isomorphic well-ordered set $vo(S)$, the **von Neumann ordinal** of S. The only property that we wish it to have is the following: if S and T are order-isomorphic sets, we want $vo(S)$ and $vo(T)$ to be not just order-isomorphic but *equal*. Let us be a bit formal and write down some axioms:

(VN1) For all well-ordered sets S, we have $vo(S) \cong S$.
(VN2) For well-ordered S and T, $S \cong T \implies vo(S) = vo(T)$.

Consider the following two additional axioms:

(VN3) $vo(\emptyset) = \emptyset$.
(VN4) For $S \neq \emptyset$, $vo(S) = \{vo(S') \mid S' < S\}$.

The third axiom is more than reasonable: it is forced upon us, by the fact that there is a unique empty well-ordered set. The fourth axiom is just expressing the order-isomorphism $I(o) \cong o$ in terms of von Neumann ordinals. Now the point is that these axioms determine all the von Neumann ordinals:

THEOREM 12.21. *(von Neumann) There is a unique correspondence $S \mapsto vo(S)$ satisfying (VN1) and (VN2).*

Before proving this theorem, let's play around with the axioms by discussing their consequences for finite ordinals. We know that $vo(\emptyset) = \emptyset = [0]$. What is $vo([1])$?

[1] In fact this only begins to express ω_1's "inaccessibility from the left"; the correct concept, that of **cofinality**, will be discussed later.

Well, it is supposed to be the set of von Neumann ordinals strictly less than it. There is in all of creation exactly one well-ordered set which is strictly less than [1]: it is $\emptyset$. So the axioms imply

$$vo([1]) = \{\emptyset\}.$$

How about $vo([2])$? The axioms easily yield:

$$vo([2]) = \{vo[0], vo[1]\} = \{\emptyset, \{\emptyset\}\}.$$

Similarly, for any finite number n, the axioms give:

$$v0([n]) = \{vo[0], vo[1], \ldots, vo[n-1]\},$$

or in other words,

$$vo([n]) = \{vo[n-1], \{vo[n-1]\}\}.$$

More interestingly, the axioms tell us that the von Neumann ordinal ω is precisely the set of all the von Neumann numbers attached to the natural numbers. And we can track this construction "by hand" up through the von Neumann ordinals of 2ω, ω^2, ω^ω and so forth. But how do we know the construction works (i.e., gives a unique answer) for every ordinality?

The answer is simple: by induction. We have seen that the axioms imply that at least for sufficiently small ordinalities there is a unique assignment $S \mapsto vo(S)$. If the construction does not always work, there will be a smallest ordinality o for which it fails. But this cannot be, since it is clear how to define $vo(o)$ given definitions of all von Neumann ordinals of ordinalities less than o: indeed, (VN4) tells us exactly how to do this.

This construction is an instance of **transfinite induction**. This is the extension to general well-ordered sets of the principle of complete induction for the natural numbers: if S is a well-ordered set and T is a subset which is (i) nonempty and (ii) for all $sinS$, if the order ideal $I(s)$ is contained in T, then s is in T; then T must in fact be all of S. We trust the proof is clear.

Note that transfinite induction generalizes the principle of *complete* induction, not the principle of mathematical induction which says that if 0 is in S and $n \in S \implies n+1 \in S$, then $S = \mathbb{N}$. This principle is not valid for any ordinality larger than ω, since indeed ω is (canonically) an initial segment of every larger ordinality and the usual axioms of induction are satisfies for ω itself. All this is to say that in most applications of transfinite induction one must distinguish between the case of successor ordinals and the case of limit ordinals. For example:

We should remark that this is not a foundationalist treatment of von Neumann ordinals. It would also be possible to define a von Neumann ordinal as a certain type of set, using the following exercise.

Exercise 1.5.2: Show that a set S is a von Neumann ordinal iff:
(i) if $x \in S$ implies $x \subseteq S$;
(ii) the relation $\subset$ is a well-ordering on elements of S.

For the rest of these notes we will drop the term "ordinality" in favor of "ordinal." The reader who wants an ordinal to be something in particular can thus take it to be a von Neumann ordinal. This convention has to my knowledge no real mathematical advantage, but it has some very convenient notational consequences, as for instance the following definition of "cardinal."

3.1. A definition of cardinals. Here we allow ourselves the following result, which we will discuss in more detail later on.

THEOREM 12.22. *(Well-ordering theorem) Assuming the Axiom of Choice, every set S can be well-ordered.*

We can use this theorem ("theorem"?) to reduce the theory of cardinalities to a special case of the theory of ordinalities, and thus, we can give a concrete definition of cardinal numbers in terms of Von Neumann's ordinal numbers.

Namely, for any set S, we define its cardinal $|S|$ to be the smallest von Neumann ordinal o such that o is equivalent to (i.e., in bijection with) S.

In particular, we find that the finite cardinals and the finite ordinals are the same: we have changed our standard n element set from $[1, n]$ to the von Neumann ordinal n, so for instance $3 = \{\emptyset, \{\emptyset\}, \{\emptyset, \{\emptyset\}\}\}$. On purely mathematical grounds, this is not very exciting. However, if you like, we can replace our previous attitude to what the set $[n] = \{1, \ldots, n\}$ "really is" (which was, essentially, "Why are you bothering me with such silly questions?") by saying that, in case anyone asks (we may still hope that they do not ask), we identify the non-negative integer n with its von Neumann ordinal. Again, this is not to say that we have discovered what 3 really is. Rather, we noticed that a set with three elements exists in the context of **pure set theory**, i.e., we do not have to know that there exist 3 objects in some box somewhere that we are basing our definition of 3 on (like the definition of a meter used to be based upon an actual meter stick kept by the Bureau of Standards). In truth 3 is not a very problematic number, but consider instead $n = 10^{10^{10^{10}}}$; the fact that n is (perhaps) greater than the number of distinct particles in the universe is, in our account, no obstacle to the existence of sets with n elements.

