MATHEMATICAL SKILLS I

REASONING AND COMMUNICATION

Notes for the Course

Brent Everitt, Michael Bate & Chris Wood

I have had my results for a long time, but I do not yet know how to arrive at them.

Carl Friedrich Gauss (1777–1855)

If only I had the theorems! Then I should find the proofs easily enough.

Bernhard Riemann (1826–1866)

Contents

List of Figures

List of Tables

Introduction

Mathematical Skills I is a module that runs throughout Year 1 of the Mathematics degree. The first part of the course, taught in the Autumn Term, is lecture based and covers four main topics, each corresponding to a chapter of these notes: “Sets and Functions”; “Logic and Proof”; “Equivalence Relations”; “Axiomatic Mathematics (mainly Group Theory)”. Each topic/chapter has its own problem sheet (available separately) and tutorial during the term. The exam in January will consist of four questions, one for each topic.

These notes, whilst assuming a basic A-level (or equivalent) maths background, are intended to be as self-contained and comprehensive as possible. However, there are inevitably going to be some omissions. Luckily there are some very good textbooks available, of which we have recommended three: [3, 17, 29]. In particular, the book by Franklin and Daoud [17] is available as a free download. At the start of each chapter there are references to relevant sections of the three recommended texts, in case alternative explanations or more extensive discussions are desired. (Relevant “episodes” of an accompanying video series made during Autumn 2020 are also highlighted; these are available online.) There is no need to stick to the recommended texts either: there are hundreds of books in the University library covering the material in this course, approaching it from different directions and taking it further as well.

One of the main themes of the course is “reasoning”, by which we mean the ability to handle the logically rigorous type of argument that is needed to prove mathematical theorems. The importance of this should be self-evident: theorems are the currency of mathematics; or as someone[1] once said: “A mathematician is a machine for turning coffee into theorems”! Apart from drinking copious amounts of coffee (or a beverage of choice), there are some basic principles and techniques that can be applied to help us in this task. These will be mentioned throughout the notes, and particularly in Chapter 2 on Logic and Proof. Some will perhaps already be familiar, others less so. However at the end of the day the only way to develop this vital mathematical skill is by seeing lots of examples and doing plenty of exercises. There’s a good supply of these throughout the notes, and on the problem sheets.

On many occasions, not only will a proof be fully written out but there will also be a brief accompanying commentary (usually right at the beginning), describing what’s going on and what type of argument is being used. (Warning: don’t expect to see this

[1] Alfréd Rényi (1921–1970): Hungarian mathematician and coffee addict.

to quite the same extent out “in the wild”; most mathematical expositors expect their readers to figure out such details for themselves!) Some results have several alternative proofs, each making use of a different principle or idea. These all “do the job” perfectly well, but some may be more “efficient”, or “elegant”, or provide more insight, than others. So ultimately there’s a lot more going on in a proof than “just” a logical verification of a statement of mathematical fact, and it’s perfectly legitimate to have personal preferences for proofs of a particular style or type. *(Idea: why not keep a “mathematical diary”, so that as your degree progresses you can record which theorems and proofs you particularly like!)*

At the end of each chapter there is what we call a “Set Piece”, that brings together the main topics from the chapter in order to showcase a significant, interesting and useful mathematical theorem. The proofs of these results typically introduce some clever new ideas, and serve as examples of longer, more involved or elaborate arguments, some of which are quite subtle. They also illustrate the level of “rigour” that we expect to encounter in a modern proof. Set Pieces 1 and 2 both deal with aspects of infinity: “Counting Infinity”, and “Different Sizes of Infinity”, respectively. These are topics that have fascinated mathematicians for centuries, if not millennia! They invoke a variety of mind-twisting ideas which, when pursued to their logical conclusion, reveal some fundamental mathematical paradoxes. When these first came to light, in the early twentieth century, they precipitated a “quiet revolution” in the foundations of mathematics. Set Piece 3 deals with another fundamental mathematical question: “What is a number?” More precisely, we show how to construct the “rational numbers” from the integers, followed by some detailed hints for constructing the integers from the “natural numbers”, which we then show how to define at the beginning of Chapter 4 from a collection of fundamental properties, or “axioms”. Finally, in Set Piece 4 we ponder the “symmetries of symmetry”. This sounds rather nebulous, but when turned into a precise mathematical concept (“group theory”) it constitutes one of the driving forces of modern mathematics, and is a first step into the vast area of “abstract algebra”.

The other main theme of this course is mathematical “communication”, which is the art of explaining mathematics to others, ranging from brand new pieces of research to the kind of problems and exercises that we find in textbooks. Whilst we certainly hope that these notes successfully communicate the mathematical ideas we’re attempting to describe, they’re also intended to provide a sort of blueprint and guide for how to develop this important mathematical skill. In the first part of the course (Autumn Term) we will be concentrating on written mathematics, which is really the mathematician’s primary means of communication.

Writing mathematics in a precise yet comprehensible and attractive way is always a challenge, and something that to many of us doesn’t come naturally, at least not without a fair amount of practice. The exercises on the problem sheets are designed to help with this, and we advise taking the opportunity, particularly when writing up assignments, to think carefully not only about how to solve the questions mathematically but also the way in which the solutions are expressed and presented. Having written up a piece of mathematics it's always a good idea to leave it for a day (or night) to "rest"; then read through it again, checking that it's mathematically correct, and clearly expressed. The point here is that what may seem like a mathematical masterpiece in the heat of the moment can all too often lose some of its shine in the cold light of day; so a second look is almost always beneficial.

There are certain rules and conventions of mathematical writing that it's helpful to be aware of (and occasionally break), such as:

> *Never start a sentence with a mathematical symbol!*

We'll touch on these from time to time, but won't go into all the details. For anyone wanting a closer look, there are comprehensive accounts of all the "dos and don'ts" in books such as [23, 38]. The reference [27] is also well regarded, although it's aimed primarily at professional mathematicians.

Finally, it's worth mentioning that most of the mathematics we'll encounter in this course is "foundational", and in all likelihood will pop up from time to time in many other modules throughout the degree. So, even when this module is done and dusted, we hope that these notes will still come in handy, to refer to as and when needed.

Conventions. Throughout these notes the four symbols ■ ♦ □ ◊ are used as "bookends" to indicate the following:

■ · · · end of Proof (of Theorem, Proposition, Corollary or Lemma);
♦ · · · end of Definition;
□ · · · end of Example;
◊ · · · end of Note or Remark.

Defined terms are highlighted using this *sepia slant* font.

We confer the title "theorem"[2] on a mathematical statement that is provably true, and which carries some weight; the mathematical version of a "show stopper". Most (but not all) of our theorems appear in the Set Pieces at the end of each chapter. A "proposition"[3] is similar to a theorem, carrying the same logical status, but not quite the same "wow factor". Propositions tend to be steps on a mathematical journey rather than destinations, and their proofs are generally (but not always) more straightforward. A "lemma"[4] is a comparitively small-scale result whose main purpose is to assist in proving a something bigger, either a proposition or a theorem. Lemmas typically appear when it's beneficial to break up a long proof into smaller pieces. So some theorems may be preceded by one or

[2] From the Greek $\theta\varepsilon\omega\rho\eta\mu\alpha$, meaning "a spectacle"; also the root of the English word "theatre".

[3] From the Latin *proponere,* meaning "to set forth".

[4] From the Greek $\lambda\eta\mu\mu\alpha$, meaning "something received".

more lemmas, which end up doing lot of the mathematical “heavy lifting”, allowing the proof of the main result to be more concise and tightly focussed. Finally, a “corollary”[5] is a result that follows relatively effortlessly from a theorem (or even a proposition), requiring no significant new input; it’s as close as we get to a “freebie” in mathematics.

[5] From the Latin *corollarium,* meaning “garland”.

Preliminaries and Notation

There are various bits and pieces of notation that we'll use often, which are common and fairly universal throughout mathematics. Here's a collection of a few of them. (There is a comprehensive Index of Notation right at the end of the Notes.)

- $\mathbb{N}$ denotes the set of *natural numbers*. These are the counting numbers $1, 2, 3, 4,$ and so on. We do not include 0 as a natural number in this course (although many people do, so watch out). These numbers go on forever (there are infinitely many of them), but there is no such number as ∞.

- $\mathbb{Z}$ (for the German noun "Zahlen") denotes the set of *integers*. These are the positive and negative whole numbers, together with 0. They can be constructed from the natural numbers $\mathbb{N}$ (we'll show how to do this in Section 3.7.2 of Set Piece 3), and contain $\mathbb{N}$ as the subset of positive integers.

- $\mathbb{Q}$ (for "quotient") denotes the set of *rational numbers*. We'll construct these properly later in the course (Set Piece 3: Construction of the rational numbers), but for now we can just think of them as the numbers that can be written as fractions a/b where a, b are integers and $b \neq 0$. We identify $\mathbb{Z}$ as a subset of $\mathbb{Q}$ in the usual way; ie. the fractions with a 1 on the bottom (denominator), these usually being written without the 1.

- $\mathbb{R}$ denotes the set of *real numbers.* Not surprisingly, these feature heavily in the mathematical discipline of "real analysis", one of whose tasks is to rigorously construct them from the rational numbers. This turns out to be rather complicated, so for the purposes of this course we'll treat them slightly less formally, as the numbers that can be expressed by a decimal expansion. We can then identify $\mathbb{Q}$ as a subset of $\mathbb{R}$ by expressing a fraction in decimal form. However, there is still no such real number as ∞, or $-\infty$. It's very useful to visualise $\mathbb{R}$ geometrically, as an infinite straight line: the "real line".

- $\mathbb{C}$ denotes the set of *complex numbers.* These will be fully explained in the Algebra module, but won't feature in this course. Suffice it to say that the complex numbers $\mathbb{C}$ can be constructed (rather easily, in fact) from the real numbers $\mathbb{R}$, which then form a natural subset of $\mathbb{C}$. There's also a very useful geometric visualisation of $\mathbb{C}$, as an infinite plane: the "complex plane" or "Argand plane".

- $\mathbb{H}$ (for "Hamilton"[6]) denotes the *quaternions,* and $\mathbb{O}$ denotes the *octonions,* or *Cayley*[7] *numbers.* We won't discuss these rather exotic number systems at all (although they might make an interesting group project); in fact, we will be re-purposing the symbol $\mathbb{O}$ to denote the odd natural numbers instead.

The rather curious typeface ("font") used to denote the above sets is called "blackboard bold". Once upon a time these sets were denoted by upper case bold typeface characters, and examples of this still appear in some textbooks. However, for blackboard presentation (during a lecture, for example) it was found much more convenient and effective (and still is) to use double strokes, and the popularity of this serendipitous practice eventually resulted in its own typeface. It's an interesting example of how mathematical subculture occasionally finds its way into the mainstream.

- Symbols:

  * $\in$ stands for "(is) in". Synonyms are: "(is) a member of", "(is) contained in", "belongs to". This symbol belongs to set theory.

  * $\exists$ stands for "there exists", or "there is". This symbol is a "quanitifier", the *existential quantifier,* and belongs to formal logic.

  * $\forall$ stands for "for all", or "for every". This is the second widely used logical quantifier, the *universal quantifier.*

  * ∞ stands for "infinity". This is arguably one of the mathematician's favourite symbols, with a popular folklore that extends well beyond mathematics. Its use requires considerable care, and varies depending on context. We remarked earlier that ∞ is not a real number. However, we will give three examples of how the symbol can be used legitimately (Example 1.2, Section 3.7.1 and Definition 4.5), each of which is different and comes with its own definition. (Precise definitions are always important in mathematics, but particularly so in this case, where the folklore surrounding ∞ gives rise to all sorts of preconceptions, and misconceptions.) The biggest "client" of the ∞ symbol is mathematical analysis, the area of mathematics that deals with limiting processes, such as differentiation and integration. We will not have much to say about this, which is best left to more specialised courses such as Calculus and Real Analysis.

  * $\sum$ (capital Greek letter "sigma") usually indicates *summation.* Sums should usually be taken over some sort of *index set* which labels the things to be summed. Often this index set is a subset of the natural numbers, but it doesn't have to be. If the index set is infinite (like $\mathbb{N}$) then the precise meaning of the sum has to be carefully considered. (This is precisely the kind of question that arises in mathematical

[6] William Rowan Hamilton (1805–1865): Irish mathematician with interests in both pure mathematics and mathematical physics.

[7] Arthur Cayley (1821–1895): British mathematician, whose primary interest was abstract algebra.

    analysis, and figures prominently in Real Analysis.) The summation symbol will appear quite often in these notes, and is a very useful labour saving device for performing complex calculations. A particularly good example of this can be found in Appendix A.4 (although this isn't an essential part of the course).

    - $\prod$ (capital Greek "pi") often indicates a *product.* As with sums, products should usually be taken over some sort of indexing set, and subtleties again arise when this set is infinite. Again, its primary purpose is to simplify complex calculations, although it's not as common as the summation symbol, and in fact won't appear at all in the main part of these notes. However, for anyone with a good head for heights there are nice examples of its use in Appendices A.5 and A.6 (which again aren't essential parts of the course).

  There are lots (literally hundreds) more specialist symbols that have been adopted by mathematicians. We'll be introducing quite a few of them as the course progresses.

- Greek letters. We've just mentioned two in the previous list of symbols, and there are others that will probably already be familiar, such as $\alpha, \beta, \gamma, \delta, \epsilon$, etc. *(Recognisable handwritten versions of some of these can be quite tricky to achieve; for example, try your hand at ζ and ξ. If you don't recognise a letter that someone (a lecturer or seminar leader) is using, please ask!)* Why Greek? Pragmatically, because there aren't enough characters in the ordinary (Roman) alphabet to provide the variety of different letters that mathematicians require[8]. And in fulfilling this rôle many have become rather popular; for example, can anyone imagine denoting the real number $3.151592\cdots$ by anything other than π? Romantically, because ancient Greece was the cradle of mathematics, and Greek mathematicians pioneered the use of deductive reasoning to prove "incontrovertible truths" (ie. mathematical theorems), the same approach that we use today! Indeed, as noted earlier, the word "theorem" comes from the Greek word meaning "a spectacle", and in fact "mathematics" itself comes from the Greek root $\mu\alpha\theta\eta\mu\alpha$ which means "learning".

  For reference here's the mathematician's (lower case) Greek alphabet, from "alpha" to "omega":

  $$\alpha,\ \beta,\ \gamma,\ \delta,\ \varepsilon,\ \zeta,\ \eta,\ \theta,\ \iota,\ \kappa,\ \lambda,\ \mu,\ \nu,\ \xi,\ \pi,\ \rho,\ \sigma,\ \tau,\ \upsilon,\ \varphi,\ \chi,\ \psi,\ \omega$$

  (Can you name each letter?) In fact, the full Greek alphabet has 24 letters (not including variants, such as φ and ϕ), so there is one letter missing. *(Do you know which one, and why?)*

  The upper case Greek alphabet contains many letters that were later adopted into the Roman alphabet, and are therefore familiar to us; but we'll list them all to help see

[8] Not only mathematicians: in 2020 the US National Hurricane Center began using Greek letters to name the increasing number of hurricanes and tropical storms; and in 2021 they were used to label Covid variants.

how the more exotic symbols correspond to their lower case counterparts, and in turn how some of those correspond to certain Latin letters:

$$A,\ B,\ \Gamma,\ \Delta,\ E,\ Z,\ H,\ \Theta,\ I,\ K,\ \Lambda,\ M,\ N,\ \Xi,\ \Pi,\ P,\ \Sigma,\ T,\ Y,\ \Phi,\ X,\ \Psi,\ \Omega$$

It may be surprising to see Z appear in sixth place, and Y before X!

Despite the use of two alphabets (Greek and Roman), and the occasional foray into others (for example, Hebrew), plus a truck load of specialist symbols, mathematicians frequently run out of notational rope! It's quite common to find the same symbol "recycled", or used to denote different things depending on the area of mathematics. For example, in the theory of partial differential equations Δ invariably means "Laplace's operator", whereas in numerical analysis it's the "difference operator", and in set theory the symbol for "symmetric difference". Even in this (short) course we'll come across a couple of examples! Provided the usage is made clear, all should be well.

Good notation should do more than simply provide a collection of symbols as placeholders for specific mathematical objects, operations, relations or properties; it should also provide a clue about what's being abbreviated (if only because, like everyone else, mathematicians have finite memories, so any help in "recovering" the underlying ideas from the notation is always welcome). We can see this with the notation for the various sets such as $\mathbb{N}, \mathbb{Z}, \mathbb{Q}, \ldots$, and there will be other examples as the course moves forward.

An area of mathematics whose notation is strikingly good is calculus, where we can often get the right results by formal manipulation of the symbols. Of course, this can be a double-edged sword, if it leads to a superficial mindset of "mathematics by symbol shuffling", rather than a proper understanding of what's really going on. The symbols or notation that we're using should only ever be seen as a means to an end: a concise and effective way to represent and discuss the vast collection of ideas that mathematicians have come up with during the course of mathematical history, and continue to develop in the present day.

1. Sets and Functions

References.
Liebeck: Chapter 1 pp. 1–2, Chapter 19
Allenby: Chapter 4
Franklin and Daoud: Chapter 7, Chapter 12.
Episodes 1–4 of video lectures.

A fundamental skill for any mathematician to acquire is the ability to use coherently and correctly the basic ideas of set theory. We won't attempt to give a rigorous definition of what we mean by a "set", but rather concentrate on some basic properties that are useful to the working mathematician; so we're going to describe what is sometimes called "naive set theory" [21]. (However, look out for "Russell's paradox" at the end of Set Piece 2 for an idea of why this topic really needs a more thorough, and ultimately axiomatic treatment.) Once we've got a bit of set theory under our belts, we're going to talk about how we can compare different sets, using functions.

1.1. Sets

The theory of sets is usually understood to have originated with the now classic work of Cantor[1] on different "sizes" of infinity [7]. We will take a look at this in Set Piece 1, after we have laid out the basic ideas of sets and functions. (There will be more to follow in Set Piece 2.)