Let's not overstate the significance of this for finite sets: with anything like a mainstream opinion on mathematical objects[2] this is completely obvious: we could also have defined 0 as $\emptyset$ and n as $\{n - 1\}$, or in infinitely many other ways. It becomes more interesting for infinite sets, though.

That is, we can construct a theory of sets without *individuals* – in which we never have to say what we mean by an "object" as an element of a set, because the only elements of a set are other sets, which ultimately, when broken up enough (but possibly infinitely many) times, are lots and lots of braces around the empty set. This is nice to know, most of all because it means that in practice we don't have to worry one bit about what the elements of are sets are: we can take them to be whatever we want, because each set is equivalent (bijective) to a *pure set*. If you would like (as I would) to take a primarily Bourbakistic view of mathematical structure – i.e., that the component parts of any mathematical object are of no

[2]The only contemporary mathematician I know who would have problems with this is Doron Zeilberger.

importance whatsoever, and that mathematical objects matter only as they relate to each other – then this is very comforting.

Coming back to the mathematics, we see then that any set of cardinals is in particular a set of ordinals, and the notion of $<$ on cardinals induced in this way is the same as the one we defined before. That is, if α and β are von Neumann cardinals, then $\alpha < \beta$ holds in the sense of ordinals if and only if there exists an injection from α to β but not an injection from β to α.

Thus we have now, at last, proved the Second Fundamental Theorem of Set Theory, modulo our discussion of Theorem 12.22.

4. The Axiom of Choice and some of its equivalents

4.1. Introducing the Axiom of Choice.

Now we come clean. Many of the results of Chapter II rely on the following "fact":

FACT 12.23. *(Axiom of Choice (AC)): For any nonempty family I of nonempty sets S_i, the product $\prod_{i \in I} S_i$ is nonempty.*

Remark: In other words, any product of nonzero cardinalities is itself nonzero. This is the version of the axiom of choice favored by Bertrand Russell, who called it the "multiplicative axiom." Aesthetically speaking, I like it as well, because it seems so simple and self-evident.

Exercise 2.1: Show that if (AC) holds for all families of pairwise disjoint sets S_i, it holds for all nonempty families of nonempty sets.

However, in applications it is often more convenient to use the following reformulation of (AC) which spells out the connection with "choice".

(AC$'$): If S is a set and $I = \{S_i\}$ is a nonempty family of nonempty subsets of S, then there exists a **choice function**, i.e., a function $f : I \to S$ such that for all $i \in I$, $f(S_i) \in S_i$.

Let us verify the equivalence of (AC) and (AC$'$).

(AC) $\implies$ (AC$'$): By (AC), $\mathcal{S} = \prod_{i \in I} S_i$ is nonempty, and an element f of $\mathcal{S}$ is precisely an assignment to each $i \in I$ of an element $f(i) \in S_i \subseteq S$. Thus f determines a choice function $f : I \to S$.

(AC$'$) $\implies$ (AC): Let $I = \{S_i\}$ be a nonempty family of nonempty sets. Put $S = \bigcup_{i \in I} S_i$. Let $f : I \to S$ be a choice function: for all $i \in I$, $f(S_i) \in S_i$. Thus $\{f(i)\}_{i \in I} \in \prod_{i \in I} S_i$.

The issue here is that if I is infinite we are making infinitely many choices – possibly with no coherence or defining rule to them – so that to give a choice function f is in general to give an infinite amount of information. Have any of us in our daily lives ever made infinitely many independent choices? Probably not. So the worry that making such a collection of choices is not possible is not absurd and should be taken with some seriousness.

Thus the nomenclature *Axiom* of Choice: we are, in fact, asserting some feeling about how infinite sets behave, i.e., we are doing exactly the sort of thing we had earlier averred to try to avoid. However, in favor of assuming AC, we can say: (i) it is a fairly basic and reasonable axiom – if we accept it we do not, e.g., feel the need to justify it in terms of something simpler; and (ii) we are committed to it, because most of the results we presented in Chapter II would not be true without it, nor would a great deal of the results of mainstream mathematics.

Every student of mathematics should be aware of some of the "facts" that are equivalent to AC. The most important two are as follows:

FACT 12.24. *(Zorn's Lemma) Let S be a partially ordered set. Suppose that every chain C – i.e., a totally ordered subset of S – has an upper bound in S. Then S has a maximal element.*

THEOREM 12.25. *The axiom of choice (AC), Zorn's Lemma (ZL), and the Well-Ordering Theorem (WOT) are all equivalent to each other.*

Remark: The fact that we are asserting the logical equivalence of an axiom, a lemma and a theorem is an amusing historical accident: according to the theorem they are all on the same logical footing.

Well-Ordering Theorem implies Axiom of Choice: It is enough to show WOT $\implies$ AC$'$, which is easy: let $\{S_i\}_{i \in I}$ be a nonempty family of nonempty subsets of a set S. Well-order S. Then we may define a choice function $f : I \to S$ by mapping i to the least element of S_i.

Axiom of Choice implies Zorn's Lemma: Strangely enough, this proof will use transfinite induction (so that one might initially think WOT would be involved, but this is absolutely not the case). Namely, suppose that S is a poset in which each chain C contains an upper bound, but there is no maximal element. Then we can define, for every ordinal o, a subset $C_0 \subseteq S$ order-isomorphic to o, in such a way that if $o' < o$, $C_{o'} \subseteq C_o$. Indeed we define $C_\emptyset = \emptyset$, of course. Assume that for all $o' < o$ we have defined $C_{o'}$. If o is a limit ordinal then we define $C_o := \bigcup_{o' < o} C_{o'}$. Then necessarily C_0 is order-isomorphic to o: that's how limit ordinals work. If $o = o' + 1$, then we have $C_{o'}$ which is assumed not to be maximal, so we choose an element x of $S \setminus C_{o'}$ and define $x_o := x$. Thus we have inside of S well-ordered sets of all possible order-isomorphism types. This is clearly absurd: the collection $o(|S|)$ of ordinals of cardinality $|S|$ is an ordinal of cardinality greater than the cardinality of S, and $o(|S|) \hookrightarrow S$ is impossible.

But where did we use AC? Well, we definitely made some choices, one for each non-successor ordinal. To really nail things down we should cast our choices in the framework of a choice function. Suppose we choose, for each well-ordered subset W of X, an element $x_W \in X \setminus W$ which is an upper bound for W. (This is easily phrased in terms of a choice function.) We might worry for a second that in the above construction there was some compatibility condition imposed on our choices, but this is not in fact the case: at stage o, any upper bound x for C_o in $S \setminus C_o$ will do to give us $C_{o+1} := C_o \cup \{x\}$. This completes the proof.

Remark: Note that we showed something (apparently) slightly stronger: namely,

that if every well-ordered subset of a poset S has an upper bound in S, then S has a maximal element. This is mildly interesting but apparently useless in practice.

Zorn's Lemma implies Well-Ordering Theorem: Let X be a non-empty set, and let $\mathcal{A}$ be the collection of pairs $(A, \leq)$ where $A \subseteq X$ and $\leq$ is a well-ordering on A. We define a relation $<$ on $\mathcal{A}$: $x < y$ if and only if x is equal to an initial segment of y. It is immediate that $<$ is a strict partial ordering on $\mathcal{A}$. Now for each chain $C \subset \mathcal{A}$, we can define x_C to be the union of the elements of C, with the induced relation. x_C is itself well-ordered with the induced relation: indeed, suppose Y is a nonempty subset of x_C which is not well-ordered. Then Y contains an infinite descending chain $p_1 > p_2 > \ldots > p_n > \ldots$. But taking an element $y \in C$ such that $p_1 \in y$, this chain lives entirely inside y (since otherwise $p_n \in y'$ for $y' > y$ and then y is an initial segment of y', so $p_n \in y'$, $p_n < p_1$ implies $p_n \in y$), a contradiction.

Applying Zorn's Lemma, we are entitled to a maximal element $(M, \leq_M)$ of $\mathcal{A}$. It remains to see that $M = X$. If not, take $x \in X \setminus M$; adjoining x to $(M, \leq_M)$ as the maximum element we get a strictly larger well-ordering, a contradiction.

Remark: In the proof of AC $\implies$ ZL we made good advantage of our theory of ordinal arithmetic. It is possible to prove this implication (or even the direct implication AC $\implies$ ZL) directly, but this essentially requires proving some of our lemmata on well-ordered sets on the fly.

4.2. Some equivalents and consequences of the Axiom of Choice. Although disbelieving AC is a tenable position, mainstream mathematics makes this position slightly unpleasant, because Zorn's Lemma is used to prove many quite basic results. One can ask which of these uses are "essential." The strongest possible case is if the result we prove using ZL can itself be shown to imply ZL or AC. Here are some samples of these results:

FACT 12.26. *For any infinite set A, we have $\#A = \#A \times \#A$.*

FACT 12.27. *For sets A and B, there is an injection $A \hookrightarrow B$ or an injection $B \hookrightarrow A$.*

FACT 12.28. *Every surjective map of sets has a section.*

FACT 12.29. *For any field k, every k-vector space V has a basis.*

FACT 12.30. *Every proper ideal in a commutative ring is contained in a maximal proper ideal.*

FACT 12.31. *The product of any number of compact spaces is itself compact.*

Even more commonly one finds that one can make a proof work using Zorn's Lemma but it is not clear how to make it work without it. In other words, many statements seem to require AC even if they are not equivalent to it. As a simple example, try to give an explicit well-ordering of $\mathbb{R}$. Did you succeed? In a precise formal sense this is impossible. But it is intuitively clear (and also true!) that being able to well-order a set S of any given infinite cardinality is not going to tell us that we can well-order sets of all cardinalities (and in particular, how to well-order 2^S), so the existence of a well-ordering of the continuum is not equivalent to AC.

Formally, speaking one says that a statement *requires* AC if one cannot prove that

statement in the Zermelo-Fraenkel axiomation of set theory (ZF) which excludes AC. (The Zermelo-Fraenkel axiomatization of set theory including the axiom of choice is abbreviated ZFC; ZFC is essentially the "standard model" for sets.) If on the other hand a statement requires AC in this sense but one cannot deduce AC from ZF and this statement, we will say that the statement *merely requires* AC. There are lots of statements that merely require AC:[3]

THEOREM 12.32. *The following facts merely require AC:*

a) *The countable union of countable sets is countable.*
b) *Every infinite set is Dedekind infinite.*
c) *There is a non(-Lebesgue-)measurable subset of* $\mathbb{R}$.
d) *The Banach-Tarski paradox.*
e) *Every field has an algebraic closure.*
f) *Every field extension has a relative transcendence basis.*
g) *Every Boolean algebra contains a prime ideal (BPIT).*
h) *Every Boolean algebra is isomorphic to a Boolean algebra of sets (Stone representation theorem).*
i) *Every subgroup of a free group is free.*
j) *The Hahn-Banach theorem (on extension of linear functionals), the open mapping theorem, the closed graph theorem, the Banach-Alaoglu theorem.*
k) *The Baire category theorem.*
l) *The existence of a Stone-Cech compactification of every completely regular space.*

Needless to say the web of implications among all these important theorems is a much more complicated picture; for instance, it turns out that the BPIT is an interesting intermediate point (e.g. Tychonoff's theorem for Hausdorff spaces is equivalent to BPIT). Much contemporary mathematics is involved in working out the various dependencies.