For the purposes of this course (and for most of the degree) it's safe to think of a set rather informally as a "collection" of "objects", called the *elements, members,* or *points* of the set. In a maths context, these will usually be things like numbers, or functions, or even other sets. We've already met some sets: $\mathbb{N}$, $\mathbb{Z}$, $\mathbb{Q}$ and $\mathbb{R}$ are all sets. Most of the other sets we'll see in this course will be built out of these in some way.

Sets are often denoted with curly brackets "{" and "}", particularly at the point of first definition; thereafter they're usually denoted by capital letters $A, B, C, \ldots$, although as we saw in our preliminary section on Notation, special sets get special symbols! The

[1] Georg Cantor (1845–1918): German mathematician, pioneer of set theory and transfinite arithmetic, whose work provoked incomprehension, hostility and ridicule amongst several of his leading mathematical contemporaries.

members of a set are sometimes presented as a (finite or infinite) list:

$$\{a, b, c\} \quad \text{(finite)}, \qquad \{a, b, c, \dots\} \quad \text{(infinite)},$$

or, more commonly, by some defining property:

$$\{x \mid P(x)\},$$

also written with a colon:

$$\{x : P(x)\},$$

where $P(x)$ is a statement about x (more about "statements" in Chapter 2, Section 2.1). For example:

$$\begin{aligned} \mathbb{R}^+ &= \{x \mid x \text{ is a positive real number}\}, \\ \mathbb{R}_0^+ &= \{x \mid x \text{ is a non-negative real number}\}, \\ \mathbb{R}^* &= \{x : x \text{ is a non-zero real number}\}. \end{aligned}$$

When interpreting set definitions like these, we mentally substitute the "|" or ":" symbol by the phrase "such that"; this should produce a recognisable sentence, and everything should then make sense!

Remark 1.1. There are limits to the statements that can be used to define sets; see Theorem 2.15 and "Russell's paradox" at the end of Set Piece 2 ♢

There is a very important set called the *empty set*. It has no elements, and is usually denoted by the symbol ∅. Using brackets, we'd write:

$$\emptyset = \{\ \},$$

but this looks absurd, so we don't! Despite its intrinsic lack of interest, the empty set plays an essential rôle in set theory, rather like the number 0 in arithmetic, or the identity element in group theory (which we will be studying in Chapter 4). Slightly more interesting is a set $\{a\}$ containing just one element, which is called a *singleton.*

1.1.1. Membership and equality.

The fundamental relations in set theory are *membership* and *equality.* We write:

- $x \in A$ if x is an element of A;
- $A = B$ if the sets A and B contain precisely the same elements[2].

The definition of set equality is sometimes called the *Principle for Equality of Sets*[3]. We've already used it in our definitions of the sets $\mathbb{R}^+$, $\mathbb{R}^*$ etc. It's so natural that it's difficult not to!

[2] Perhaps surprisingly, the "=" sign was only introduced into mathematics in the 16-th century, by Welsh mathematician Robert Recorde (1512–1558). Originally intended for use in algebraic equations, it is now used much more widely.

[3] Also known as the Principle of Extension.

Both of these relations may be negated; so we write:

- $x \notin A$ if x is *not* an element of A;
- $A \neq B$ if *at least one* element of A is *not* a member of B, or vice versa.

Example 1.1 (Sets of prime factors).
Let's define sets A and B as follows:

$$A = \{p \mid p \text{ is a prime factor of } 6\} = \{2, 3\},$$
$$B = \{q \mid q \text{ is a prime factor of } 10\} = \{2, 5\}.$$

Note that 1 is not considered to be a prime number (see Remark 2.9), so doesn't appear in either set:

$$1 \notin A, \qquad 1 \notin B.$$

Notice also that we've chosen to use p and q rather than x as the "dummy variable" when defining A and B. This is simply because it is more common to use these symbols (particularly p) to denote prime numbers. However, this doesn't affect the validity of the definitions; we are free to use whatever symbol(s) we like! Now, since A has at least one element (namely 3) that doesn't belong to B, the sets are not equal:

$$A \neq B.$$

Let's now introduce a third set:

$$C = \{p \mid p \text{ is a prime factor of } 20\} = \{2, 2, 5\} = \{2, 5\}.$$

An important point to note here, which can sometimes cause confusion, is that sets ignore repeated elements! This is an inevitable consequence of our definitions. Perhaps the most compelling way of seeing this is to suppose for argument's sake that $\{2, 2, 5\} \neq \{2, 5\}$. Then applying the above definition would mean that one of these sets has an element which doesn't belong to the other, and this is not the case. So in fact $B = C$, even though the defining conditions for these sets are different.

In view of this, it is correct to say:

> *10 and 20 have the same set of prime factors.*

We may even abbreviate this to:

> *10 and 20 have the same prime factors.*

However, we should take care not to say that 10 and 20 have the same *prime factorisation,* which takes into consideration the *multiplicity* of each prime, and is unique to each natural number; this is the Fundamental Theorem of Arithmetic (see Theorem A.4). □

The relations of membership and equality are the starting point for constructing new sets from old ones. There are many ways to do this, and we now describe those that appear most often.

1.1.2. Subsets.

Otherwise known as *inclusion,* this is the third fundamental relation of set theory. We say that a set A is a *subset* of another set B, and write $A \subseteq B$, if every element of A is also an element of B:

$$\text{if } x \in A \text{ then } x \in B.$$

Every set is a subset of itself: $A \subseteq A$, as indicated by the "subset or equals" notation. (Warning: many mathematicians use the "$\subset$" symbol in the same way, so it's also acceptable to write $A \subset A$.)

The subset relation may be negated:

$$A \nsubseteq B,$$

which means there exists an element of A that doesn't belong to B. This is not to be confused with the pair of relations $A \subseteq B$ and $A \neq B$, which when taken together mean something quite different: every element of A belongs to B but there is some element of B that doesn't belong to A. In this case A is called a *proper subset* of B; we will use the notation $A \subset B$ only when this is the case. (Mathematicians who use "$\subset$" in all instances are obliged to use the rather unsightly notation $A \subsetneq B$, risking a mix up with the negation!) There are times when it's convenient to use the *superset* symbols: $A \supseteq B$ and $A \supset B$. These mean exactly the same as $B \subseteq A$ and $A \subset B$, respectively.

The empty set $\emptyset$ is a subset of *every* set. The nicest way to see this is to assume the contrary, that $\emptyset$ is *not* a subset of some set: $\emptyset \nsubseteq A$ for some set A. This would mean that $\emptyset$ has an element which doesn't belong to A, contradicting the fact that $\emptyset$ has no elements! The only way this logical *impasse* can be resolved is if $\emptyset \subseteq A$ after all. (This method of proof, which is called "proof by contradiction", is slightly devious but very effective and hugely popular; we will have a lot more to say about it in Chapter 2 (Section 2.4.4), and throughout the notes thereafter.)

It is very common to define a "new" set as a subset of some "old" one. This leads to a third notation for sets:

$$\{x \in X \mid P(x)\},$$

where X is some predefined set. For example:

$$\begin{aligned} \mathbb{R}^+ &= \{x \in \mathbb{R} \mid x > 0\}, && \cdots \text{ the positive reals,} \\ \mathbb{R}^+_0 &= \{x \in \mathbb{R} \mid x \geqslant 0\}, && \cdots \text{ the non-negative reals,} \\ \mathbb{R}^* &= \{x \in \mathbb{R} : x \neq 0\}, && \cdots \text{ the non-zero reals.} \end{aligned}$$

There is also a fourth notation, where belonging to the predefined set X appears as part of the defining property:

$$\{E(x) \mid x \in X\},$$

where $E(x)$ is an expression in x. For example:

$$\mathbb{E} = \{2n \mid n \in \mathbb{N}\}$$

is the set of all even natural numbers, defined as a subset of $\mathbb{N}$.

It's really important to understand the different notational conventions for sets, and try to use whichever is best suited to the case at hand. For example, instead of the above definition of $\mathbb{E}$ we could have given the following:

$$\mathbb{E} = \{n \mid n \in \mathbb{N} \text{ and } n = 2m \text{ for some } m \in \mathbb{N}\},$$

which although correct is decidedly clunkier. There are exercises on the problem sheets to help with this.

Example 1.2 (Intervals)**.**
A useful class of subsets of $\mathbb{R}$ are the *intervals,* which in contrast to the subsets $\mathbb{N}$, $\mathbb{Z}$ and $\mathbb{Q}$ are "continuous pieces" of the "real line" $\mathbb{R}$. They come in two basic flavours, *open* and *closed,* and a hybrid *half open* (or *half closed*) combination. They can also be *finite* or *infinite.* Here are the precise definitions.

Let a, b be real numbers with $a < b$. Then:

$$\begin{aligned}
(a,b) &= \{x \in \mathbb{R} \mid a < x < b\}, && \cdots \text{ open, finite,}\\
[a,b] &= \{x \in \mathbb{R} \mid a \leqslant x \leqslant b\}, && \cdots \text{ closed, finite,}\\
[a,b) &= \{x \in \mathbb{R} \mid a \leqslant x < b\}, && \cdots \text{ half open/closed, finite,}\\
(a,b] &= \{x \in \mathbb{R} \mid a < x \leqslant b\}, && \cdots \text{ half open/closed, finite,}\\
(a,\infty) &= \{x \in \mathbb{R} : a < x\}, && \cdots \text{ open, infinite,}\\
(-\infty,a) &= \{x \in \mathbb{R} : x < a\}, && \cdots \text{ open, infinite,}\\
[a,\infty) &= \{x \in \mathbb{R} : a \leqslant x\}, && \cdots \text{ closed, infinite,}\\
(-\infty,a] &= \{x \in \mathbb{R} : a \leqslant x\}, && \cdots \text{ closed, infinite,}\\
(-\infty,\infty) &= \mathbb{R}, && \cdots \text{ open and closed, infinite.}
\end{aligned}$$

Remarks *1.2.*

1) The descriptions "finite" or "infinite" refer to the length of the interval, not the number of points it contains; all intervals contain infinitely many points!
2) Use of the symbols ±∞ here is purely formal. As stated previously, these are *not* real numbers; ie. not elements of the set $\mathbb{R}$. So we don't allow intervals of the form $(a,\infty]$, for example. Nevertheless intervals such as $[a,\infty)$ and $(-\infty,\infty)$ are referred to as "closed". (There is a reason for this; see the end of Section 1.1.7.) ◊

Intervals are of particular importance in calculus and mathematical analysis. □

There are many situations in mathematics where we need to show that one set is a subset of another. There is a "standard routine" for this, which amounts to nothing other than verifying the definition: to show $A \subseteq B$ we verify that *every* element of A is also an element of B.

Example 1.3 (Standard routine for subsets)**.**
Let $A = \{n^2 \mid n \in \mathbb{E}\}$. We claim that $A \subseteq \mathbb{E}$. To prove this, suppose $a \in A$. Then $a = n^2$ for some $n \in \mathbb{E}$. Since $n = 2m$ for some $m \in \mathbb{N}$ we have:

$$n^2 = (2m)^2 = 4m^2 = 2(2m^2) = 2k,$$

for some $k \in \mathbb{N}$. Therefore $a \in \mathbb{E}$. Having thus shown that every element of A is an element of $\mathbb{E}$ we can write $A \subseteq \mathbb{E}$. Since there are even numbers that aren't squares (for example, 2), A is a proper subset. We can indicate this, if we wish, by writing $A \subset \mathbb{E}$. □

1.1.3. Intersections.

The *intersection* $A \cap B$ of two sets A and B consists of all the elements that A and B have in common, thus:

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

For example:

$$\{0, 1, 2\} \cap \{1, 2, 3\} = \{1, 2\},$$
$$\mathbb{R}^* \cap \mathbb{R}_0^+ = \mathbb{R}^+,$$
$$(0, 2] \cap [1, 3) = [1, 2].$$

It follows directly from the definition that:

$$A \cap B \subseteq A \quad \text{and} \quad A \cap B \subseteq B.$$

In fact $A \cap B$ is the "largest" subset of A and B, by which we mean that if $C \subseteq A$ and $C \subseteq B$ then $C \subseteq A \cap B$. *(Can you prove this, using the "standard routine" for subsets?)*

If $A \cap B = \emptyset$ we say that A and B are *disjoint.* For example, $\{0, 1\}$ and $\{2, 3\}$ are disjoint, as are the intervals $[0, 1)$ and $[1, 2)$.

1.1.4. Unions.

The *union* $A \cup B$ of two sets A and B consists of all the elements of A together with all the elements of B:

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

For example:

$$\{0, 1, 2\} \cup \{1, 2, 3\} = \{0, 1, 2, 3\},$$

$$\mathbb{R}^+ \cup \{0\} = \mathbb{R}_0^+,$$
$$(0,2] \cup [1,3) = (0,3).$$

It follows immediately from the definition that:

$$A \subseteq A \cup B \quad \text{and} \quad B \subseteq A \cup B.$$

In fact $A \cup B$ is the "smallest" set containing both A and B as subsets, by which we mean that if $A \subseteq C$ and $B \subseteq C$ then $A \cup B \subseteq C$. *(Again, this is a property that you might like to prove, using the "standard routine".)*

If A and B are disjoint we say that $A \cup B$ is a *disjoint union.* For example, $\{0,1,2,3\}$ is the disjoint union of $\{0,1\}$ and $\{2,3\}$, and also the disjoint union of $\{0,2\}$ and $\{1,3\}$. The interval $[0,2)$ is the disjoint union of $[0,1)$ and $[1,2)$, and also of $[0,1]$ and $(1,2)$. As these examples show, there are many ways of expressing a set as a disjoint union.

Remark 1.3. Use of the logical connectives "and" and "or" in the definitions of intersection and union hints at deep connections between set theory and formal logic. This was a view propounded by Bertrand Russell[4] in his 1903 treatise: *"The Priniciples of Mathematics"* [32], and pursued relentlessly in the epic follow-up: *"Principia Mathematica"* by Whitehead[5] and Russell [40] (*not* a recommended textbook; it takes several hundred pages just to prove that $1+1=2$). ♢

Remark 1.4 (Infinite unions and intersections).
We can take intersections and unions of infinitely many sets; as long as we're careful, this shouldn't cause any problems. For example, if we have sets $A_1, A_2, \ldots$ then we define their intersection and union:

$$I = A_1 \cap A_2 \cap \cdots = \bigcap_{n\in\mathbb{N}} A_n, \qquad U = A_1 \cup A_2 \cup \cdots = \bigcup_{n\in\mathbb{N}} A_n,$$

by the condition that $a \in I$ if and only if $a \in A_n$ for *all* $n \in \mathbb{N}$, and $a \in U$ if and only if $a \in A_n$ for *some* $n \in \mathbb{N}$. The same principle applies to intersections and unions of arbitrary families of sets. We'll be using such a union in Chapter 3 when we discuss *partitions* of sets (Definition 3.8). ♢

1.1.5. Set difference.

For any sets A and B, the *set difference,* or *relative complement* $A \smallsetminus B$ consists of all the elements of A that are not in B:

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

[4] Bertrand Russell (1872–1970): Welsh mathematician, logician and philosopher, multiple prize winner, recipient of Order of Merit, and Nobel laureate in literature!

[5] Alfred North Whitehead (1861–1947): English mathematician, logician and philosopher, and mentor to Bertrand Russell.

For example:

$$\{0,1,2\} \smallsetminus \{1,2,3\} = \{0\},$$
$$\mathbb{Z} \smallsetminus \mathbb{N} = \{0,-1,-2,-3,\dots\},$$
$$(0,2] \smallsetminus [1,3) = (0,1).$$

Notice that $A \smallsetminus B$ is *always* a subset of A. Furthermore it is not assumed that B is a subset of A; in the extreme case where A and B are disjoint we have $A \smallsetminus B = \emptyset$, and $\emptyset \subseteq A$ as we noted in Section 1.1.2.

1.1.6. Power sets.

The *power set* of a set A, denoted $\mathscr{P}(A)$, is the set of all subsets of A:

$$\mathscr{P}(A) = \{B \mid B \subseteq A\}.$$

The possibility of having "sets of sets" can be a little confusing; indeed a particularly devious attempt to construct such a set forms the basis for the mind-expanding "Russell's paradox" (described at the end of Set Piece 2). However, here's a much simpler example:

$$\mathscr{P}\{1,2,3\} = \{\{1,2,3\},\{1,2\},\{1,3\},\{2,3\},\{1\},\{2\},\{3\},\emptyset\},$$

and there are questions on the problem sheets that should help this to sink in. Notice that we have "improved" the notation, writing $\mathscr{P}\{1,2,3\}$ instead of $\mathscr{P}(\{1,2,3\})$. (Omitting a pair of brackets may cause grief to sticklers for notation, but many mathematicians observe the unwritten and informal *Principle of Minimal Bracketing* whenever safe and sensible to do so!) Notice also that, apart from $\mathscr{P}(\emptyset) = \{\emptyset\}$, the power set $\mathscr{P}(A)$ always contains at least *two* elements: A and $\emptyset$.

Remark 1.5. In our example, the set A had 3 elements and the number of elements in $\mathscr{P}(A)$ was $8 = 2^3$. By looking at a few more examples it is not hard to "guess" a simple formula for the number of elements of $\mathscr{P}(A)$ if A has N elements. *(We will leave you to think about this, and return to it in Section 2.4.5.)* The challenge is then to find a proof that validates the formula in all cases. (This approach to discovering/creating mathematics—looking at some examples, "guessing" a possible generalisation, and finally putting together a proof—is a common *modus operandi* for mathematicians.) There are in fact many possible proofs. (Again, this is not unusual, and mathematicians are often on the lookout for new "improved" proofs of old results, particularly where existing proofs are complicated or "messy".) We will give one of these in Proposition 2.8.