In summary, if your beliefs about sets are the same as the standard ones except that you do not admit any form of AC (again, exactly what this means is something that we have not spelled out), then you will find that there is an amazing array of mathematical theorems that you will not be able to prove. If instead of being entirely agnostic about AC you believe a strong enough condemnation of it (i.e., you believe one of the many axioms which is independent of ZF and contradicts AC), then you will be able to prove false some of the results in standard mathematics.

Notable here is the existence of a relatively mild denial of AC which allows most familiar analytic results to remain true but implies that every subset $\mathbb{R}$ is Lebesgue measurable. There are analysts who advocate the use of this axiom, noting that it simplifies the theory: in proving Fubini-type theorems on integrals over product measure spaces, one has to verify that the measurability of the given functions implies the measurability of certain auxiliary functions, a verification which is tedious and unpleasant (and nontrivial). Like most people who lost an hour of their lives somewhere in their early 20's sitting through the proof of Fubini's theorem, I have some sympathy for this position.

What should your attitude be towards AC? You will, of course, have to decide

[3]This list was compiled with the help of the Wikipedia page on the Axiom of Choice.

for yourself, although again a sincere agnosticism or disbelief could lead you to state and prove different theorems. My own take on AC (which is rather standard to the extent of coming uncomfortably close to parroting the corresponding paragraph in Kaplansky's book [**K**], but it is nevertheless how I feel) is a sort of middle-ground: when you use a result which requires (merely or otherwise) AC, you should acknowledge this – not necessarily with a large fanfare; if you used Zorn's Lemma somewhere it is plausible that your result requires AC, whether or not the set theorists have proven its independence from ZF – and take mental note: it means that there is some obstacle to making your result explicit in full generality. Now if you are working in some fairly concrete area of mathematics (like number theory), perhaps there are some interesting special cases of your general result which you might be able to make explicit with a different and more perspicuous argument. In general when you prove a theorem asserting the existence of an object, it is good to know whether or not you can actually *construct*, in some algorithmic sense, such an object. The advent of computers has done wonders for constructive mathematics, a philosophy which only 50 years ago looked rather eccentric.

One thing that the majority of working mathematicians would probably agree with is that while uncountable sets exist in the sense of convenience and noncontradiction, they do not exist in the same *visceral* sense of things that you can get your computer to spit out. Twenty-first century mathematics is at the same time more abstract *and* more concrete than mathematics one hundred years before.

5. A Universal Countable Ordered Set

If X and Y are totally ordered sets, an **order embedding** $\iota : X \hookrightarrow Y$ is an injective order-preserving map: equivalently, for all $x_1, x_2 \in X$, if $x_1 < x_2$ then $\iota(x_1) < \iota(x_2)$.

Let $\mathcal{F}$ be a set of linearly ordered sets. We say that a totally ordered set Y if $\mathcal{F}$-**universal** if for all $X \in \mathcal{F}$ there is an order embedding $X \hookrightarrow Y$. In studying the class of $\mathcal{F}$-universal ordered sets, we may pass from $\mathcal{F}$ to any subset $\mathcal{F}'$ in which every order-isomorphism class of elements of $\mathcal{F}$ appears exactly once: then for any linearly ordered set Y, Y is $\mathcal{F}$-universal if and only if it is $\mathcal{F}'$-universal. So without loss of generality we may, and shall, assume that the elements of $\mathcal{F}$ are pairwise non-order-isomorphic.

It is not hard to see that for any set $\mathcal{F}$ of linearly ordered sets, $\mathcal{F}$-universal subsets exist. Indeed, if we linearly order the elements X_f of $\mathcal{F}$, we may take $Y \coloneqq \bigoplus_{f \in \mathcal{F}} X_f$ to be the *ordered sum* of the sets X_f: this is an ordering on the disjoint union $\coprod_{f \in \mathcal{F}} X_f$ that restricts to the given ordering on each X_f and for $f_1 \neq f_2$ and $x_1 \in X_{f_1}$ and $x_2 \in X_{f_2}$ we put $x_1 < x_2$ if and only if $f_1 < f_2$. However, if each element of $\mathcal{F}$ satisfies a certain property, we may ask for an $\mathcal{F}$-universal set that has that property, and this construction may or may not give that: for instance, if each element of $\mathcal{F}$ is well-ordered, then $\bigoplus_{f \in \mathcal{F}} X_f$ is well-ordered if and only if the chosen ordering on $\mathcal{F}$ is a well-ordering. We may also be able to find a smaller $\mathcal{F}$-universal subset.

PROPOSITION 12.33. *Let $\mathcal{F}_f$ be the family of all finite ordinals. Then a linearly ordered set Y is $\mathcal{F}_f$-universal if and only if it is infinite.*

PROOF. If Y is finite of cardinality n, then the finite ordinal $[n+1]$ does not order-embed in Y. Conversely, if Y is infinite, then for all $n \in \mathbb{N}$ it contains a subset Y_n of cardinality n, and there is a unique order isomorphism $\iota : [n] \to Y_n$. □

Now consider a set $\mathcal{F}_c$ such that every countable linearly ordered set is order-isomorphic to exactly one element of $\mathcal{F}_c$, and consider ω_1, the set of all countable ordinals. In Exercise 12.17 you are asked to show that if X and Y are well-ordered sets, then there is an order embedding from X to Y if and only if $X \leq Y$: that is, X is order-isomorphic to an initial segment of Y. It follows that no countable ordinal o if ω_1-universal: we would then have to have $\alpha \leq o$ for every $\alpha \in \omega_1$ and thus $\omega_1 \leq o$, so o is uncountable. Moreover, because there are uncountably many countable ordinals, the ω_1-universal ordered set $\bigoplus_{\kappa<\omega_1} \kappa$ mentioned above has continuum cardinality, as does $\bigoplus_{X_f \in \mathcal{F}_c} X_f$. We claim however that there is a countable linearly ordered set that is not just ω_1-universal but even $\mathcal{F}_c$-universal.

THEOREM 12.34. *For a linearly ordered set Y, the following are equivalent:*

(i) *The set Y is $\mathcal{F}_c$-universal: that is, it contains an order-isomorphic copy of every countable linearly ordered set.*
(ii) *There is an order embedding $\iota : \mathbb{Q} \hookrightarrow Y$.*

PROOF. (i) $\implies$ (ii) is immediate: since $\mathbb{Q}$ (with its standard ordering) is a countable lienearly ordered set, if Y is $\mathcal{F}_c$-universal, then by definition we must have an order embedding $\iota : \mathbb{Q} \hookrightarrow Y$.
(ii) $\implies$ (i): In general, if $\mathcal{F}$ is a set of linearly ordered sets, X is $\mathcal{F}$-universal and X order embeds in Y, then just because a composition of order embeddings is an order embedding it follows that also Y is $\mathcal{F}$-universal. So it suffices to show that $\mathbb{Q}$ is $\mathcal{F}_c$-universal.

By Proposition 12.33, every finite linearly ordered set order embeds in $\mathbb{Q}$, so let $X = \{x_n \mid n \in \mathbb{Z}^+\}$ be a countably infinite linearly ordered set. Let $r : \mathbb{Z}^+ \to \mathbb{Q}$ be a bijection. Having made these choices we will now define a specific order embedding from X into $\mathbb{Q}$.
Step 1: We map x_1 to r_1, and put $n_1 := 1$.
Step 2: If $x_2 < x_1$, let n_2 be the least positive integer such that $r_{n_2} < r_1$, and we map x_2 to r_{n_2}. If $x_2 > x_1$, let n_2 be the least positive integer such that $r_{n_2} > r_1$, and we may x_2 to r_{n_2}.
Step N+1: Now let $N \geq 2$ and suppose that we have defined an order embedding ι from $\{x_1, \ldots, x_N\}$ into $\mathbb{Q}$ and we put $r_{n_i} = \iota(x_i)$ for $1 \leq i \leq N$. Then $\mathbb{Q} \setminus \{r_{n_1}, \ldots, r_{n_N}\}$ is a union of $N+2$ open intervals $I_1, \ldots, I_{N+1}$ in $\mathbb{Q}$, the first and last being unbounded, and each of these intervals containing infinitely many points of $\mathbb{Q}$. There is a unique $1 \leq j \leq N+1$ such that if we map x_{N+1} to the least integer n_{N+1} such that $r_{n_{N+1}} \in I_j$, then $\iota : \{x_1, \ldots, x_{N+1}\} \to \mathbb{Q}$ is an order-isomorphism: indeed, this j is one more than $\#\{1 \leq i \leq N \mid x_{N+1} > x_i\}$.

This defines a mapping $\iota : X \hookrightarrow \mathbb{Q}$. A map of linearly ordered sets is an order embedding if and only if its restriction to each finite subset is an order embedding, and every finite subset of X is contained in $\{x_1, \ldots, x_N\}$ for some N, so ι is an order embedding. □

Since $\mathbb{Q}$ itself is countable, the linearly ordered sets that order embed in $\mathbb{Q}$ are precisely the countable ones. If we pass from $\mathbb{Q}$ to $\mathbb{R}$, then since $\mathbb{Q}$ order embeds

in $\mathbb{R}$, it follows from Theorem 12.34 that every countable linearly ordered set order embeds in $\mathbb{R}$. The following consequence of this is already very striking:

COROLLARY 12.35. *Every countable ordinal can be order embedded into $\mathbb{R}$. Thus the number of isomorphism types of well-ordered subsets of $\mathbb{R}$ is uncountable.*

On the other hand, clearly some uncountable linearly ordered subsets embed in $\mathbb{R}$...e.g. $\mathbb{R}$ itself! This makes one wonder whether any uncountable ordinals order embed into $\mathbb{R}$. The answer is negative:

THEOREM 12.36. *Let X be a well-ordered set, and let $\iota : X \to \mathbb{R}$ be an order-embeddding. Then X is countable.*

PROOF. Without loss of generality, we may assume that X is an ordinal. If X is a successor ordinal – i.e., if X has a maximum – then from the order embedding $\iota : X \hookrightarrow \mathbb{R}$ we can easily build an embedding $\tilde{\iota} : X + \omega \to \mathbb{R}$, so we may assume that X is an infinite limit ordinal. For $x \in X$, we denote x' by the least element of x that is larger than x.

For every $x \in X$, we have $\iota(x) < \iota(x')$; since $\mathbb{Q}$ is dense in $\mathbb{R}$, there is a rational number r_x such that $\iota(x) < r_x < \iota(x')$. We claim that $x \mapsto r_x$ defines an injection from X to $\mathbb{Q}$; if so, we get that X is countable, completing the proof. As for the claim: from r_x we can recover x as the largest element $y \in X$ such that $\iota(y) < r_x$, so the map must be injective. □

6. Exercises

EXERCISE 12.1. *Prove Proposition 12.1.*

EXERCISE 12.2. *Let $f : S \to T$ and $g : T \to U$ be order homomorphisms of partially ordered sets.*

a) *Show that $g \circ f : S \to U$ is an order homomorphism.*
b) *Note that the identity map from a partially ordered set to itself is an order homomorphism. (It follows that there is a **category** whose objects are partially ordered sets and whose morphisms are order homomorphisms.)*

EXERCISE 12.3. *Find partially ordered sets $(X, \leq)$, $(Y, \leq)$ and an order-preserving bijection $f : X \to Y$ such that $f^{-1} : Y \to X$ is not order-preserving.*