Many important sets have infinitely many elements, and in this case so too will the power set. Saying something meaningful in this case is more challenging, and will form the centrepiece of Set Piece 2 (Set Piece Theorem 2). ◊

Example 1.4 (Power sets and probability).
Power sets occur naturally in probability theory. To take a very simple example, if we roll a standard (6-sided) dice then the set of all possible outcomes is:

$$A = \{1, 2, 3, 4, 5, 6\},$$

which is referred to as the *sample space* for this "experiment". However we may want to consider certain types of result; for example, the possible outcomes for "roll an even number" constitute the subset $\{2, 4, 6\}$, which is referred to as an *event.* The power set $\mathscr{P}(A)$ may therefore be interpreted as the set of all possible dice-rolling events (including the "no roll" event, which corresponds to $\emptyset$). Of course, the next step is to assign a "probability" to each event, which requires the notion of a "function" (see Section 1.2). □

1.1.7. Complements.

Informally, the *complement* A^c of a set A is the set of all things not in A. Really this only makes sense if we've previously agreed to work inside some bigger set, which may be regarded as a "universe" for our theory (not to be confused with the notion of a "universal set", more about which later; see Theorem 2.15). In other words, if $A \subseteq X$ then:

$$A^c = \{x \in X \mid x \notin A\}.$$

Notice that this is nothing other than $X \smallsetminus A$; it's just a more convenient notation if we have no intention of ever leaving X. For example, if we're working in the set $\mathbb{R}$ of all real numbers then $\mathbb{Q}^c$ is the set of *irrational numbers;* ie. those real numbers that can't be written as a fraction, and therefore do not belong to the subset $\mathbb{Q}$ of rational numbers. (We will see in Set Piece 2 (Corollary 2.14) that "most" real numbers are in fact irrational!) Also inside $\mathbb{R}$, the complement of an open interval is either a closed interval, a disjoint union of two closed intervals, or the empty set; the complement of a closed interval is a similar configuration of open intervals.

1.1.8. Cartesian (or direct) products.

If A and B are sets, then $A \times B$ is the set of all *ordered pairs* (a, b) with $a \in A$ and $b \in B$:

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

For example:

$$\{0, 1, 2\} \times \{1, 2, 3\} = \{(0, 1), (0, 2), (0, 3), (1, 1), (1, 2), (1, 3), (2, 1), (2, 2), (2, 3)\}.$$

This construction originates from the (at the time, revolutionary) idea of Descartes[6] to use what are now known as "Cartesian coordinates" (x, y) to locate points in the plane;

[6] René Descartes (1595–1650): French mathematician and philosopher, famous for (amongst other things) his saying: "Cogito, ergo sum", meaning: "I think, therefore I am".

hence the terminology. Since $x, y \in \mathbb{R}$ the Cartesian plane itself can be presented as the Cartesian product $\mathbb{R} \times \mathbb{R}$, which is usually abbreviated to $\mathbb{R}^2$.

With the Cartesian plane in mind, it should be reasonably clear that $(a, b) \neq (b, a)$, unless $a = b$; ie. the order is important! Indeed, the relation of equality between ordered pairs is:

$$(a, b) = (c, d) \quad \text{if and only if} \quad a = c \ \text{ and } \ b = d. \tag{1.1}$$

Remark 1.6 (Definition of ordered pair).
Ordered pairs can be defined precisely, as follows:

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

Thus $(a, b) \in \mathscr{P}(A \cup B)$. Every element of $A \times B$ is therefore an element of $\mathscr{P}(A \cup B)$, which means that $A \times B \subset \mathscr{P}(A \cup B)$ by the definition of inclusion (Section 1.1.2). This may seem slightly strange! However, it allows us to deduce the rule (1.1) for equality of ordered pairs, which is really the only thing we need. For, suppose:

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

If $a \neq b$ then it follows from the Principle for Equality of Sets that $\{a\} = \{c\}$ and $\{a, b\} = \{c, d\}$, and then that $a = c$ and $b = d$. If $a = b$ then the equation collapses to:

$$\{\{a\}\} = \{\{c\}, \{c, d\}\},$$

and the “Principle” implies $d = c = a$; hence $a = c$ and $b = d$ once again.

The essential point here is that when ordered pairs are viewed “concretely” in this way, the rule (1.1) for determining when two ordered pairs are equal is simply a consequence of the more fundamental rule for determining when two sets are equal. ♢

We can extend the Cartesian product construction to three (or more) sets A, B, C by defining:

$$A \times B \times C = \{(a, b, c) : a \in A,\ b \in B,\ c \in C\},$$

where (a, b, c) is an *ordered triple.* (A precise definition can be given, along the lines of that of an ordered pair in Remark 1.6.) Again, if we take $A = B = C = \mathbb{R}$ then we end up with a set that is usually abbreviated to $\mathbb{R}^3$, whose elements are interpreted as the Cartesian coordinates (x, y, z) of points in space.

Remarks 1.7.

1) We can extend the definition of Cartesian product to infinitely many sets, just as we did for intersection and union (Remark 1.4). However, this can be a bit hairier, since it involves the “Axiom of Choice” (which hints at the deeper waters of axiomatic set theory).

2) There is one further important technique for constructing new sets from old: the quotient construction. This is a lot more technical than the others, and we will not meet until Chapter 3 (Definition 3.9). ♢

1.1.9. Principle of Mutual Containment

This is a fairly obvious but surprisingly useful way to show two sets are equal: if A and B are two sets and we can show $A \subseteq B$ *and* $B \subseteq A$, then we can conclude that $A = B$. This is really nothing more than a restatement of the Principle for Equality of Sets. Proofs often adopt the following template.

To show $A = B$:

- Let $x \in A$ and show that $x \in B$. Conclude that $A \subseteq B$.
- Conversely, let $y \in B$ and show that $y \in A$. Conclude that $B \subseteq A$.
- Put the two steps together to conclude that $A = B$.

If we're lucky we may be able to get the second step from the first step by simply reversing the argument; but this isn't always possible.

Example 1.5 (Principle of mutual containment)**.**
Let $A = \{n - m : m, n \in \mathbb{N}\}$. We prove that $A = \mathbb{Z}$.

($\subseteq$) The difference of two integers is always an integer, so $A \subseteq \mathbb{Z}$.

($\supseteq$) Suppose $a \in \mathbb{Z}$. There are three cases to consider.

- If $a > 0$ then $a = n - m$ where $n = a + 1$ and $m = 1$ (for example); so $a \in A$.
- If $a = 0$ then $a = 1 - 1$ (for example); so $a \in A$.
- If $a < 0$ then $a = n - m$ where $n = 1$ and $m = 1 - a$ (for example); so $a \in A$.

Hence $a \in A$. Therefore $\mathbb{Z} \subseteq A$.

It follows from the Principle of Mutual Containment that $A = Z$. □

We recommend trying out this methodology to prove some (or all!) of the basic laws of set theory (see Examples 1.6 and 1.7 below). These are given in the next section.

1.1.10. Basic laws of set theory

For any sets A, B, C we have the following properties of intersections, unions and complements. (There are also laws for Cartesian products, which we won't list.)

- The *commutative law of intersection:*

$$A \cap B = B \cap A.$$

- The *commutative law of union:*

$$A \cup B = B \cup A.$$

- The *associative law of intersection:*

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

- The *associative law of union:*

$$A \cup (B \cup C) = (A \cup B) \cup C.$$

- The *first distributive law:*

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

- The *second distributive law:*

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

- *De Morgan's*[7] *first law:*

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

- *De Morgan's second law:*

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

- The *double complement law:*

$$(A^c)^c = A.$$

Thankfully, it's not necessary to learn these rules off by heart, although in fact they're not difficult to remember; in particular, the commutative, associative and distributive laws are rather similar to those for addition and multiplication of integers. A good way to "reconstruct" the laws is to use *Venn diagrams*, such as Figure 1.1 for the first distributive law. However, these don't count as "proper proofs"! (There's a general mathematical principle: pictures are good for showing us what to prove, but don't prove it!) We will illustrate this by proving the first distributive law (Example 1.6) and the double complement law (Example 1.7).

[7] Augustus De Morgan (1806–1871): English mathematician, who was an early pioneer in formal logic, and first president of the London Mathematical Society.

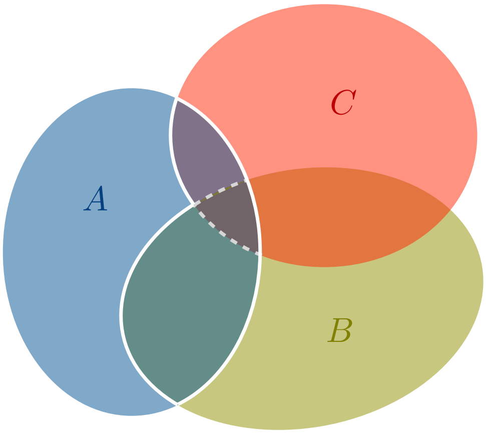


Figure 1.1.: Venn diagram for the first distributive law

Example 1.6 (First distributive law)**.**
Repeated application of the definitions of intersection and union (Sections 1.1.3 and 1.1.4) allow us to "repackage" the elements of one set into those of the other, as follows:

$$\begin{aligned} & x \in A \cap (B \cup C) \\ \iff \quad & x \in A \text{ and } x \in B \cup C \\ \iff \quad & x \in A, \text{ and } x \in B \text{ or } x \in C \\ \iff \quad & x \in A \text{ and } x \in B, \text{ or } x \in A \text{ and } x \in C \\ \iff \quad & x \in A \cap B \text{ or } x \in A \cap C \\ \iff \quad & x \in (A \cap B) \cup (A \cap C). \end{aligned}$$

Note (by reading backwards from the end) that all the implications are reversible. Note also the subtle use of punctuation to clarify the meaning of some of the statements in the sequence, which could otherwise be ambiguous.

This shows that the sets $A \cap (B \cup C)$ and $(A \cap B) \cup (A \cap C)$ have precisely the same elements, and are therefore equal, by the Principle for Equality of Sets. □

Example 1.7 (Double complement law)**.**
For this law to make sense we assume that $A \subseteq X$. Now suppose $x \in X$. By the definition of set complement (Section 1.1.7) we have:

$$x \in (A^c)^c \iff x \notin A^c.$$

Since to be an element of A^c is to not be an element of A, to not be an element of A^c is to "not not be" an element of A; in other words, to be an element of A. (We've used the logical principle of the double negative; more of this in Chapter 2). Therefore:

$$x \notin A^c \iff x \in A.$$

Putting these two bi-implications together gives us:

$$x \in (A^c)^c \iff x \in A.$$

Hence A and $(A^c)^c$ have precisely the same elements, and are therefore equal, by the Principle for Equality of Sets. □

Remarks 1.8.

1) The two proofs given in Examples 1.6 and 1.7 boil down to manipulation of statements involving the logical operations "and", "or" and "not". This is evidence of the close relationship between set theory and formal logic, mentioned in Remark 1.3.

2) Writing a formal proof for a result that may seem obvious, such as Example 1.7, can sometimes be surprisingly tricky. Nevertheless, all mathematical statements require proof, even the "obvious" ones! ◊

1.2. Functions

Functions tell us how to move between sets, and provide the basis for comparing one set with another.

Definition 1.1 (Function)**.**
A *function* $f\colon A \to B$ consists of three things:

- A set A called the *domain* of the function (sometimes known as its *source*).
- Another set B called the *codomain* of the function (also known as its *target*). It's allowed that $B = A$.
- A *rule* f which describes unambiguously how to associate to each element $a \in A$ a *unique* element $f(a) \in B$. ♦

Note. This is our first formal definition! Definitions have an extremely important status in mathematics, and because they need to be highly accurate can often make simple ideas appear rather complicated. Nevertheless, don't read any more (or less) into a definition than is actually there! ◊

Remarks 1.9 (Functions).

1) The notation $A \xrightarrow{f} B$ is occasionally used.

2) The terminology *map, mapping,* or *transformation* is also used.

3) The notation $a \mapsto f(a)$ is often used when defining the rule f; for example, the rule $f(x) = x^2$ (for a function $f\colon \mathbb{R} \to \mathbb{R}$) is also denoted $x \mapsto x^2$.

4) If $a \in A$ then $f(a)$ is sometimes called the *value* of f at a, and a itself is (rather quaintly) the *argument* for this value. It's also common to refer to a as the *input,* or *independent variable,* and $f(a)$ as the *output,* or *dependent variable.*

5) Despite the above Note, there is a whiff of vagueness in Definition 1.1; in particular, what exactly is meant by a "rule"? (If everything smells fine, then read no further!) In fact it's possible to give a completely rigorous definition; see Remark 3.3. However, this turns out to be considerably more abstract than the idea it's trying to encapsulate, which as we noted is sometimes the price we pay for mathematical precision. ◊

We'll see many examples of functions during the course. A function should *always* have a domain and a codomain specified as well as a rule. The rule f doesn't have to be an algebraic expression, but it does have to be unambiguous:

> *If a function is handed an element a of the domain, then it hands back a* single *element $f(a)$ of the codomain.*

Example 1.8 (Square root function)**.**
The rule "take the square root of a non-negative real number x" is ambiguous, since if $x > 0$ there are two possibilities: $\pm\sqrt{x}$. So this rule doesn't define a function. However, if we agree to choose the non-negative square root, which by convention is denoted $\sqrt{x}$, then this is unambiguous and therefore defines the following function:

$$f\colon \mathbb{R}_0^+ \to \mathbb{R}; x \mapsto \sqrt{x}.$$

The domain of this function is $\mathbb{R}_0^+$ and its codomain is $\mathbb{R}$. □

As we get more relaxed with functions, it's common to just talk about "the function f". This is okay, but shouldn't be taken to mean that we can forget the domain and codomain, even though it may seem that they have been swept under the carpet!

1.2.1. Principle for Equality of Functions

This principle helps us to clarify what we mean by saying that two functions are "the same", and equally importantly when they're not. We say that functions $f\colon A \to B$ and $g\colon C \to D$ are equal if:

- $A = C$; ie. they have the same domain;
- $B = D$; ie. they have the same codomain;
- $f = g$; ie. they have the same rule.

In other words, each of the three ingredients of a function (domain, codomain, rule) are the same.

Having the same rule means:

$$f(a) = g(a), \quad \text{for } \textit{all } a \in A. \tag{1.2}$$

Note that for (1.2) to make sense requires f and g to have the same domain and codomain, so the three parts of the definition follow on from each other naturally.

What notation should we use for equality of functions? Perhaps:

$$(f\colon A \to B) = (g\colon C \to D).$$

However, this is incredibly "clunky". So we usually just write $f = g$. (There is another unwritten mathematical principle at work here: the *Principle of Notational Simplicity,* a generalisation of the "Prinicple of Minimal Bracketing". Provided it's used with care, it makes mathematical life easier for everyone.)

1.2.2. Images

Given a function $f\colon A \to B$ and an element $a \in A$ we sometimes refer to the value $f(a) \in B$ as the *image of a* under f. We then use the same terminology in a collective sense, as follows.

Definition 1.2 (Image)**.**

Suppose $f\colon A \to B$ is a function. If $X \subseteq A$ then the *image of X* under f is the subset $f(X) \subseteq B$ defined:

$$f(X) = \{f(x) \mid x \in X\}.$$

In particular, the set $f(A)$ is usually called simply the *image of f,* and often also denoted by $\operatorname{im}(f)$. It is the set of values in the codomain B that the function takes as we plug in *all* the elements of the domain A. ♦

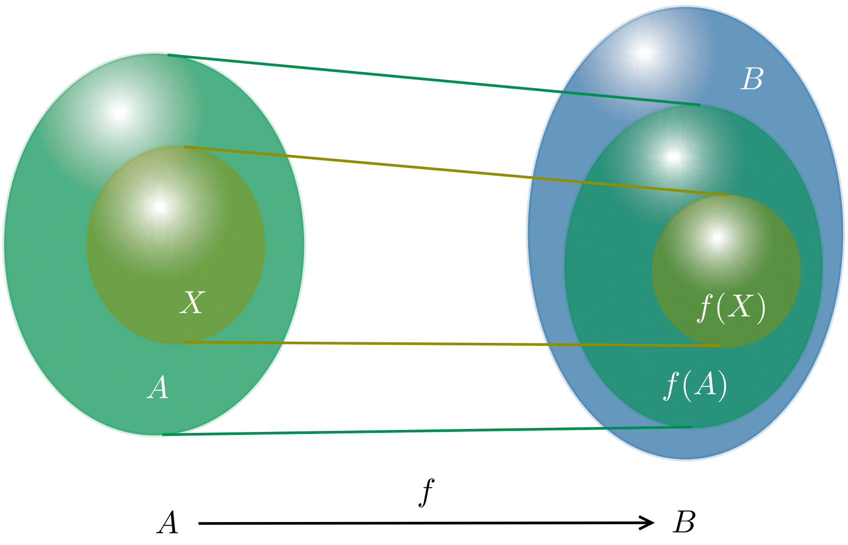


Figure 1.2.: Function images

Remarks 1.10 (Images).

1) We must be careful not to confuse a function with its set of values; ie. its image. There may be many rules that produce this set; in other words, *different* functions can have the *same* image. For example, the functions $f, g\colon \mathbb{R} \to \mathbb{R}$ with rules $f(x) = x$ and $g(x) = x^3$ both have image $\mathbb{R}$.

2) The codomain can be *much* larger than the image; an extreme example is $f\colon \mathbb{R} \to \mathbb{R}$ defined $f(x) = 0$ for all $x \in \mathbb{R}$, with codomain $\mathbb{R}$ and image $\{0\}$. Thus, the function $f\colon A \to B$ is strictly speaking different to the function $f\colon A \to f(A)$. Often a major problem is to determine what $f(A)$ actually is!

3) Some people use the word "range" to describe the image, and others use the word "range" to describe the codomain. It's confusing, so we're not going to use the word "range" at all! Nor are we going to talk about the equally confusing notion of "maximal domain" (which is sometimes used out "in the wild").

4) If $X \subset A$ (a proper subset) it is nevertheless possible for $f(X) = f(A)$. *(Can you think of any (really simple) examples?)* $\Diamond$

Example 1.9 (Square root function; images).
Continuing Example 1.8, the codomain of the function:

$$f\colon \mathbb{R}_0^+ \to \mathbb{R}; x \mapsto \sqrt{x}$$

is (clearly) $\mathbb{R}$, whereas its image is $\mathbb{R}_0^+$ (since none of its values is negative, and every non-negative real number y is the square root of a non-negative real number x, namely

$x = y^2$). Notice that determining the image requires more thought than the codomain, which we simply read off from the definition of f.