EXERCISE 12.4. *Prove Lemma 12.2.*

EXERCISE 12.5. *Let S be a partially ordered set.*

a) *Show that the order isomorphisms $f : S \to S$ form a group, the **order automorphism group** $\mathrm{Aut}(S)$ of S. (The same holds for any object in any category.)*
b) *Notice that Lemma 12.3 implies that the automorphism group of a well-ordered set is the trivial group.*[4]
c) *Suppose S is linearly ordered and f is an order automorphism of S such that for some positive integer n we have $f^n = \mathrm{Id}_S$, the identity map. Show that $f = \mathrm{Id}_S$.*
*(Thus the automorphism group of a linearly ordered set is **torsionfree**.)*
d) *For any infinite cardinality κ, find a linearly ordered set S with $\#\,\mathrm{Aut}(S) \geq \kappa$. Can we always ensure equality?*

[4]One says that a structure is **rigid** if it has no nontrivial automorphisms.

e) *Show that every group G is isomorphic to the automorphism group of some partially ordered set.*

EXERCISE 12.6. *For any ordering $\leq$ on a set X, we have the opposite ordering $\leq'$, defined by $x \leq' y$ if and only if $y \leq x$.*

a) *If $\leq$ is a linear ordering, so is $\leq'$.*
b) *If both $\leq$ and $\leq'$ are well-orderings, then X is finite.*

EXERCISE 12.7. *Prove Proposition 12.9.*

EXERCISE 12.8. *Prove Proposition 12.11.*

EXERCISE 12.9. *Let $\alpha_1 = o(X_1), \ldots, \alpha_n = o(X_n)$ be ordinalities.*

a) *Show that $\alpha_1 \times (\alpha_2 \times \alpha_3)$ and $(\alpha_1 \times \alpha_2) \times \alpha_3$ are each order isomorphic to the set $X_1 \times X_2 \times X_3$ endowed with the ordering $(x_1, x_2, x_3) \leq (y_1, y_2, y_3)$ if $x_1 < y_1$ or ($x_1 = y_1$ and ($x_2 < y_2$ or ($x_2 = y_2$ and $x_3 \leq y_3$))). In particular ordinal multiplication is associative.*
b) *Give an explicit definition of the product well-ordering on $X_1 \times \ldots \times X_n$, another "lexicographic ordering."*

EXERCISE 12.10. *Let α and β be ordinalities.*

a *) Show that $0^\beta = 0$, $1^\beta = 1$, $\alpha^0 = 1$, $\alpha^1 = \alpha$.*
b) *Show that the correspondence between finite ordinals and natural numbers respects exponentiation.*
c) *For an ordinal α, the symbol α^n now has two possible meanings: exponentiation and iterated multiplication. Show that the two ordinalities are equal. (The proof requires you to surmount a small left-to-right lexicographic difficulty.)*
d) *If $\alpha > 0$ and β is infinite, show: $\#(\alpha^\beta) = \max(\#\alpha, \#\beta)$.*

EXERCISE 12.11. *Let $f : S_1 \to S_2$ and $g : T_1 \to T_2$ be order isomorphisms of linearly ordered sets.*

a) *Suppose $s \in S_1$. Show that $f(I(s)) = I(f(s))$ and $f(I[s])) = I(f[s])$.*
b) *Suppose that $S_1 < T_1$ (resp. $S_1 \leq T_1$). Show that $S_2 < T_2$ (resp. $S_2 \leq T_2$).*
c) *Deduce that $<$ and $\leq$ give well-defined relations on any set $\mathcal{F}$ of ordinalities.*

EXERCISE 12.12.

a) *Show that if $i : X \to Y$ and $j : Y \to Z$ are initial embeddings of linearly ordered sets, then $j \circ i : X \to Z$ is an initial embedding.*
b) *Deduce that the relation $<$ on any set of ordinalities is transitive.*

EXERCISE 12.13. *Let α and β be ordinalities. Show that if $\#\alpha > \#\beta$, then $\alpha > \beta$. (Of course the converse does not hold: there are many countable ordinalities.)*

EXERCISE 12.14. *Let S be a totally ordered set. We endow S with the **order topology**, which is the topology generated by by infinite rays of the form*

$$(a, \infty) = \{s \in S \mid a < s\}$$

and

$$(-\infty, b) = \{s \in S \mid s < b\}.$$

Equivalently, the open intervals $(a, b) = (a, \infty) \cap (-\infty, b)$ together with the above rays and $X = (-\infty, \infty)$[5] form a basis for the topology. A topological space which arises (up to homeomorphism, of course) from this construction is called a ***linearly ordered space****.*

a) *Show that the order topology on an ordinal o is discrete if and only if $o \leq \omega$. What is the order topology on $\omega + 1$? On 2ω?*
b) *Show that order topologies are Hausdorff.*
c) *Show that an ordinality is compact if and only if it is a successor ordinality. In particular $I[o]$ is the one-point compactification of $I(o) \cong o$; deduce that the order topology on an ordinality is Tychonoff.*
d) *Show that, in fact, any linearly ordered space is normal, and moreover all subspaces are normal.*
e) *A subset Y of a linearly ordered set X can be endowed with two topologies: the subspace topology, and the order topology for the ordering on X restricted to Y. Show that the subspace topology is always finer than the order topology; by contemplating $X = \mathbb{R}$, $Y = \{-1\} \cup \{\frac{1}{n}\}_{n \in \mathbb{Z}^+}$ show that the two topologies need not coincide.*
f) *Show that it may happen that a subspace of a linearly ordered space need not be a linearly ordered space (i.e., there may be* no *ordering inducing the subspace topology). Suggestion: take $X = \mathbb{R}$, $Y = \{-1\} \cup (0, 1)$. One therefore has the notion of a* ***generalized order space****, which is a space homeomorphic to a subspace of a linearly ordered space. Show that no real manifold of dimension greater than one is a generalized order space.*
g) *Let X be a well-ordered set and Y a nonempty subset. Show that the embedding $Y \to X$ may be viewed as a* net *on X, indexed by the (nonempty well-ordered, hence directed) set Y. Show that for any ordinality o the net $I(o)$ in $I[o]$ converges to o.*

EXERCISE 12.15. *Let $\mathcal{F}$ be a set of ordinalities. As we have seen, $\mathcal{F}$ is well-ordered under our initial embedding relation $<$ so gives rise to an ordinality $o(\mathcal{F})$. In fact there is another way to attach an ordinality to $\mathcal{F}$.*

a) *Show that there is a least ordinality s such that $\alpha \leq s$ for all $\alpha \in \mathcal{F}$. (Write $\alpha = o(X_\alpha)$ and apply the Burali-Forti theorem to $\#2^{\coprod_{\alpha \in \mathcal{F}} X_\alpha}$.) We call this s the* ***ordinal supremum*** *of the ordinalities in $\mathcal{F}$.*
b) *Show that an ordinality is a limit ordinality if and only if it is the supremum of all smaller ordinalities.*
c) *Recall that a subset T of a partially ordered set S is* ***cofinal*** *if for all $s \in S$ there exists $t \in T$ such that $s \leq t$. Let α be a limit ordinality, and $\mathcal{F}$ a subset of $I(\alpha)$. Show that $\mathcal{F}$ is cofinal if and only if $\alpha = \sup \mathcal{F}$.*
d) *For any ordinality α, we define the* ***cofinality*** $\mathrm{cf}(\alpha)$ *to be the minimal ordinality of a cofinal subset $\mathcal{F}$ of $I(\alpha)$. E.g., an ordinality is a successor ordinality if and only if it has cofinality 1. Show that $\mathrm{cf}(\omega) = \omega$ and $\mathrm{cf}(\omega_1) = \mathrm{cf}(\omega_1)$. What is $\mathrm{cf}(\omega^2)$?*
e) *An ordinality is said to be* ***regular*** *if it is equal to its own cofinality. Show that for every cardinality κ, there exists a regular ordinality o with $|o| > \kappa$.*

[5]This calculus-style interval notation is horrible when S has a maximal or minimal element, since it – quite incorrectly! – seems to indicate that these elements "$\pm\infty$" should be excluded. We will not use the notation enough to have a chance to get tripped up, but beware.

f) *For a cardinality κ, let o be a regular ordinality with $|o| > \kappa$. Show that any linearly ordered subset of cardinality at most κ has an upper bound in o, but $I(\kappa)$ does not have a maximal element.*[6]

EXERCISE 12.16. *Show that for any well-ordered set S, we have*

$$\mathrm{vo}(S^+) = \{\mathrm{vo}(S), \{\mathrm{vo}(S)\}\}.$$

EXERCISE 12.17. *Let X be a well-ordered set, and let Y be a subset of X, so Y is also well-ordered with the induced ordering. Show: $Y < X$, i.e., Y is an order-isomorphic to an initial segment of X.*

EXERCISE 12.18. *The notion of $\mathcal{F}$-universality can be extended to partially ordered sets. If $\mathcal{F}$ is a set of* linearly *ordered sets and Y is an $\mathcal{F}$-universal* partially *ordered set, one may wonder if Y necessarily has a $\mathcal{F}$-universal linearly ordered subset. Let $\mathcal{F}_c$ be the set of all finite ordinals. Show: there is an $\mathcal{F}_c$-universal partially ordered set that has no $\mathcal{F}_c$-universal linearly ordered subset. (Hint: use Proposition 12.33.)*

[6]This shows that one must allow chains of arbitrary cardinalities, and not simply ascending sequences, in order for Zorn's Lemma to hold.

Bibliography

[Ac00] F. Acerbi, *Plato: Parmenides 149a7-c3. A Proof by Complete Induction?* Archive for History of the Exact Sciences 55 (2000), 57–76.

[Ap79] R. Apéry, *Irrationalité de ζ_2 et ζ_3.* Luminy Conference on Arithmetic. Astérisque No. 61 (1979), 11–13.

[BMRS19] L. Boza, J.M. Marín, M.P. Revuelta, and M.I. Sanz. 3-*color Schur numbers.* Discrete Appl. Math. 263 (2019), 59–68.

[Bo65] B. Bollobás, *On generalized graphs.* Acta Mathematica Academiae Scientarum Hungaricae 16 (1965), 447–452.

[BR01] K. Ball and T. Rivoal, *Irrationalité d'une infinité de valeurs de la fonction zêta aux entiers impairs.* Invent. Math. 146 (2001), 193–207.

[CFIM16] M. Codish, M. Frank, A. Itzhakov and A. Miller, *Computing the Ramsey number $R(4,3,3)$ using abstraction and symmetry breaking.* Constraints 21 (2016), 375–393.

[Cl15] P.L. Clark, *A note on Euclidean order types.* Order 32 (2015), 157–178.

[Cl-CA] P.L. Clark, *Commutative Algebra.* `http://alpha.math.uga.edu/~pete/integral2015.pdf`

[Cl-DC] P.L. Clark, *Discrete calculus.* In preparation. Draft available on request.

[Cl-FT] P.L. Clark, *Field Theory.* `http://alpha.math.uga.edu/~pete/FieldTheory.pdf`

[Cl-GT] P.L. Clark, *General Topology.* `\http://alpha.math.uga.edu/~pete/pointset.pdf`

[Cl-HC] Pete L. Clark, *Honors Calculus.* `http://alpha.math.uga.edu/~pete/2400full.pdf`

[Cl-NT] Pete L. Clark *Number Theory: A Contemporary Introduction.* `http://alpha.math.uga.edu/~pete/4400FULL.pdf`

[Cl13] P.L. Clark, *Graph derangements.* Open Journal of Discrete Mathematics 3 (2013), 183–191.