Now, here are the images of certain subsets of $\mathbb{R}_0^+$ under f:

$$\begin{aligned} f(\mathbb{N}) &= \{1, \sqrt{2}, \sqrt{3}, 2, \sqrt{5}, \sqrt{6}, \sqrt{7}, \sqrt{8}, 3, \sqrt{10}, \dots\}, \\ f(a,b) &= (\sqrt{a}, \sqrt{b}), \quad \text{for } 0 \leqslant a < b, \\ f[a,\infty) &= [\sqrt{a}, \infty), \quad \text{for } a \geqslant 0. \end{aligned}$$

It's interesting to note that $f(\mathbb{N}) \supset \mathbb{N}$. This may seem counter-intuitive, since functions associate one and only one element of the codomain to each element the domain and therefore cannot increase the "size" of any subset. However, $\mathbb{N}$ and $f(\mathbb{N})$ are infinite sets, so great care is required when measuring their "size"; we will take a closer look at this, and further counter-intuitive situations that can arise, in Section 1.2.7 and Set Piece 1. Note also that stictly speaking we should write $f((a,b))$ and $f([a,\infty))$, but we don't; the Principle of Minimal Bracketing comes to our rescue!

What about the image of $\mathbb{Q}^+$ (the positive rationals)? The ancient Greeks thought, at least in the early days, that $f(\mathbb{Q}^+) = \mathbb{Q}^+$ (in other words, that the square root of any rational number is always rational), which very much suited their mathematical world view. But they were in for a rather rude awakening (see Theorem 2.5). □

1.2.3. One-to-one and onto functions

There are two very important, and desirable properties, that a function may (or may not) have.

Definition 1.3 (One-to-one and/or onto)**.**
Suppose $f\colon A \to B$ is a function.

- We say that f is *one-to-one* (synonym: *injective*), often abbreviated *1-1,* if *distinct* elements of the domain A are sent to *distinct* elements of the codomain B by the rule f; more precisely:

  $$\textit{If } x, y \in A \textit{ with } x \neq y \textit{ then } f(x) \neq f(y). \tag{1.3}$$

  This eliminates the "collapsing" behaviour illustrated in the first diagram of Figure 1.3 below. However, in practice (and this is really the key point) we prefer to check the following logically equivalent statement (known officially as the "contrapositive"; we'll have more to say about this in Chapter 2, Section 2.4.2):

  $$\textit{If } x, y \in A \textit{ satisfy } f(x) = f(y) \textit{ then } x = y. \tag{1.4}$$

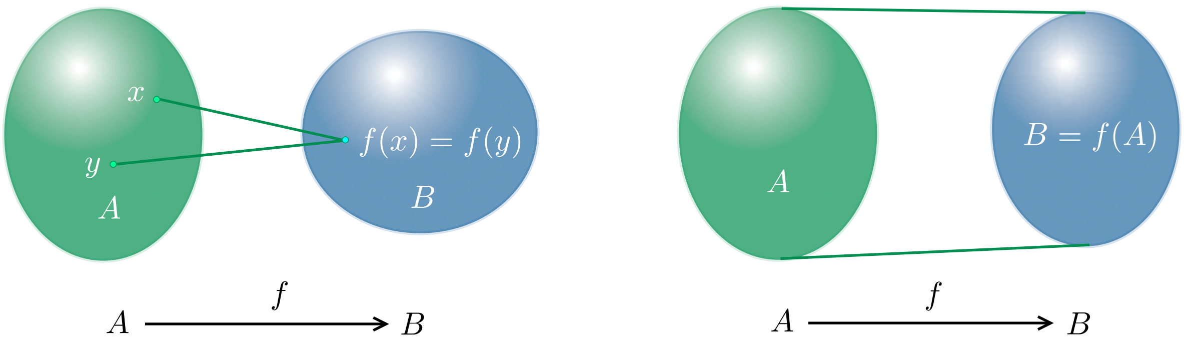
  This is preferable because it involves equations, which are much easier to work with than "non-equations". We refer to (1.4) as the "standard routine" for showing one-to-one-ness, and will use it many times throughout the course.

- We say that f is *onto* (synonym: *surjective*) if *every* element of B comes from *some* element of A via f. More precisely:

  Given any $b \in B$ there exists some $a \in A$ such that $f(a) = b$.

  The element a isn't necessarily unique, and if there are several then we don't have to give all of them in order to verify the condition; one is enough!

  The "onto" property is illustrated in the second diagram of Figure 1.3. If we imagine f as "painting" B with points of A then f being onto says that all of B gets painted (it might even get multiple "coats"!). ♦



Not one-to-one

Onto

Figure 1.3.: Behaviour and misbehaviour of functions

Remark 1.11. The condition for $f\colon A \to B$ to be surjective can be written in an abbreviated form, using quantifiers:

$$(\forall b \in B)(\exists a \in A)\; f(a) = b.$$

This is convenient if we need logical clarity (we'll show why this could be useful in Example 2.2 of Chapter 2). However, the shortest and perhaps most elegant form of the definition is:

$$f(A) = B.$$

We are free to use whichever version we prefer! ◊

Example 1.10 (Increasing and decreasing functions)**.**
Suppose $f\colon A \to B$ where A, B are subsets of $\mathbb{R}$. (This is probably the type of function we're most familiar with.) In this case it makes sense to talk about f being "increasing", or "decreasing". More precisely:

- ○ f is *increasing* if $x < y$ implies $f(x) < f(y)$. (Sometimes we say f is *strictly increasing* to emphasise the strict inequalities.)

- f is *(strictly) decreasing* if $x < y$ implies $f(x) > f(y)$.

These conditions become very clear geometrically when we draw the graph of f.

In either case, f is one-to-one. Interestingly, to show this it's easier to use the original definition (1.3) of one-to-one-ness, rather than the "standard routine". For, if f is increasing and $x \neq y$ then either $x < y$ or $y < x$, in which case $f(x) < f(y)$ or $f(y) < f(x)$, hence $f(x) \neq f(y)$. □

Remark 1.12. If f is differentiable then calculus provides a neat way to check whether f is increasing or decreasing, by inspecting the sign of its derivative: positive implies (strictly) increasing, negative implies (strictly) decreasing. ◊

If a function is neither one-to-one nor onto (see Example 1.11 below) all is not lost: we may be able to salvage injectivity by restricting the domain, and surjectivity by reducing the size of the codomain, without having to change the rule at all.

Example 1.11 (Injective and surjective functions)**.**
Consider $f\colon \mathbb{R} \to \mathbb{R}$ defined $f(x) = x^2$ (note that this rule is unambiguous and therefore does indeed define a function). Then f is not one-to-one, because for example:

$$f(-1) = 1 = f(1).$$

Furthermore f isn't onto, because there is *no* real number x with $f(x) = -1$, for example.

Now if we restrict the domain of f to $\mathbb{R}_0^+$ we obtain what is strictly speaking a new function $f\colon \mathbb{R}_0^+ \to \mathbb{R}$ defined by exactly the same rule (which is why we keep the same name f for it) but which is now one-to-one. We'll verify this using the "standard routine". Suppose $f(a) = f(b)$ for $a, b \in \mathbb{R}_0^+$; thus:

$$a^2 = b^2,$$

which we rearrange to a difference of two squares and factorise:

$$(a-b)(a+b) = 0.$$

Hence either $a - b = 0$ or $a + b = 0$, so $a = b$ or $a = -b$. The latter is impossible unless $a = b = 0$, because a, b are both non-negative; therefore $a = b$.

On the other hand, if we reduce the codomain of f from $\mathbb{R}$ to the image $f(\mathbb{R})$ we immediately obtain a surjection $f\colon \mathbb{R} \to f(\mathbb{R})$. The question is then, what is $f(\mathbb{R})$? Since squares of real numbers are non-negative we have $f(\mathbb{R}) \subseteq \mathbb{R}_0^+$. And every $y \in \mathbb{R}_0^+$ can be written $y = f(x)$ where $x = \sqrt{y}$, so $\mathbb{R}_0^+ \subset f(\mathbb{R})$. Thus $f(\mathbb{R}) = \mathbb{R}_0^+$ by the Principle of Mutual Containment. So the function $f\colon \mathbb{R} \to \mathbb{R}_0^+$, defined by exactly the same rule, is onto.

Putting both these modifications in place produces a function:

$$f\colon \mathbb{R}_0^+ \to \mathbb{R}_0^+; x \mapsto x^2,$$

which is both one-to-one and onto. □

Remark 1.13 (Restrictions of functions).
There are several reasons why we might want to restrict the domain of a function $f\colon A \to B$ to a subset $X \subset A$ (such as the situation described in Example 1.11), and in effect create a new function $g\colon X \to B$. However, more often than not we continue to use the same symbol (namely "f") for the rule, because it won't have changed. Nevertheless there are times when it's helpful to indicate notationally that the rule has been restricted to act only on elements of X, in which case we denote it by $f|_X$ and use terminology like: "f restricted to X", or: "the restriction of f to X". ◊

Example 1.12 (Function on a Cartesian product)**.**
For a sligthly different kind of function, consider $f\colon \mathbb{N} \times \mathbb{N} \to \mathbb{Z}$ defined:

$$f(m,n) = n - m, \quad \text{for all } m,n \in \mathbb{N}.$$

(Recall that the Cartesian product $\mathbb{N} \times \mathbb{N}$ is the set of all ordered pairs of natural numbers; see Section 1.1.8.) Every integer is the difference of two natural numbers (Example 1.5), and can therefore be written as $f(m,n)$ for some $(m,n) \in \mathbb{N} \times \mathbb{N}$. Therefore f is onto. However $f(1,1) = 0 = f(2,2)$, for example, which shows that f isn't one-to-one. □

1.2.4. Bijections and inverse functions

If a function is one-to-one *and* onto, then it pairs up the elements of the domain and the codomain. This means that there is another function going the other way (ie. from codomain to domain) which "undoes" the rule f. This is the "inverse function". We formalise this idea as follows.

Definition 1.4 (Bijection and inverse function)**.**
Suppose $f\colon A \to B$ is a function.

- We say that f is *bijective* (synonyms are: f is a *bijection;* f is a *one-to-one correspondence*) if f is one-to-one *and* onto. Thus, combining the two parts of Definition 1.3, f is bijective if and only if the following condition holds:

  $$\textit{For every } b \in B \textit{ there exists a } \text{unique } a \in A \textit{ such that } f(a) = b. \tag{1.5}$$

- The unique element $a \in A$ such that $f(a) = b$ is denoted $f^{-1}(b)$. The rule "associate to $b \in B$ the element $f^{-1}(b) \in A$" then defines a function $f^{-1}\colon B \to A$ called the *inverse function* of f. By definition, it satisfies the property:

  $$f(f^{-1}(b)) = b, \quad \text{for all } b \in B. \tag{1.6}$$

  Thus "f undoes f^{-1}".

Note. If we take $b = f(a)$ in (1.5) then we see immediately that:

$$f^{-1}(f(a)) = a, \quad \text{for all } a \in A. \tag{1.7}$$

Thus "f^{-1} undoes f". ♢

The relationship between bijections and their inverses can be taken one step further, as described in the following result.

Proposition 1.1 (Inverse of a bijection)**.**

i) *The inverse of a bijection $f\colon A \to B$ is also a bijection.*

ii) *In this situation, the inverse of $f^{-1}\colon B \to A$ is $f\colon A \to B$; ie. $(f^{-1})^{-1} = f$.*

Proof. For notational convenience we denote $g = f^{-1}$; so $g\colon B \to A$.

i) We first verify that g is one-to-one, using the "standard routine". Suppose $g(b_1) = g(b_2)$ for $b_1, b_2 \in B$. Then, applying f to both sides of this equation and using (1.6) immediately gives us $b_1 = b_2$, as required.

We now verify that g is onto. Thus, suppose $a \in A$. Then by (1.7) we have $a = g(b)$ where $b = f(a)$.

It follows by definition that g is a bijection.

ii) Since g is a bijection it has an inverse $g^{-1}\colon A \to B$ which by (1.6) satisfies:

$$g(g^{-1}(a)) = a, \quad \text{for all } a \in A.$$

Applying f to both sides of this equation and using (1.6) gives us:

$$g^{-1}(a) = f(a), \quad \text{for all } a \in A.$$

Therefore $g^{-1} = f$ by the Principle for Equality of Functions. ■

Note. This is our first formal result! It comes in two pieces: a precise description (the "statement") of what we claim to be true (which in this case we've split into two parts, to make things clearer), followed by an argument (the "proof") that justifies our claim. It's important that the proof uses only concepts and facts that have already been established, and is unerringly accurate whilst being as simple, comprehensible and elegant as possible! Achieving all this can often be quite tricky. Some would say that it's an art! Indeed, when Kurt Gödel received and read Paul Cohen's proof of the independence of the Continuum Hypothesis [10] (more about this in Chapter 2; Remarks 2.13 and 2.16) this was his reaction:

> *Let me repeat that it is really a delight to read your proof ... I think that in all essential respects you have given the best possible proof ... Reading your proof had a similarly pleasant effect on me as seeing a really good play.*

Whilst the proof in question lies beyond the scope of these notes, there will be many other examples that give some idea of what we're aiming for. ◊

Remarks 1.14.

1) We used subscripts during the proof! This form of notational decoration can be very handy (it helps us to observe the unwritten and entirely informal *Principle of Conservation of Symbols*), and we'll be seeing a lot more of it.

2) Taking the inverse of a function is one of many examples of a mathematical operation that when performed twice gets us back where we started; other examples are taking the complement of a subset (Example 1.7), and (for those familiar with matrices) taking the transpose of a matrix. In general, such operations are called *involutions.* ◊

There remains the practical question of how we go about finding inverse functions, which usually boils down to determining their rule. In general, this is exceedingly tricky; in fact, it could be said (with only a certain amount of exaggeration) that a large proportion of mathematics is concerned with doing just this, in one context or another. For example, integration is the "inverse of differentiation" (and can indeed be viewed as an inverse function if we set things up carefully enough), and we all know how vexatious working out integrals can be! Here's something far simpler.

Example 1.13 (Linear bijections)**.**
Define $f\colon \mathbb{R} \to \mathbb{R}$ by:

$$f(x) = ax + b,$$

where $a, b \in \mathbb{R}$ with $a \neq 0$. (We recognise f as a *linear,* or more precisely an *affine* function.) Then f is one-to-one, because if $f(x) = f(y)$ then cancelling the b's and dividing through by a gives us $x = y$. Furthermore f is onto. For, given any $y \in \mathbb{R}$ we can solve the equation $ax + b = y$ to obtain:

$$x = \frac{y-b}{a};$$

thus $y = f(x)$ with x as above. So f is a bijection, and therefore has an inverse function $f^{-1}\colon \mathbb{R} \to \mathbb{R}$, which by definition has rule:

$$f^{-1}(y) = \frac{y-b}{a}, \quad \text{for all } y \in \mathbb{R}.$$

Notice that we effectively found the rule for the inverse function when showing that f is onto; this is often the case. □

Sometimes we are able to spot a rule for "undoing" a function relatively easily, without having formally checked that our function is a bijection (see Example 1.14 below). In such situations we would like to know that it's "mission accomplished", making such a check unnecessary.

Definition 1.5 (Invertible function)**.**
A function $f\colon A \to B$ is *invertible* if there exists a function $g\colon B \to A$ such that:

$$g(f(a)) = a, \qquad f(g(b)) = b, \tag{1.8}$$

for all $a \in A$ and all $b \in B$. ◆

Remarks 1.15.

1) A function g satisfying both of the equations (1.8) is necessarily unique. For, if h is another such function then it follows from (1.8) (applied to both h and g) that for all $b \in B$:
$$h(b) = h\big(f(g(b))\big) = g(b).$$
Hence $h = g$, by the Principle for Equality of Functions.

2) By definition, the function g satisfying (1.8) is also invertible. ◇

It follows from equations (1.6) and (1.7) that bijections are invertible. The following result shows that the converse is also true; so bijections and invertible functions are precisely the same thing.

Proposition 1.2 (Invertible functions are bijections)**.**
Suppose $f\colon A \to B$ is invertible. Then f is a bijection, and $f^{-1} = g$, where g is the (unique) function satisfying equations (1.8).

Proof. We first show that f is one-to-one, using the "standard routine". If $f(a_1) = f(a_2)$ for $a_1, a_2 \in A$ then by applying g to both sides of this equation and using (1.8) it follows immediately that $a_1 = a_2$.

Now to show f is onto, simply observe that if $b \in B$ then it follows from (1.8) that $b = f(a)$ where $a = g(b)$.

It follows that f is a bijection (Definition 1.4), with inverse function f^{-1} satisfying equations (1.6) and (1.7). Comparing with equations (1.8) and noting the uniqueness of g (Remark 1.15 (1)), we infer that $f^{-1} = g$. ■

Remark 1.16 (Left-invertible and right-invertible functions).
Close-up inspection of the proof of Proposition 1.2 reveals that when showing f is one-to-one we used only the left hand equation of (1.8), whereas we used the right hand equation to show that f is onto. When we want to distinguish between these two equations we say that f is *left-invertible,* or *right-invertible,* respectively.

It's possible for only one of equations (1.8) to be satisfied. For example, if $f\colon \mathbb{N} \to \mathbb{N} \times \mathbb{N}$ and $g\colon \mathbb{N} \times \mathbb{N} \to \mathbb{N}$ are defined by the rules:

$$f(n) = (n, n), \qquad g(m, n) = m,$$

then for all $n \in \mathbb{N}$ we have:

$$g(f(n)) = g(n,n) = n.$$

So f is left-invertible, and g is a *left-inverse* of f (or, equivalently, g is right-invertible, and f is a *right-inverse* of g). However, for all $(m,n) \in \mathbb{N} \times \mathbb{N}$ we have:

$$f(g(m,n)) = f(m) = (m,m),$$

so g isn't a right-inverse of f (or, equivalently, f isn't a left-inverse of g).

Unlike inverse functions, left-inverse functions (or right-inverses) need not be unique. For example, if f is the function defined in the previous paragraph then:

$$h\colon \mathbb{N} \times \mathbb{N}; (m,n) \mapsto n,$$

is another left-inverse of f.

It turns out that left-invertibility is is equivalent to being one-to-one, whereas right-invertibility is equivalent to being onto. *(Can you prove this?)* ♢

Example 1.14 (Set complement as a function on the power set)**.**
Let A be any set and define $f\colon \mathscr{P}(A) \to \mathscr{P}(A)$ by the rule:

$$f(X) = X^c,$$

for all $X \in \mathscr{P}(A)$. (Recall from Section 1.1.6 that the power set $\mathscr{P}(A)$ is the set of all subsets of A, and from Section 1.1.7 that X^c is the complement $A \smallsetminus X$.) Then by the double complement law (Example 1.7):

$$f(f(X)) = f(X^c) = (X^c)^c = X.$$

Hence f is invertible (since equations 1.8 are are satisfied with $g = f$), and it follows from Proposition 1.2 that f is a bijection with $f^{-1} = f$. □

1.2.5. Composition of functions

Suppose $f\colon A \to B$ and $g\colon B \to C$ are functions. Then since the domain of g is the codomain of f we can apply the rules f and g consecutively (ie. one after the other) to get a rule taking us from A to C:

$$A \xrightarrow{f} B \xrightarrow{g} C.$$

This leads us to the following important construct.

Definition 1.6 (Composition)**.**
Suppose $f\colon A \to B$ and $g\colon B \to C$ are functions. Then the *composition* is the function denoted by $g \circ f : A \to C$ whose rule is defined:

$$(g \circ f)(a) = g(f(a)), \qquad \text{for all } a \in A.$$

Thus $g \circ f$ is the rule: "do f then do g". ♦

Remarks 1.17 (Elementary properties of composition).

1) We can only compose two functions if the codomain of the first is the same as (or more generally, a subset of) the domain of the second; otherwise, the rule for $g \circ f$ doesn't make sense!

2) Composition of functions is associative:

$$h \circ (g \circ f) = (h \circ g) \circ f,$$

for any three functions:

$$A \xrightarrow{f} B \xrightarrow{g} C \xrightarrow{h} D.$$

In other words, for all $a \in A$ we have:

$$(h \circ (g \circ f))(a) = ((h \circ g) \circ f)(a). \tag{1.9}$$

For, (1.9) expands out to:

$$h\big((g \circ f)(a)\big) = (h \circ g)(f(a)),$$

and ultimately:

$$h\big(g(f(a))\big) = h\big(g(f(a))\big).$$

3) Having established associativity, we can use it to remove brackets completely and write:

$$h \circ g \circ f,$$

without any danger of ambiguity.

4) Composition of functions isn't commutative:

$$g \circ f \neq f \circ g.$$

Indeed, in general the composition $f \circ g$ won't make sense, unless $B = A$. However, even in this case, two functions $f, g \colon A \to A$ won't in general commute; see Example 1.15.

5) The composition notation allows us to invoke the Principle for Equality of Functions to "clean up" the two equations (1.6) and (1.7) for the inverse function, as follows:

$$f \circ f^{-1} = 1_B, \qquad f^{-1} \circ f = 1_A,$$

where $1_A \colon A \to A$ and $1_B \colon B \to B$ are the *identity functions:*

$$1_A(a) = a, \qquad 1_B(b) = b,$$

for all $a \in A$ and $b \in B$. The identity function on a set is innocuous, but plays a very important rôle. We will meet it again in Chapter 4 (Section 4.5). ◊

Composition is a fundamental technique for building new functions from old. In all likelihood, it's something we first met (perhaps unwittingly) when studying elementary calculus with functions $f, g\colon \mathbb{R} \to \mathbb{R}$, where it's often referred to as "a function of a function". In particular, the Chain Rule of differential calculus tells us how to differentiate a composition of differentiable functions, which when written out using the "dash for derivative" notation of Lagrange[8] looks like:

$$(g \circ f)'(x) = g'(f(x))f'(x).$$

Despite appearances, this is exactly the same as the "traditional" version, expressed using the "d by dx" notation of Leibniz[9]:

$$\frac{dv}{dx} = \frac{dv}{du}\frac{du}{dx},$$

once the meaning of this notation has been properly established; ie. we set $u = f(x)$ and $v = g(u) = g(f(x))$, then use v to denote the two different functions f and $g \circ f$, and treat u as both a function and an independent variable! One of the challenges of mastering calculus and analysis is to reconcile this sort of "doublethink" with the rigorous ideas that we're describing in this course.

Here are some simple examples of "a function of a function", which also show that composition is not commutative (Remark 1.17 (4)).

Example 1.15 (Function of a function)**.**
Suppose $f, g\colon \mathbb{R} \to \mathbb{R}$ are given by $f(x) = x^2$ and $g(x) = x + 1$. Then for all $x \in \mathbb{R}$:

$$g \circ f(x) = g(x^2) = x^2 + 1,$$

whereas:

$$f \circ g(x) = f(x+1) = (x+1)^2.$$

For these two (composite) functions to be decreed equal they have to agree for *all* x, by the Principle for Equality of Functions. However $f \circ g(1) = 4$ whereas $g \circ f(1) = 2$. Thus $g \circ f \neq f \circ g$, so this pair of functions do not commute. □

Composition preserves some of our favourite properties of functions, as summarised by the following result.

[8] Joseph-Louis Lagrange (1736–1813): Franco/Italian mathematician, whose mathematical interests ranged from calculus and mechanics through to number theory and algebra.

[9] Gottfried Wilhelm Leibniz (1646–1716): German mathematician and philosopher, who developed the main ideas of differential and integral calculus independently from Newton, and whose notation is far better!

Proposition 1.3 (Composition of bijections)**.**
The composition of injections is injective and the composition of surjections is surjective. Therefore, the composition of bijections is bijective, and in this case:

$$(g \circ f)^{-1} = f^{-1} \circ g^{-1}.$$

Proof. Suppose f, g are injective. We show that $g \circ f$ is injective, using the "standard routine":

$$\begin{aligned} g \circ f(x) = g \circ f(y) &\implies g(f(x)) = g(f(y)) \\ &\implies f(x) = f(y), \quad \text{because } g \text{ is 1-1} \\ &\implies x = y, \quad \text{because } f \text{ is 1-1.} \end{aligned}$$

Thus $g \circ f$ is one-to-one.

Now suppose f, g are surjective. Given $c \in C$, there is an element $b \in B$ with $g(b) = c$, since g is onto. Furthermore there is an element $a \in A$ with $f(a) = b$, since f is onto. Hence:

$$c = g(b) = g(f(a)) = g \circ f(a).$$

Thus $g \circ f$ is onto.

It follows immediately that if f, g are bijective then so is $g \circ f$. The inverse function $(g \circ f)^{-1} \colon C \to A$ is characterised by the condition that its value at any point $c \in C$ is the unique element $a \in A$ such that $g \circ f(a) = c$; in other words $g(f(a)) = c$. By (1.6) this equation is satisfied if $a = f^{-1}(g^{-1}(c)) = (f^{-1} \circ g^{-1})(c)$, hence by uniqueness:

$$(g \circ f)^{-1}(c) = (f^{-1} \circ g^{-1})(c), \quad \text{for all } c \in C,$$

which by the Principle for Equality of Functions allows us to write:

$$(g \circ f)^{-1} = f^{-1} \circ g^{-1}.$$ ■

Example 1.16 (Inverse of a composition)**.**
Suppose $h \colon \mathbb{R} \to \mathbb{R}$ is defined $h(x) = 2x + 5$ for all $x \in \mathbb{R}$. Define $f \colon \mathbb{R} \to \mathbb{R}$ by $f(x) = 2x$ and $g(x) = x + 5$. Then $h = g \circ f$. It is easily checked *(and you should do this)* that f, g are bijections, with inverse functions:

$$f^{-1} \colon \mathbb{R} \to \mathbb{R}; f^{-1}(x) = \tfrac{1}{2}x, \qquad g^{-1} \colon \mathbb{R} \to \mathbb{R}; g^{-1}(x) = x - 5.$$

Then Proposition 1.3 tells us that h is a bijection with inverse $h^{-1} \colon \mathbb{R} \to \mathbb{R}$ defined:

$$h^{-1}(x) = (f^{-1} \circ g^{-1})(x) = f^{-1}((g^{-1}(x)) = f^{-1}(x - 5) = \tfrac{1}{2}(x - 5).$$

This agrees with Example 1.13, where h was shown to be bijective and its inverse calculated in one go. □

1.2.6. Preimages

Let $f\colon A \to B$ be a function, and suppose $Y \subseteq B$. We can ask which elements of A end up in the subset Y when we apply the rule f. This is a very common question in mathematics, so the concept has a name and a notation.

Definition 1.7 (Preimage)**.**
Suppose $f : A \to B$ is a function, and $Y \subseteq B$. We define the *preimage* of Y under f to be the following subset of A:

$$f^{-1}(Y) = \{a \in A \mid f(a) \in Y\}.$$

This is the subset of all points of A that map into Y (see Figure 1.4). ♦

Note. The notation can be confusing, because we use the symbol f^{-1} even though f might not have an inverse! We just have to live with this—the preimage of a subset of the codomain *always* makes sense. However, if f is a bijection then $f^{-1}(Y)$ coincides with the image of Y under the inverse function f^{-1} (Definition 1.2), so the two notations are consistent. It should be apparent from the context what is going on. ♢

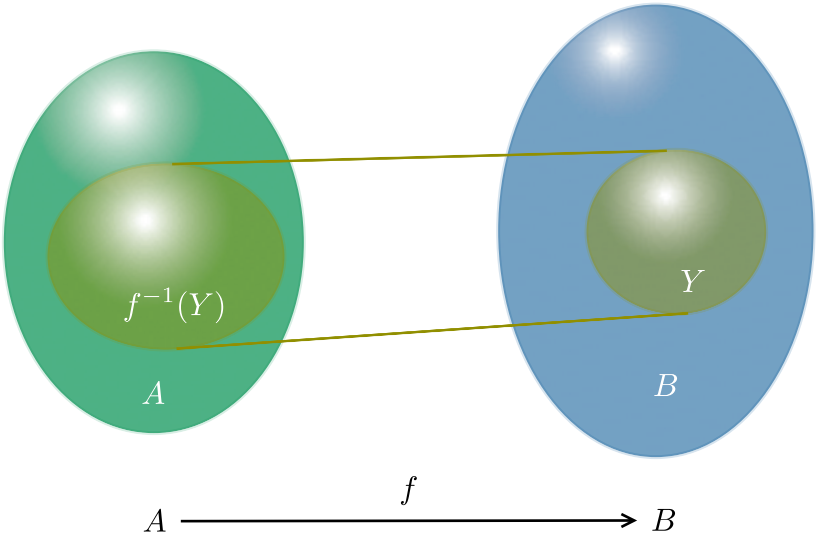


Figure 1.4.: Function preimage

Remark 1.18 (Images and preimages).
By definition, the image of a preimage is the original subset:

$$f(f^{-1}(Y)) = Y,$$

for all $Y \subseteq B$. However, if $X \subseteq A$ then the preimage of its image certainly contains X, but could in fact be larger. So in general we can only say:

$$X \subseteq f^{-1}(f(X)).$$

It's not hard to construct examples where the inclusion is strict; see Example 1.17 (2) below. ♢

Remark 1.19 (Fibres of a function).
If $Y = \{y\}$ for some $y \in B$, then we usually write $f^{-1}(y)$ rather than $f^{-1}\{y\}$ or $f^{-1}(\{y\})$; thus:

$$f^{-1}(y) = \{x \in X \mid f(x) = y\}.$$

We refer to this as the *fibre* of f over y, or (particularly in calculus; see Example 1.18) the *level set* set of f with value y. ♢

Examples 1.17 (Preimages).
We return to the function $f\colon \mathbb{R} \to \mathbb{R}; x \mapsto x^2$. Note that f is not bijective, so there is no inverse function $f^{-1}\colon \mathbb{R} \to \mathbb{R}$. Nevertheless, we can use the symbol f^{-1} when dealing with preimages.

1) Here are a few preimages:

$$\begin{aligned}
f^{-1}(\mathbb{R}) &= \mathbb{R}, \\
f^{-1}(\mathbb{R}_0^+) &= \mathbb{R}, \\
f^{-1}(\mathbb{R}^+) &= \mathbb{R}^*, \\
f^{-1}(a) &= \{-a, a\}, \quad \text{for } a > 0, \\
f^{-1}[0, a) &= (-\sqrt{a}, \sqrt{a}\,), \quad \text{for } a > 0, \\
f^{-1}[a, b) &= (-\sqrt{b}, -\sqrt{a}\,] \cup [\sqrt{a}, \sqrt{b}\,), \quad \text{for } 0 < a < b, \\
f^{-1}(-\infty, 0] &= \{0\}, \\
f^{-1}(-\infty, 0) &= \emptyset.
\end{aligned}$$

(Can you verify all of these?)

2) Here's a counter-example to the statement $f^{-1}(f(X)) = X$ for all subsets X of the domain (see Remark 1.18). In the spirit of making counter-examples as simple as possible, let $X = \{1\}$. Then $f(X) = \{1\}$ hence:

$$f^{-1}(f(X)) = f^{-1}(1) = \{-1, 1\} \supset X.$$

On the other hand, if we take the preimage first then:

$$f(f^{-1}(X)) = f\{-1, 1\} = \{1\} = X,$$

as noted in Remark 1.18. □

Example 1.18 (Circles as preimages)**.**
This example shows that the idea of preimages is more familiar than it may appear. Let's define $f\colon \mathbb{R}^2 \to \mathbb{R}$ by:

$$f(x,y) = x^2 + y^2,$$

and suppose $k \in \mathbb{R}$. If $k < 0$ then $f^{-1}(k) = \emptyset$, whereas if $k = 0$ then $f^{-1}(k) = \{(0,0)\}$. However, if $k > 0$ and we write $k = r^2$ for $r > 0$ then the fibre of f over k is:

$$f^{-1}(k) = \{(x,y) : x^2 + y^2 = r^2\}.$$

Visualising $\mathbb{R}^2$ as the Cartesian plane, we recognise this as the circle of radius r centred at the origin $(0,0)$; see Figure 1.5. So in geometry and calculus, fibres of functions are nothing other that what we normally call *level curves* (or in higher dimensions, *level surfaces,* etc.) defined by an "implicit" equation. In fact, there is a whole area of mathematics—algebraic geometry—devoted to the study of level sets of polynomials! □

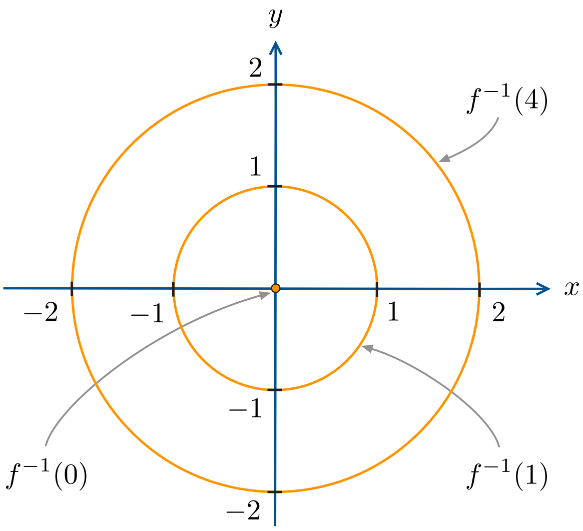


Figure 1.5.: Circles as preimages

Example 1.19 (Fermat's Last Theorem)**.**
Let n be a positive integer. Then Fermat's Last Theorem says that the equation:

$$x^n + y^n = z^n$$

has no solutions $x, y, z \in \mathbb{N}$ if $n > 2$. (The equation is easy to solve if $n = 1$, and if $n = 2$ there are also infinitely many solutions, commonly known as "Pythagorean triples"; see [36], for example.) We can rephrase this statement by defining a function $f\colon \mathbb{N}^3 \to \mathbb{Z}$ (where $\mathbb{N}^3 = \mathbb{N} \times \mathbb{N} \times \mathbb{N}$, the set of ordered triples of natural numbers) by:

$$f(x,y,z) = x^n + y^n - z^n.$$

Then the theorem says that $f^{-1}(0) = \emptyset$ if $n > 2$.

Despite the simplicity of its statement (whichever version we choose), Fermat's Last Theorem is immensely difficult to prove. First conjectured by Fermat[10] in 1637, it suffered the indignity of attracting more erroneous proofs than any other mathematical result (literally thousands!), before a complete proof finally emerged in 1995 after a remarkable sustained and persistent effort by Andrew Wiles[11] [41, 33]. □

1.2.7. Cardinality

The notion of a bijection allows us to define what it means for two sets to have the same "size".

We first consider the seemingly innocuous, perhaps even silly-sounding, question: "How should we define the notion of a *finite* set?" We can perhaps agree that for any natural number n the following set:

$$[n] = \{1, 2, \ldots, n\}$$

is finite, and has n elements. This will serve as a "model" for more general finite sets. So, suppose A is a set. To determine whether or not A is finite we attempt to "count" its elements, by first picking an element of A (any element will do), which for convenience we relabel a_1. We now pick any element of $A \setminus \{a_1\}$, and relabel it as a_2, then any element of $A \setminus \{a_1, a_2\}$, relabelling it as a_3, and so on. If there exists a natural number n such that $A \setminus \{a_1, \ldots, a_n\} = \emptyset$ then we can agree that A is finite, with n elements:

$$A = \{a_1, \ldots, a_n\}.$$

We now define a function $f_A \colon [n] \to A$ by $f(i) = a_i$. Then f_A is clearly onto, and one-to-one since by construction $a_i = a_j$ only if $i = j$. This suggests the following rather elegant way of defining finite sets, and the number of elements they contain.

Definition 1.8 (Finite and infinite sets)**.**
A non-empty set A is *finite* if there exists a bijection $f_A \colon [n] \to A$ for some $n \in \mathbb{N}$. The *cardinality* $|A|$ is then defined:

$$|A| = n.$$

We say A is *infinite* if A is not finite. ♦

[10] Pierre de Fermat (1607–1665): French mathematician, of whose own "last theorem" he famously wrote: "I have discovered a truly marvelous proof of this, which this margin is too narrow to contain". Unfortunately this proof has never been found, or reconstructed, leading many mathematicians to believe it was likely to have been one of the many false proofs that have appeared over the ages.