[CPZ] G. Chartrand, A.D. Polimeni and P. Zhang, *Mathematical Proofs: A Transition to Advanced Mathematics.*

[DHM24] E. Dueñez, A.S. Hamakiotes and S.J. Miller, *Sums of powers by L'Hopital's rule.* `https://arxiv.org/pdf/2302.03624`

[Di50] R.P. Dilworth, Dilworth, *A decomposition theorem for partially ordered sets.* Ann. of Math. (2) 51 (1950), 161–166.

[D077] G. Dobiński, *Summirung der Riehe $\sum \frac{n^m}{n!}$ für $m = 1, 2, 3, 4, 5 \ldots$.* Grunert's Archiv (1877), 333-336.

[EPS81] R.J. Evans, J.R. Pulham and J. Sheehan, *On the number of complete subgraphs contained in certain graphs.* Journal of Combinatorial Theory. Series B. 30 (1981), 364–371.

[Ge34] A. Gelfond, *Sur le septiéme Probléme de Hilbert.* Bulletin de l'Académie des Sciences de l'URSS. Classe des sciences mathématiques et na. VII (4): 623–634.

[GG55] R.E. Greenwood and A.M. Gleason, *Combinatorial relations and chromatic graphs.* Canadian J. Math. 7 (1955), 1–7.

[GR82] C.M. Grinstead and S.M. Roberts, *On the Ramsey numbers $R(3,8)$ and $R(3,9)$.* J. Combin. Theory Ser. B 33 (1982), 27–51.

[GY68] J.E. Graver and J. Yackel, *Some graph theoretic results associated with Ramsey's theorem.* J. Combinatorial Theory 4 (1968), 125–175.

[H] G.H. Hardy, *A mathematician's apology.* With a foreword by C. P. Snow. Reprint of the 1967 edition. Canto. Cambridge University Press, Cambridge, 1992.

[HV50] P.R Halmos and H.E. Vaughan, *The marriage problem.* Amer. J. Math. 72 (1950), 214–215.

[HW09] M. Hardy and C. Woodgold, *Prime simplicity.* Math. Intelligencer 31 (2009), 44–52.

[Jo] D. Joyce, `https://mathcs.clarku.edu/~djoyce/elements/bookIX/propIX20.html`.

[K] I. Kaplansky, *Set theory and metric spaces.* Allyn and Bacon Series in Advanced Mathematics. Allyn and Bacon, Inc., Boston, Mass., 1972.

[Ke64] G. Kéry, *On a Theorem of Ramsey*, *Matematikai Lapok* 15 (1964), 204–224.

[Li33] F.A Lindemann, *The Unique Factorization of a Positive Integer*. Quart. J. Math. 4, 319–320, 1933.

[Lu66] D. Lubell, *A short proof of Sperner's lemma.* Journal of Combinatorial Theory 1 (1966), 299.

[Me63] L.D. Meshalkin, *Generalization of Sperner's theorem on the number of subsets of a finite set.* Theory of Probability and Its Applications 8 (1963), 203–204.

[Mi71] L. Mirsky, *A dual of Dilworth's decomposition theorem.* Amer. Math. Monthly 78 (1971), 876–877.

[MM92] B.D. McKay and Z.K. Min, *The value of the Ramsey number* $R(3,8)$. J. Graph Theory 16 (1992), 99–105.

[MR95] B.D. McKay and S.P. Radziszowski, $R(4,5) = 25$. J. Graph Theory 19 (1995), 309–322.

[Mo61] L.J. Mordell, *The congruence* $(\frac{p-1}{2})! \equiv \pm 1 \pmod{p}$. Amer. Math. Monthly 68 (1961), 145–146.

[MR14] K. Mamakani and F. Ruskey, *New roses: simple symmetric Venn diagrams with 11 and 13 curves.* Discrete Comput. Geom. 52 (2014), 71–87.

[MSE] `https://math.stackexchange.com/questions/75906/`

[Mu63] A.A. Mullin, *Recursive function theory (A modern look at a Euclidean idea)*. Bulletin of the American Mathematical Society 69 (1963), 737.

[Ro64] G.-C. Rota, *The number of partitions of a set.* Amer. Math. Monthly 71 (1964), 498—504.

[R] W. Rudin, *Principles of mathematical analysis.* Third edition. International Series in Pure and Applied Mathematics. McGraw-Hill Book Co., New York-Auckland-Düsseldorf, 1976.

[Ra30] F.P. Ramsey, *On a problem of formal logic*. Proceedings of the London Mathematical Society 30 (1930), 264–286.

[Ro63] K. Rogers, *Classroom Notes: Unique Factorization*. Amer. Math. Monthly 70 (1963), no. 5, 547–548.

[Sch16] I. Schur, Über die kongruenz $x^m + y^m = z^m \pmod{p}$, Jahresber. Deutsche Math.-Verein., 25 (1916), 114–116.

[Sp28] E. Sperner, *Ein Satz über Untermengen einer endlichen Menge.* Mathematische Zeitschrift 27 (1): 544–548.

[Tu53] W.T. Tutte, *The* 1*-factors of oriented graphs.* Proc. Amer. Math. Soc. 4 (1953), 922–931.

[Ya54] K. Yamamoto, *Logarithmic order of free distributive lattice.* Journal of the Mathematical Society of Japan 6 (1954), 343–353.

[Ze34] E. Zermelo, *Elementare Betrachtungen zur Theorie der Primzahlen*. Nachr. Gesellsch. Wissensch. Göttingen 1, 43–46, 1934.