[11] Andrew Wiles (b. 1953): English mathematician, specialising in number theory, who has devoted much of his life to proving Fermat's Last Theorem.

Remarks 1.20.

1) The empty set $\emptyset$ doesn't satisfy our definition of a finite set (since it doesn't have any elements!), and we would certainly not want to categorise it as an infinite set! So we simply decree $\emptyset$ to be finite, with cardinality $|\emptyset| = 0$.

2) The definition of an infinite set may seem rather trite; it is more or less a tautology! We will express it in a more tangible (and useful) way after we've developed a bit of formal logic in Chapter 2 (see Example 2.4). ◊

Our definition of the cardinality of a finite set has a potential banana skin: could counting the elements of a set in different ways result in different cardinalities? More precisely, the procedure for constructing f_A involved making choices, and therefore if applied differently could conceivably produce different values of n. Unlikely as this may seem, were it possible then our definition would be worthless!

To see that this cannot happen, suppose we have two bijections:

$$f\colon [n] \to A, \qquad g\colon [m] \to A.$$

Then the composition:

$$F = g \circ f^{-1}\colon [n] \to [m]$$

is also a bijection (Propositions 1.1 and 1.3). We would like to be able to say that this implies $m = n$, something that seems almost obvious, but nevertheless requires proof! We include it as the punchline of our next result, which describes how the twin concepts of injective/surjective functions (Definition 1.3) pan out in the context of finite sets. The most compelling proof is by "mathematical induction", which we won't meet officially until Chapter 2, so we defer it for the time being and just give the statement.

Proposition 1.4 (Functions between finite sets)**.**
Suppose $f\colon [m] \to [n]$.

i) *If f is one-to-one then $m \leqslant n$.*

ii) *If f is onto then $m \geqslant n$.*

Therefore if f is a bijection then $m = n$.

Proof. See Proposition A.1 in Appendix A.1. ■

When we formulate mathematical definitions, particularly those such as Definition 1.8 that are strongly motivated by intuition, it is often a good idea to check that they allow us to deduce properties that we would expect. This is not because mathematicians take a perverse delight in proving things that are blindingly obvious, but rather because it assures us that our "mathematical model" is behaving in the right way and is therefore likely to be well-constructed. Our next result is one such example. The most compelling proof again involves mathematical induction, and once again we defer it for now.

Proposition 1.5 (Subsets of finite sets)**.**
If A is a finite set and $B \subseteq A$ then B is finite and $|B| \leqslant |A|$.

Proof. See Proposition A.2 in Appendix A.1. ■

Suppose A, B are finite sets, with cardinalities $|A| = n$ and $|B| = m$. So (Definition 1.8) there are bijections $f_A\colon [n] \to A$ and $f_B\colon [m] \to B$. If $m = n$ then it follows from Propositions 1.1 and 1.3 that the composition:

$$f_B \circ f_A^{-1}\colon A \to B$$

is a bijection between A and B. Conversely, if $f\colon A \to B$ is a bijection then the triple composition:

$$f_B^{-1} \circ f \circ f_A\colon [n] \to [m]$$

is a bijection, again using Propositions 1.1 and 1.3, and it follows from Proposition 1.4 that $n = m$. So, in summary, if A, B are finite sets then $|A| = |B|$ if and only if there exists a bijection $f\colon A \to B$. This is a characterisation of cardinality that generalises to infinite sets.

Definition 1.9 (Cardinality)**.**
Two sets A and B have the *same cardinality* if there exists a bijection $f\colon A \to B$. In this situation we write $|A| = |B|$. ♦

Note. The definition doesn't require us to assign individual meaning to the quantities $|A|$ and $|B|$, and for infinite sets we make no attempt to do so! ◊

Since bijections pair up the elements of one set with another, having the same cardinality carries the same meaning as it does in the finite case, of having the same number of elements, even though that number is not defined in absolute terms. In the context of infinite sets this has some interesting and slightly counter-intuitive consequences, which we'll begin to explore in our first Set Piece (!) of the course (see Examples 1.20, 1.21 and 1.22).

Remark 1.21. Having confessed to not defining $|A|$ and $|B|$ in general, it is perhaps rather presumptuous to write an "equation" $|A| = |B|$. However, since a bijection has an inverse which is also a bijection (Proposition 1.1), the statement "$|A| = |B|$" is true if and only if the statement "$|B| = |A|$" is also true. So our "equation" is behaving in the way we expect equations to behave; ie. symmetrically with respect to the "=" sign. Furthermore, our "equation" also behaves "transitively"; that is:

If $|A| = |B|$ and $|B| = |C|$ then $|A| = |C|$.

This follows from the fact that a composition of bijections is a bijection (Proposition 1.3), and is another basic property of equality that we take for granted. What we have here is in fact an example of an "equivalence relation", a generalised form of equality that we will see much more of in Chapter 3 (in particular, Example 3.2). ◊

1.3. Set Piece 1: Counting Infinity

References
Liebeck: Chapter 21, pp. 179–183.
Allenby: Chapter 9, pp. 154–155.
Franklin and Daoud: Chapter 11, Exercise 6.
Episode 4 of video lectures.

In Section 1.2.7 we learned what it means for two sets to have the same cardinality. A particularly interesting case is when one of those sets happens to be the natural numbers $\mathbb{N}$, which of course is an infinite set.

Definition 1.10 (Countable Set)**.**
Let A be a set. We say that:

- A is *countably infinite* if $|A| = |\mathbb{N}|$; ie. there exists a bijection between A and $\mathbb{N}$.
- A is *countable* if A is finite or countably infinite; ie. $|A| = n$ for some natural number n, or $|A| = |\mathbb{N}|$. (Sometimes the terminology *denumerable* is used.) ♦

Note. Some people (eg. Liebeck in his book [29]) use the word "countable" as a synonym for "countably infinite"; however most mathematicians use it to mean "countably infinite or finite". ♢

We're going to prove the following result, which might be a little bit surprising at first sight. It says that, although there are clearly more rational numbers than integers, there are in fact just as many! (Actually, it is even more surprising than that; see Remark 1.25).

Set Piece Theorem 1 (Countability of the rationals)**.**
The set $\mathbb{Q}$ of rational numbers is countable; ie. $|\mathbb{Q}| = |\mathbb{N}|$.

Before getting into the proof, we get warmed-up with several examples. The first shows that it is possible to have a countably infinite set with *fewer* elements than $\mathbb{N}$ (in some sense, half as many).

Example 1.20 (Countability of the even numbers)**.**
Let $\mathbb{E} = \{2n \mid n \in \mathbb{N}\}$, the set of even natural numbers. We claim that $|\mathbb{E}| = |\mathbb{N}|$. There's a natural function $f\colon \mathbb{N} \to \mathbb{E}$ defined:

$$f(n) = 2n, \quad \text{for all } n \in \mathbb{N},$$

which is a bijection; it's one-to-one because if $f(n) = f(m)$ then $2n = 2m$ so $n = m$ (this is the "standard routine"); and it's onto by definition of $\mathbb{E}$. So $\mathbb{E}$ is countable. □

Remark 1.22 (Countability of the odd numbers).
A similar argument can be used to construct a bijection between $\mathbb{N}$ and the set $\mathbb{O}$ of odd natural numbers *(we'll leave this to you);* so we have $|\mathbb{O}| = |\mathbb{N}|$ also. Composing such a bijection with that of Example 1.20 gives a bijection between $\mathbb{O}$ and $\mathbb{E}$ (by Proposition 1.3), so $|\mathbb{O}| = |\mathbb{E}|$. This is perhaps less surprising, since our intuition strongly suggests there should be equally many even and odd numbers. ◊

Our second example illustrates the opposite phenomenon: a countable set with *more* elements than $\mathbb{N}$ (in some sense, twice as many).

Example 1.21 (Countability of the integers)**.**
We claim that $|\mathbb{Z}| = |\mathbb{N}|$. The following diagram (Figure 1.6) shows how we prove this.

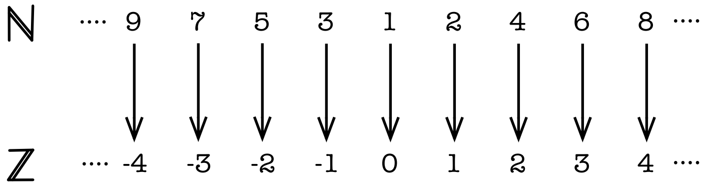


Figure 1.6.: Counting the integers

We use the even natural numbers to count the positive integers, and the odd ones to count the negative integers and 0. (We could equally well have decided to do this the other way round.) An explicit function is $f\colon \mathbb{N} \to \mathbb{Z}$ given by the rule:

$$f(1) = 0, \qquad f(2n) = n, \qquad f(2n+1) = -n,$$

for all $n \in \mathbb{N}$. We can formally check that f is a bijection in a similar way to Example 1.20. However, because the function rule comes in several pieces there are various cases to untangle.

- f is onto.

  Suppose $a \in \mathbb{Z}$. If $a > 0$ then $2a \in \mathbb{N}$ and $a = f(2a)$; if $a = 0$ then $a = f(1)$; and if $a < 0$ then $1 - 2a \in \mathbb{N}$ and $a = f(1 - 2a)$. So in all cases $a = f(m)$ for some $m \in \mathbb{N}$, which means that f is onto (Definition 1.3).

- f is one-to-one.

  We apply the "standard routine" (1.4). So, suppose there exist $p, q \in \mathbb{N}$ such that $f(p) = f(q) = a$, say. If $a = 0$ then $p = q = 1$. If $a > 0$ then p, q are both even and:

  $$p = 2a = q.$$

  If $a < 0$ then p, q are both odd and:

  $$p - 1 = -2a = q - 1,$$

  from which it follows that $p = q$ once again. Thus $p = q$ in all cases, as required □

Our final warm-up is distinctly "toastier": we show how to count a set that has in some sense "infinitely times as many" elements as $\mathbb{N}$.

Example 1.22 (Countability of the positive integer lattice)**.**

We claim that $|\mathbb{N} \times \mathbb{N}| = |\mathbb{N}|$, where $\mathbb{N} \times \mathbb{N}$ denotes the set of ordered pairs of natural numbers; ie. the Cartesian product (see Section 1.1.8):

$$\mathbb{N} \times \mathbb{N} = \{(a, b) : a, b \in \mathbb{N}\}.$$

Since $\mathbb{N}$ is a subset of $\mathbb{R}$, $\mathbb{N} \times \mathbb{N}$ is a subset of $\mathbb{R} \times \mathbb{R} = \mathbb{R}^2$, the Cartesian plane, so we may visualise this set geometrically (Figure 1.7):

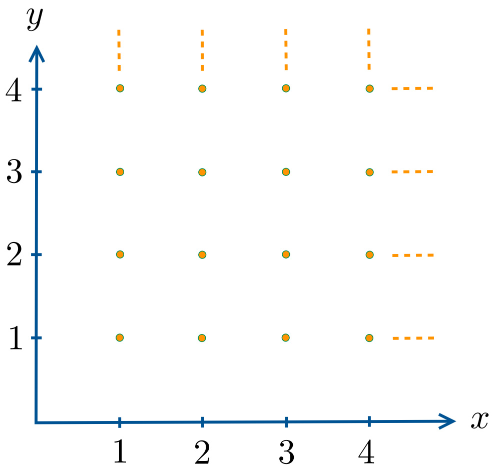


Figure 1.7.: Integer lattice

This configuration is sometimes referred to as the *(positive) integer lattice.* Think of it as the corners of the squares on an infinite chess (or draughts/chequers) board, if this helps. We can see from the diagram that each column (or row) contains a copy of $\mathbb{N}$, so the set $\mathbb{N} \times \mathbb{N}$ is "infinitely times bigger" than $\mathbb{N}$.

We define a bijection $f : \mathbb{N} \to \mathbb{N} \times \mathbb{N}$ by following the path through the lattice indicated in Figure 1.8.

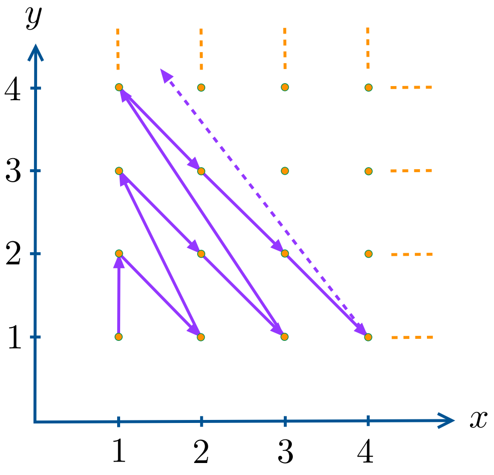


Figure 1.8.: Counting lattice points

So, the idea is to "count" the lattice points diagonally, rather than by rows or columns (which would mean never getting out of the first row or column!).

We organise the values of f into the various diagonals:

$$\begin{aligned} &f(1) = (1,1) \quad \cdots \text{ 1st diagonal} \\ &f(2) = (1,2), \quad f(3) = (2,1) \quad \cdots \text{ 2nd diagonal} \\ &f(4) = (1,3), \quad f(5) = (2,2), \quad f(6) = (3,1) \quad \cdots \text{ 3rd diagonal} \\ &f(7) = (1,4), \ \ldots \end{aligned}$$

To write down a formula for $f(n)$ we need to know those natural numbers s_k where $f(s_k)$ sits at the bottom (ie. lower right) of the k-th diagonal. Since there are precisely r points on the r-th diagonal, these are:

$$s_1 = 1, \quad s_2 = 1+2 = 3, \quad s_3 = 1+2+3 = 6, \ \ldots, \ s_k = 1 + \cdots + k = \tfrac{1}{2}k(k+1), \ \ldots$$

where we've used the formula for the sum of the first k natural numbers. (We'll take a closer look at this formula in Chapter 2; see Propositions 2.6) and 2.10.) The natural numbers that get mapped onto the k-th diagonal ($k \geqslant 2$) are therefore:

$$n = s_{k-1} + s, \quad s = 1, \ldots, k,$$

and their destinations are:

$$f(n) = f(s_{k-1} + s) = (s, k - s + 1).$$

(Note that the x- and y-coordinates sum to $k+1$, as they should.) This defines f.

In summary, we have constructed a function $f\colon \mathbb{N} \to \mathbb{N} \times \mathbb{N}$ whose rule is:

$$f(n) = (s, k - s + 1), \quad \text{for } n = s_{k-1} + s, \tag{1.10}$$

where $s_{k-1} = \frac{1}{2}k(k-1)$ and $s = 1, \ldots, k$.

Remark 1.23. At the risk of stating the obvious, this formula for f is quite complicated, and constructing it from Figure 1.8 was quite a hassle. This is not an uncommon experience when turning pictures into equations! Although our diagram makes everything we're trying to show seem "obvious" (and therefore suggests that our method has a high chance of success), we need the formula to nail down a proof. $\Diamond$

We now verify that f is a bijection.

- To show f is onto, if $(a, b) \in \mathbb{N} \times \mathbb{N}$ then comparison with (1.10) shows that $(a, b) = f(n)$ for some $n \in \mathbb{N}$ if and only if:

  $$s = a, \qquad k = a + b - 1.$$

  Therefore $(a, b) = f(n)$ where:

  $$n = a + \tfrac{1}{2}(a + b - 1)(a + b - 2).$$

- To show f is one-to-one we use the "standard routine". If $f(n) = f(m)$ then writing $m = s_{\ell-1} + t$ for $\ell \geqslant 2$ and $t = 1, \ldots, \ell$ it follows from (1.10) that:

  $$(s, k - s + 1) = (t, \ell - t + 1),$$

  which implies $s = t$ and $k = \ell$; thus $n = m$. $\square$

Examples 1.20, 1.21 and 1.22 already demonstrate the "more (or less), but just as many" kind of strangeness that arises when counting the elements of infinite sets.

We now turn to the proof of the main result. For this we will use a couple of lemmas, the first of which is a formalisation of Remark 1.21.

Lemma 1.6 (Countability and cardinality).
If A is countable and $|A| = |B|$ then B is countable.

Proof. Since A is countable there exists a bijection $f\colon S \to A$ where either $S = [n]$ or $S = \mathbb{N}$. Since $|A| = |B|$ there's also a bijection $g\colon A \to B$. Therefore $g \circ f\colon S \to B$ is a bijection, by Proposition 1.3, so B is countable. $\blacksquare$

Our second lemma may be regarded as a generalisation of both Example 1.20, Remark 1.22 and Proposition 1.5.

Lemma 1.7 (Subsets of countable sets)**.**
If A is countable and $B \subseteq A$ then B is countable.

Proof. We subdivide the proof into two parts: a special case (Part 1), followed by the general case (Part 2).

Part 1. We first show that the result is true if $A = \mathbb{N}$; ie. every subset of $\mathbb{N}$ is countable. So, suppose $B \subseteq \mathbb{N}$. The result is clearly true if B is finite. If B is infinite we construct a bijection with $\mathbb{N}$ as follows. Since the elements of B are natural numbers we can list them in order of increasing size, and label the n-th number on the list by b_n. More precisely, let b_1 be the smallest element of B, remove it from B and let b_2 be the smallest element of $B \setminus \{b_1\}$, then let b_3 be the smallest element of $B \setminus \{b_1, b_2\}$, etc. Thus:

$$B = \{b_1, b_2, b_3, \ldots\},$$

with $b_1 < b_2 < b_3 < \cdots$. We can now define a function $f \colon \mathbb{N} \to B$ by the rule $f(n) = b_n$. This function is a bijection: it's onto by definition, and one-to-one because it's strictly increasing (see Example 1.10). Hence $|B| = |\mathbb{N}|$.

Part 2. We now prove the general case. If A is finite then so is B (Proposition 1.5), and there is nothing further to prove. If A is countably infinite then by definition (Definition 1.10) there exists a bijection $G \colon A \to \mathbb{N}$. By applying G just to elements of B we obtain a "new" function $g \colon B \to G(B)$, where $G(B)$ is the image of B (see Section 1.2.2), with the same rule but different domain and codomain (see also Remark 1.13). (Recall from the Principle for Equality of Functions that two functions are equal only when they have the same domain, codomain and rule; so technically speaking g is different from G.) Now, g retains the property of being one-to-one, and because the codomain has been adjusted to contain *only* numbers of the form $G(b) = g(b)$ for elements $b \in B$, it's also onto (see Definition 1.3). Therefore g is a bijection. Since $G(B)$ is a subset of $\mathbb{N}$ it's countable, by the first part of the proof. Hence B is countable by Lemma 1.6. ■

Remark 1.24 (Well-ordering principle).
The proof of Lemma 1.7 made subtle use of the *well-ordering principle.* This is a fundamental property of the natural numbers that we tend to take for granted, which says: every non-empty subset of $\mathbb{N}$ has a least element (see Remark 4.2). ◊

We now prove our main result of this Set Piece.

Proof. [Set Piece Theorem 1]
Like our proof of Lemma 1.7, we will subdivide the proof into two parts.

Part 1. We first show that the set $\mathbb{Q}^+$ of all positive rational numbers is countable.

Every element of $\mathbb{Q}^+$ can be expressed uniquely as a/b where a, b are natural numbers with no common factors, and with this understanding we define a function $G \colon \mathbb{Q}^+ \to \mathbb{N} \times \mathbb{N}$ by:

$$G(a/b) = (a, b).$$

Then G is one-to-one; for, if $G(a/b) = G(c/d)$ then $(a, b) = (c, d)$, and equating x- and y-coordinates gives $a = c$ and $b = d$, so $a/b = c/d$. However, G is *not* onto; for example, the point $(2, 4) \in \mathbb{N} \times \mathbb{N}$ is not in the image of G (neither is any point (a, b) where a, b have a common factor). It's easy to fix this; we simply adjust the codomain of G to be the subset $Q = \operatorname{im}(G) \subset \mathbb{N} \times \mathbb{N}$ (cf. the proof of Lemma 1.7). This produces a "new" function $g\colon \mathbb{Q}^+ \to Q$, with the same domain and rule, but different codomain. It's still one-to-one, but is now onto by definition; hence g is a bijection. Hence $|\mathbb{Q}^+| = |Q|$ (Definition 1.9). Since Q is a subset of a countable set (by Example 1.22) it's countable by Lemma 1.7. Therefore $\mathbb{Q}^+$ is countable by Lemma 1.6.

Part 2. We now show that the entire set $\mathbb{Q}$ is countable. This is rather similar to the argument that we used to count the set $\mathbb{Z}$ in Example 1.21.

Since $\mathbb{Q}^+$ is countable there exists a bijection $F\colon \mathbb{N} \to \mathbb{Q}^+$. (Note that we don't need to construct F; it's sufficient to know that it exists!) Now define $f\colon \mathbb{N} \to \mathbb{Q}$ by:

$$f(1) = 0, \qquad f(2n) = F(n), \qquad f(2n+1) = -F(n),$$

for all $n \in \mathbb{N}$. It's easy to adapt the formal argument given in Example 1.21 to show that f is a bijection, and this completes the proof. ■

Remark 1.25. It is natural to regard the sets $\mathbb{E}$, $\mathbb{N}$, $\mathbb{Z}$ and $\mathbb{Q}$ as subsets of $\mathbb{R}$, and $\mathbb{N} \times \mathbb{N}$ as a subset of $\mathbb{R}^2$. When viewed this way, $\mathbb{Q}$ is rather different from the others, insomuch that every interval of $\mathbb{R}$ (see Example 1.2), no matter how small, always contains a rational number. *(Can you prove this? It's essentially because we can create fractions with arbitrarily large denominator.)* Thus rational numbers persist at all levels of magnification. This property is called *density.* By contrast, the other subsets are *discrete:* each of their points is surrounded by a certain amount of "empty space". It is therefore perhaps more surprising that $\mathbb{Q}$ turns out to be countable. ◊

2. Logic and Proof

References.
Liebeck: Chapter 1, pp. 3–7
Allenby: The whole book! Dip in, or use the index to find something specific.
Franklin and Daoud: Chapters 1–6, Chapter 8.
Episodes 5–8 of video lectures.

In mathematics, one of the main games we play is to deduce new statements from old ones using logical arguments. This is the process of *proof*. It is a key part of a mathematician's skill set to be able to construct and present coherent proofs, and to recognise when a proof is or isn't valid. We've already seen some examples of this. However in this chapter we're going to specifically collect together some ideas from logic and some common methods of proof. Of course, there's no "magic bullet" for proving theorems; if there was then we could all pack up and go home! Nevertheless, being aware of the main techniques puts us in a better position to formulate our own arguments. It's also worth pointing out that proofs are usually not unique! Very often there are several alternative (correct) proofs—in certain cases, multitudes—which raises the interesting question: "What makes a good proof?" We'll treat that question rhetorically!

2.1. Truth, falsehood and negation

It goes without saying that every (successful) mathematical proof depends on sound reasoning, which in turn relies on logic. The kind of logic most mathematicians use most of the time is called *first order logic,* or the *first order predicate calculus.* We will give a descriptive account, rather than attempting a fully rigorous exposition (for a deeper dive, Robert Stoll's book [35] is a great reference).

In logic, a *statement* is a sentence that is either true or false. We won't go into much detail about what we mean by a "sentence"—that's the preserve of formal logic—except to say it's more or less what we think it should be! In mathematics, sentences can of course be mixtures of words, symbols and expressions such as equations, inequalities, etc. Generally speaking, we denote statements by P, Q etc. Sometimes statements may depend on one or more *parameters* (or *variables*). For example, consider the following statement $P(x)$ about real numbers x (here x is the parameter):

$$x^2 - 3 > 0.$$

For some values of x, $P(x)$ is true (eg. when $x = 2$), and for some values it's false (eg. when $x = 1$); however, it is never true *and* false at the same time; ie. for the same x. This dichotomy is obvious, but also fundamental. In logic, it's called the "Law of the Excluded Middle"[1]. All of maths is built on the idea that we can prove things to be true or false, and that they can never be both true and false at the same time. That's why, typically, mathematical statements have to be very precise. For example, the sentence: "The sky is blue", although colloquially considered to be a statement, is so vague that it fails the Law of the Excluded Middle.

The *negation* of a statement P is the statement denoted $\neg P$ ("not P") which is true precisely when P is false, and false precisely when P is true. For the above statement $P(x)$ about a real number x, the negation $\neg P(x)$ is:

$$x^2 - 3 \leqslant 0.$$

However, for the pseudo-statement: "The sky is blue", its negation would be: "The sky is not blue", which again is neither true nor false.

2.2. Words to symbols and back again

Maths is full of shorthand. Unfortunately, many people confuse the use of shorthand with the act of writing mathematics. However, writing mathematics is *not* the business of creating reams of indecipherable hieroglyphics! Rather, notation should help us discern the underlying ideas, which are after all what we're really interested in.

Nevertheless, there is some shorthand that is so useful as to be generally accepted and widely used in mathematical writing (especially in lectures, solutions to exercises, exams, etc.). Even so, it's advisable to think very carefully about how and when to use these shorthands in formal written work, particularly if there's a danger of producing "symbol soup", or "mathematical minestrone"!

Here are some examples (we've already mentioned the first two in Preliminaries and Notation):

- $\forall$ stands for the words *for all*. Synonyms include: "for every"; "given". This is a symbol used in formal logic, called the *universal quantifier.*

- $\exists$ stands for the words *there exists*. Synonyms include: "there is"; "for some". This symbol from formal logic is called the *existential quantifier.*

- $P \wedge Q$ is the compound statement *P and Q*. It is true precisely when P and Q are *both* true.

[1] There is so-called "fuzzy logic", where this no longer holds. We won't go there!

- $P \vee Q$ is the compound statement *P or Q.* It is true precisely when *at least one* of P, Q is true (and sometimes referred to as the "inclusive or".)

The logical behaviour of "$\wedge$" and "$\vee$" can be summarised in a "truth table" (Table 2.1), which records the logical status of the compound statement for all possible combinations of the truth (T) or falsehood (F) of P and Q.

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

Table 2.1.: Truth table for "and" and "or"

- $P \Rightarrow Q$ stands for the phrase: *P implies Q*. Synonyms include: "if P then Q"; "P only if Q"; "P is sufficient for Q". The statement P is called the *premise,* and Q the *conclusion.* It's a compound statement, which is true when both P and Q are true, or when P is false: a false premise can imply a true conclusion! *(You may be able to think of an example, or see Example 2.5.)* So the implication $P \Rightarrow Q$ is false only when P is true and Q is false. As such, it's logically equivalent to the compound statement $(\neg P) \vee Q$. It's worth pausing to think about this; then take a look at Example 2.1 where we demonstrate the equivalence using a truth table. We write:

$$(P \Rightarrow Q) \equiv (\neg P) \vee Q,$$

  using the symbol "$\equiv$" for logical equivalence. This equivalence can be useful, if the statements P, Q are complicated, and in particular if we want to negate the implication.

- $P \Leftarrow Q$ stands for the phrase: *P is implied by Q*. Synonyms include: "if Q is true then so is P"; "P if Q"; "P is necessary for Q". It is the *converse* of the statement $P \Rightarrow Q$.

- $P \Leftrightarrow Q$ stands for the phrase: *P is equivalent to Q*. Synonyms include: "P if and only if Q" (which is often abbreviated to "P iff Q"); "P implies and is implied by Q"; "P is necessary and sufficient for Q". As a compound statement it's true precisely when P, Q are both true, or both false.

Mathematical sentences can be translated into logical symbols, and vice versa. This can be quite useful, especially when trying to formulate a precise a mathematical statement, or negate a statement. In this regard, the negation of $\forall$ is $\exists$, and vice versa. To be precise:

$$\neg((\forall x)P(x)) \equiv (\exists x)\,\neg P(x),$$

the left hand side of which can be read as:

> *It is not the case that $P(x)$ is true for all x,*

and whose right hand side reads:

> *There exists an x such that $P(x)$ is false.*

Similarly:

$$\neg((\exists x)P(x)) \equiv (\forall x)\,\neg P(x),$$

the left hand side of which reads:

> *There exists no x such that $P(x)$ is true,*

and whose right hand side reads:

> *For all x, $P(x)$ is false.*

These equivalences should be intuitively clear.

The negation of an "or" statement is an "and" statement, and vice versa. To be precise:

$$\neg(P \wedge Q) \equiv (\neg P) \vee (\neg Q), \qquad \neg(P \vee Q) \equiv (\neg P) \wedge (\neg Q).$$

Again, these should be intuitively clear, and it's worth pausing to be convinced of this! (Try inserting examples of statements P and Q.) It's also worth creating a "truth table" (see Example 2.1), which provides a formal verification.

In practice, we don't often use the symbols $\wedge$ and $\vee$ when writing mathematics (apart from formal logic), since the linguistic equivalents "and" and "or" are almost as efficient. Moreover, sometimes there's a danger of "over-symbolising" a piece of written mathematics, and we need to "let it breathe".

Example 2.1 (Truth table for "implies")**.**
We check the logical equivalence:

$$P \Rightarrow Q \equiv (\neg P) \vee Q,$$

by means of a truth table (Table 2.2; cf. Table 2.1). Again, the idea is to record how the logical status ("true" or "false") of the two compound statements depends on that of P and Q, and show it's the same in both cases.

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

Table 2.2.: Truth table for "implies"

We see that the two relevant (green) columns are identical, so the equivalence is proved.

This way of thinking about $P \Rightarrow Q$ can be helpful if we need to negate it. Using the rule for negating "$\vee$" we get:

$$\neg(P \Rightarrow Q) \equiv P \wedge (\neg Q). \tag{2.1}$$

For example, suppose we want to determine whether the following rather daunting statement (which it's worth trying to put into words) is true:

$$(\forall x \in \mathbb{R})\ |x| < 1 \implies |\sin(1/x)| < 1.$$

By (2.1) and the rule for negating the universal quantifier, its negation is the statement:

$$(\exists x \in \mathbb{R})\ |x| < 1 \text{ and } |\sin(1/x)| \geqslant 1.$$

Putting this into words, we see that it is true; take $x = 2/\pi$ for example. Therefore the original statement is in fact false. □

Example 2.2 (Negation of a quantified statement)**.**
The statement:

Every real number has a square root,

may be formalised as follows:

$$(\forall x \in \mathbb{R})(\exists y \in \mathbb{R})\, x = y^2.$$

Note the use of brackets to aid legibility. Nevertheless, it is very concise, and we can relax it a bit by inserting some words:

$$(\forall x \in \mathbb{R})(\exists y \in \mathbb{R}) \text{ such that } x = y^2.$$

To negate the statement we apply the "$\neg$" operator from the left and work our way through, negating each clause in turn:

$$(\exists x \in \mathbb{R})(\forall y \in \mathbb{R})\, x \neq y^2.$$

In this case the negated statement is true (take $x = -1$ for example); the original statement was of course false! □

Example 2.3 (Quantifying over different parameters)**.**
Quantifiers ($\forall$ and $\exists$) act on statements with parameters, to produce new statements that have fewer or no parameters, and which may have a different logical status. For example, if $P(x)$ is the statement $x^2 - 1 \geqslant 0$ then we may quantify it in various ways, such as:

$$(\forall x \in \mathbb{R})P(x), \qquad (\exists x \in \mathbb{R})P(x),$$

the first of which is false and the second true. Or:

$$(\forall x \in \mathbb{N})P(x), \qquad (\exists x \in \mathbb{N})P(x),$$

both of which are true. □

Example 2.4 (Precise definition of infinite sets)**.**
The condition for a non-empty set A to be finite (Definition 1.8) can be expressed using quantifier notation as follows:

$$(\exists n \in \mathbb{N})(\exists \text{ function } f_A \colon [n] \to A)\ f_A \text{ is a bijection.}$$

The negation of this statement can be obtained by applying the rules for negating quantifiers, giving us the following:

$$(\forall n \in \mathbb{N})(\forall \text{ functions } f \colon [n] \to A)\ f \text{ is not a bijection.}$$

When translated into words this reads:

> *There is no bijection between A and $[n]$, for any natural number n.*

Since A is infinite precisely when it isn't finite (Definition 1.8 again), this is a tangible characterisation of infinite sets, which can be put to practical use. □

Example 2.5 (False premise with true conclusion)**.**
To show what we mean by "a false premise can (correctly) imply a true conclusion", here's a statement: "If I want to go to Mars then I need a spaceship". This implication is true; if the premise is true (which in this case is highly questionable!) then so is the conclusion. However if the premise is false, meaning "I do not want to go to Mars" is true, then the conclusion "I need a spaceship" may still be true; for example, if I want to go to the Moon! □

2.3. Proof versus counter-example

In mathematics, as well as proving certain things are true, we spend a lot of time showing that other things are not true. This is partly in order to get a feel for a particular problem, and to avoid pitfalls when it comes to formulating statements that we feel are likely to become theorems. Luckily, in many ways disproving things is much easier: in order to show a statement is false we simply need to find a *single* concrete instance where it breaks down. This is known as a *counter-example*.

It's all too easy to forget the essential simplicity of this, and instead over-complicate the issue by appealing to abstract formulas and general equations without really pinning anything down. Since this is so important, here's an illustration (see also Example 1.15).

Question. Is it true that:

$$(x+y)^2 = x^2 + y^2$$

for all real numbers x and y?

Naive answer. Well:

$$(x+y)^2 = x^2 + 2xy + y^2 \neq x^2 + y^2,$$

since $2xy \neq 0$. So the statement is false.

Mathematician's answer. Let $x = y = 1$. Then $(1+1)^2 = 4$, whereas $1^2 + 1^2 = 2$. Since $2 \neq 4$, the statement is false.

One problem with the naive answer is that the given equation is in fact *true* for many (in fact infinitely many) choices of x and y — for example, we can choose any value of x and set $y = 0$ — so the statement "$2xy \neq 0$" is not even correct! *(Can you see any other problems with the naive answer?)*

2.4. Methods of proof

As the degree progresses the same methods of proof tend to be used over and over again. After a while, they begin to become recognisable, and it's possible to anticipate how the arguments will go before even seeing all the detail. It could be said, slightly poetically, that proofs have a rhythm of their own—if we wind them up correctly and set them off, they will often unwind themselves like clockwork, with only a minimal amount of effort required, particularly proofs of a "routine" nature. Having said that, finding an idea or set of ideas on which to base a successful proof can often involve significant amounts of insight, creativity, inspiration and, occasionally, deviousness! And it's only through familiarity with "bread and butter" mathematical proofs that we come to appreciate those that are genuinely inspired.

Here are some common styles of proof.

2.4.1. Direct proof

Does exactly what it says on the tin: proceed directly from premise P to conclusion Q via a sequence of intermediate logical steps:

$$P \implies P_1 \implies P_2 \implies \cdots \implies Q.$$

All the proofs we've seen so far have been direct, and here's another.

Proposition 2.1 (Cancellation law for integer multiplication)**.**
Suppose $a, b, c \in \mathbb{Z}$ with $c \neq 0$. If $ac = bc$ then $a = b$.

Proof. To prove this it's tempting to "divide through by c". This is possible if we view $\mathbb{Z}$ as a subset of $\mathbb{Q}$, where division is a valid operation. However, there is an alternative "self-contained" argument. We simply rewrite the equation as:

$$(a - b)c = 0,$$

from which it follows that $a - b = 0$ or $c = 0$. Since we've assumed $c \neq 0$ it follows that $a - b = 0$, hence $a = b$. ■

Remark 2.1*.* The fact that Proposition 2.1 doesn't require the use of division turns out to be very important when we approach the task of rigorously constructing the rational numbers $\mathbb{Q}$ from $\mathbb{Z}$, which we will undertake in Set Piece 3. In effect, it prevents a circular argument. It also becomes highly significant when we attempt to create an abstract algebraic theory that generalises the familiar arithmetic of addition and multiplication in $\mathbb{Z}$: this is the area of modern algebra called "ring theory". There will be much more about this in Year 2. ◊

The route from P to Q may often be some sort of calculation, and the proof has to provide enough detail to convince us that the calculation is correct, whilst not getting bogged down in its minutiae. An interesting (and rather complicated) example of this is the proof of Proposition A.9 in Appendix A.6.

2.4.2. Contrapositive

If we want to prove something of the form $P \Rightarrow Q$, it is sometimes easier to convert this into contrapositive form. The *contrapositive* is the logically equivalent statement:

$$\neg Q \Rightarrow \neg P.$$

To get a feel for this, it's worth pausing to think of an example; or see Proposition 2.2 below.

Here's a proof using formal logic:

$$\neg Q \Rightarrow \neg P \equiv \neg(\neg Q) \vee (\neg P), \quad \text{by Example 2.1}$$

$$
\begin{aligned}
&\equiv Q \vee (\neg P)\\
&\equiv (\neg P) \vee Q\\
&\equiv P \Rightarrow Q, \quad \text{by Example 2.1 again.}
\end{aligned}
$$

The power of the contrapositive is not to be underestimated. As a first example we will use it to prove the following result: if a square integer is even then so is its square root.

Proposition 2.2 (Even squares)**.**
For all $n \in \mathbb{N}$, if n^2 is even then n is even.

Proof. The contrapositive of the implication part of the statement is:

> *If n is odd then n^2 is odd.*

The proof of this is straightforward. Suppose $n = 2m - 1$ where $m \in \mathbb{N}$. Then:

$$n^2 = (2m-1)^2 = 4m^2 - 4m + 1 = 2(2m^2 - 2m) + 1,$$

which is odd. ■

Remarks 2.2.

1) It's easy to see that the converse of Proposition 2.2 is also true; ie. if n is even then n^2 is even. So this is in fact an "if and only if" result (see Section 2.5). In particular, we can immediately deduce that the square root of an odd square integer is odd, saving us from having to prove this from scratch.
2) Proposition 2.2 is valid for higher powers:

   > *For all $n, k \in \mathbb{N}$, if n^k is even then n is even.*

   The method of proof is identical, with the expansion of $(2m-1)^k$ being achieved using the Binomial Theorem. ♢

Example 2.6 (One-to-one functions)**.**
The "standard routine" (1.4) for showing that a function $f\colon A \to B$ is one-to-one is in fact the contrapositive version of the defining statement (1.3). Although these statements are logically equivalent, (1.3) better captures the intuitive idea of the concept, whereas (1.4) is more useful in practice. □

Example 2.6 suggests a wider category of proofs that are well-suited to contrapositive formulation: uniqueness proofs. Suppose $P(x)$ is a property on a set X, and we want to show that there is at most one element $a \in X$ such that $P(a)$ is true. We would therefore want to prove: if $b \neq a$ then $P(b)$ is false. (Note the similarity between this and the condition for a function to be one-to-one.) The contrapositive version is: if $P(b)$ is true then $b = a$. This is usually much easier to deal with, not least because it involves an equation ($b = a$) rather than a non-equality ($b \neq a$). Here's an example.

Proposition 2.3 (Uniqueness of the cube root)**.**
The cube root of a real number x is unique.

Proof. The contrapositive of this is: if $x = a^3$ and $x = b^3$ then $a = b$.

We begin by eliminating x to obtain $a^3 = b^3$, and then factorise:

$$0 = a^3 - b^3 = (a - b)(a^2 + ab + b^2).$$

Hence $a - b = 0$ or $a^2 + ab + b^2 = 0$. Now the signs of a, b are the same as that of x; so a and b have the same sign. This implies $ab \geqslant 0$, hence $a^2 + ab + b^2 \geqslant 0$, with equality only possible if $a = b = 0$. We therefore conclude that in all cases $a = b$. ■

Remark 2.3. Another way of stating Proposition 2.3 is that the function $\mathbb{R} \to \mathbb{R}; x \mapsto x^3$ is one-to-one. ◊

2.4.3. Proof by exhaustion

Sometimes a statement naturally breaks up into lots of smaller cases; for example, often statements about integers fragment into cases depending on whether the numbers are even or odd, composite or prime, positive or negative, etc. It may be that to prove the result we're after there is nothing for it but to exhaust all of these cases one by one, hence "proof by exhaustion[2]". The most famous (or perhaps infamous) example of this may well be the proof of the "Four Colour Theorem" [5], which involved the analysis of so many cases that it could only be done by computer! However, thankfully, the number of cases is usually much more manageable; a more typical example, which is still nevertheless slightly tedious, is the proof of Theorem 4.16 below. In Example 1.5 we needed to consider three cases, depending on whether an integer is positive, negative or zero, and these same three cases were considered as part of the formal proof that $|\mathbb{Z}| = |\mathbb{N}|$ in Example 1.21 during Set Piece 1. Sometimes, as in the proof of Theorem 2.4 below, there are just two cases to consider; further examples include the proofs of Proposition 2.9 in Section 2.4.5 and Theorem 4.8 in Chapter 4, and a small part of the proof of Proposition 2.12 in Set Piece 2.

There is a danger of a proof by exhaustion falling into the "brute force and ignorance" category, and becoming "exhausting" in more ways than one! However, a well formulated proof of this type will usually find a way of organising itself to minimise the number of cases or amount of repetition; for example, by showing that settling one particular case first of all makes the others easier to resolve.

[2] Not to be confused with the "method of exhaustion" used by ancient Greek mathematicians (such as Archimedes) for finding areas or volumes of curved shapes by polygonal approximation; the precursor of integral calculus.

As an example we describe a famous "classical" result, concerning the number of prime numbers, which dates back to ancient Greece and certainly deserves the title "theorem". It appears in Book 9 of Euclid's "Elements" (Proposition 20) [15], a 13 volume encyclopædic survey of all the mathematics known to Greek mathematicians at the time (circa 300 BC). Although compiled by Euclid, the Elements undoubtedly contains material that had been in circulation for several centuries. However there is reason to believe that the proof of this particular theorem is Euclid's own. Our version of the proof incorporates some of the ideas that we've developed so far, and therefore has a more modern "feel" to it, but the basic idea is the same. The proof makes use of another famous result concerning primes, that we borrow from Section 2.4.5 below (Proposition 2.9), where we also remind ourselves of the definition of prime numbers (Definition 2.1). We denote the set of all prime numbers by $\mathbb{P}$.

Theorem 2.4 (Infinitude of primes).
The set $\mathbb{P}$ is infinite; ie. there are infinitely many prime numbers.

Proof. We will verify that $\mathbb{P}$ satisfies the condition for infinite sets laid out in Example 2.4; ie. there is no bijection between $\mathbb{P}$ and $[n]$ for any $n \in \mathbb{N}$.

Suppose $f : [n] \to \mathbb{P}$ is any function. For notational convenience we denote the values of f by:

$$f(1) = p_1, \ldots, f(n) = p_n.$$

We claim that f cannot be onto, and is therefore not a bijection.

Let N be the following natural number:

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

There are two possibilities.

Case 1. N is prime. Then $N \neq p_1, \ldots, p_n$ (because $N > p_1, \ldots, p_n$). So N is an element of $\mathbb{P}$ that doesn't belong to the image of f.

Case 2. N is not prime. Then N has a prime factor p by Proposition 2.9 below. But $p \neq p_1, \ldots, p_n$, because the only common factor of N and $N - 1$ is 1 (since any common factor must also divide $N - (N - 1) = 1$). Therefore p is an element of $\mathbb{P}$ that doesn't belong to the image of f. ■

Remarks 2.4.

1) It follows from Theorem 2.4 that the set $\mathbb{N}$ is also infinite, by the contrapositive version of Proposition 1.5. This is of course a property that we take for granted (and indeed have already done so, in Set Piece 1). Then, being a subset of a countable set, it follows from Lemma 1.7 that $\mathbb{P}$ is countably infinite.

2) The proof of Theorem 2.4 relies on another property that we take for granted: each natural number m has a "successor" $m+1$, which is unique and not equal to any of $1, \ldots, m$. It's one of a handful of essential properties of natural numbers that may be used as axioms for defining the set $\mathbb{N}$: the "Peano axioms". We will see more of these in Chapter 4 (Section 4.1).

3) Theorem 2.4 implies that there is no greatest prime number. For, if p was the greatest prime then $\mathbb{P} \subseteq [p]$, which would imply that $\mathbb{P}$ is a finite set by Proposition 1.5. This very simple argument foreshadows our next method of proof: proof by contradiction (Section 2.4.4).

4) A common misconception arising from the proof of Theorem 2.4 is that it implies the following statement:

   If $p_1, \ldots, p_n$ are primes then $N = 1 + p_1 \cdots p_n$ is a new prime.

   However, this is only Case 1 of the proof, and can easily fail, as shown by the counterexample $p_1 = 3$, $p_2 = 5$. What *is* true is that if $p_1, \ldots, p_n$ are *successive* primes, starting from $p_1 = 2$, then N is a new prime; for, if $p_1, \ldots, p_n$ are the *only* primes less than N then N can't have any proper prime factors, by the argument given for Case 2. However, even here N may not be the *least* prime greater than $p_1, \ldots, p_n$; for example if $p_1 = 2$ and $p_2 = 3$ then $N = 7$, missing out the prime number 5. ◊

Remark 2.5 (Greatest known prime).
The previous remark addresses an often-asked question: "If there are infinitely many prime numbers, how can it be that there is a greatest *known* prime?" For example, as of December 2018 the greatest known prime is:

$$p_{\text{max}} = 2^{82,589,933} - 1,$$

a number with a mere $24,862,048$ decimal digits. This seems to fly in the face of Theorem 2.4, which guarantees that there are infinitely many primes greater than p_{max}.

The fundamental problem is that it's extremely difficult to determine in practice which of the infinitely many natural numbers greater than p_{max} are prime! One reason for this is that, according to the Prime Number Theorem[3] [42], the proportion of prime numbers amongst all integers with up to d decimal digits is approximately:

$$\frac{1}{d\ln(10)}.$$

So, as we go further out into the natural numbers the probability of randomly encountering primes decreases. To get some feel for this, a simple calculation (using a calculator!) shows that for $d = 25,000,000$ it's less than 1 in 57 million, decreasing even further as the

[3] The Prime Number Theorem was proved, independently, in 1896 by French mathematician Jaques Hadamard (1865–1963) and Belgian mathematician Charles-Jean de la Vallée Poussin (1866–1962), although it had been "in the air" for the preceding hundred years.

number of digits increases. Now, couple this with the fact that even with modern computing technology, and optimised algorithms, it takes many weeks for a single number of the order of magnitude of $p_{\max}$ to be tested for primality. The reason why computational effort was expended on $p_{\max}$ is because of its rather special nature: it's a "Mersenne number"; ie. a power of 2, minus 1. This facilitates the testing regime (for example, the Lucas-Lehmer test; see [6]). And, even though there are only 51 known Mersenne primes (including $p_{\max}$), this is enough to improve the odds of a Mersenne number being prime.

It's also worth pointing out that although we know $p_{\max}$ is prime, we don't know all the primes less than $p_{\max}$, so we can't simply generate a new prime by adding 1 to their product. ◊

2.4.4. Proof by contradiction

Also known by the Latin phrase *reductio ad absurdum,* some of the most famous "classical" proofs use this technique, and it remains a very popular and effective strategy. The idea dates back to the ancient Greeks (and possibly even further), and is both logically compelling and delightfully devious. It goes like this. In order to prove that a statement P is true we begin by assuming it isn't; in other words, we assume that the negated statement $\neg P$ is true. We then argue our way to something that is obviously false, or that contradicts one of our other assumptions. Given that our reasoning is correct, since a true premise cannot imply a false conclusion (see Table 2.2) the only way to resolve this contradiction (or "absurdity") is by realising that the assumption "$\neg P$ is true" must be false, and therefore by the Law of the Excluded Middle that P is true!

The best known application of proof by contradiction is probably the following result, whose fame guarantees it full "theorem" status!

Theorem 2.5 (Irrationality of the square root of 2)**.**
The real number $\sqrt{2}$ is irrational; ie. there is no rational number whose square is 2.

Proof. Suppose to the contrary that $\sqrt{2}$ is rational. Thus $\sqrt{2} = a/b$ where $a, b \in \mathbb{N}$. By cancelling any common factors of 2, we may assume that at least one of a, b is odd. Now, by squaring and rearranging the equation for $\sqrt{2}$ we obtain:

$$a^2 = 2b^2. \tag{2.2}$$

This shows a^2 is even, hence a is even (Proposition 2.2); therefore b must be odd. Since $a = 2c$ for some $c \in \mathbb{N}$ equation (2.2) becomes:

$$4c^2 = 2b^2,$$

which after cancelling a factor of 2 shows b^2 is even, hence b is even. We have arrived at a contradiction (the existence of a natural number that is both even and odd), and since the reasoning that led to it is squeaky clean our hypothesis (that $\sqrt{2}$ is rational) must be false; so $\sqrt{2}$ is irrational. ■

Remarks *2.6.*

1) The first written record of the proof of Theorem 2.5 is found in Book 10 of Euclid's "Elements" [15]. However the result itself probably goes back considerably earlier, to the school of Pythagoras, circa 500 BC [26, 30]. It has profound ramifications for the measurement of length and area in Euclidean geometry (there's an interesting discussion of this in [20]), and its long period of dormancy may in fact have been a "cover up", possibly involving an assassination!

2) In Remark 2.2 (2) we noted that Proposition 2.2 generalises to higher powers of 2. This means that Theorem 2.5 also generalises, to:

   For all $k \in \mathbb{N}$ the real number $2^{1/k}$ is irrational.

   The proof is almost identical, but nevertheless worth briefly running through to check that everything works.

3) The statement of Theorem 2.5 implicitly assumes the existence of $\sqrt{2}$ as a real number. However, it does not prove this; rather, that if it does exist then it can't be rational. The existence of real numbers such as $\sqrt{2}$, or more generally $2^{1/k}$, is a much more subtle question, which has to be carefully considered when defining the set $\mathbb{R}$. $\Diamond$

Theorem 2.5 is an example of a "non-existence result": the non-existence of a rational number whose square is 2. Such results are well-suited to proof by contradiction, because assuming that a non-existence statement is false immediately yields an existence statement, and this gives us something tangible to work with. In fact, what is arguably the greatest non-existence result of them all—Fermat's Last Theorem (see Example 1.19)—was proved in this way, albeit within a highly sophisticated mathematical arena well beyond the scope of anything in these notes. Instead, in the following example we show how proof by contradiction helps to get us started on the journey.

Example 2.7 (First steps towards Fermat's Last Theorem)**.**

Fermat himself is given credit for proving the theorem when $n = 4$; that is, the Fermat equation:

$$x^n + y^n = z^n$$

has no solutions $x, y, z \in \mathbb{N}$ if $n = 4$. It follows from this that the theorem also holds when n is any multiple of 4. For, suppose not. Then Fermat's equation has a solution for some $n = 4m$. Elementary rules of exponentiation allow us to write this as:

$$(x^m)^4 + (y^m)^4 = (z^m)^4.$$

But then x^m, y^m, z^m is a solution of Fermat's equation when $n = 4$, contradicting Fermat's original result. In particular, this shows that Fermat's Last Theorem is true whenever the exponent n is a power of 2 (other than 2).

A similar argument shows that if Fermat's Last Theorem holds for a prime exponent $p \geqslant 3$ then it also holds for any exponent n that is divisible by p. Since the only natural numbers not to have prime factors $p \geqslant 3$ are powers of 2, this shows that it suffices to prove Fermat's Last Theorem for prime exponents $n = 3, 5, 7, \ldots$. The fact that the sequence of primes has no known pattern is one reason why this is difficult! □

2.4.5. Proof by induction

The *Principle of Mathematical Induction* is actually axiomatic in our formulation of mathematics; we will take a closer look at it in Chapter 4 (Section 4.1). Its use in proof by induction goes as follows.

Suppose we have a sequence of mathematical statements $P(n)$ indexed by natural numbers $n \in \mathbb{N}$. Then we can deduce the truth of $P(n)$ for all sufficiently large n in two steps:

- First prove a *base case.* That is, prove $P(n_0)$ is true for some natural number n_0; most often, but not always, $n_0 = 1$.
- Now prove the *induction step.* That is, prove that if $P(n)$ is true (this assumption is called the *induction hypothesis*) then $P(n+1)$ is also true.

The Principle of Mathematical Induction then allows us to conclude that $P(n)$ is true for all $n \geqslant n_0$.

Even without any formal knowledge of the Principle of Induction, the rationale for this "machine-like" method of proof seems intuitively compelling. Having said that, if the machine is not set up correctly the method collapses like a pack of cards. So we're going to be quite careful when going through proofs by induction, writing out inductive arguments in full to ensure nothing has been overlooked.

Here's our first example. At the age of 10, so the story goes, the young Gauss[4] found a way to quickly sum the first 100 natural numbers, using a method that can easily be generalised to provide a formula for the sum of the first n natural numbers for any natural number n. We will look at Gauss's method for deriving the formula in Section 2.4.6 (Proposition 2.10). In the meantime we use proof by induction to check that the formula is true.

Proposition 2.6 (Gauss' summation formula; proof by induction)**.**
For all $n \in \mathbb{N}$ we have:

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

Note. This is our first use of the "sigma summation" symbol. We've written it out in full simply as a reminder of what it means! ◊

[4] Carl Friedrich Gauss (1777–1855): German mathematician, who was prolific in just about every area of contemporaneous mathematics and physics. Probably the greatest mathematician of all time.

Proof. Let $P(n)$ be the statement:

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

The base case is $P(1)$, which is true because:

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

For the induction step consider:

$$\begin{aligned}\sum_{i=1}^{n+1} i &= (n+1) + \sum_{i=1}^{n} i \\ &= (n+1) + \frac{1}{2}n(n+1), \quad \text{by the induction hypothesis} \\ &= \frac{1}{2}(n+1)(n+2), \quad \text{by elementary algebra.}\end{aligned}$$

This is the statement $P(n+1)$. It therefore follows from the Principle of Induction that $P(n)$ is true for all $n \geqslant 1$. ■

Our second example may look unfamiliar at first sight, until we realise that the integers in question are simply those whose decimal representation is a string of 9s.

Proposition 2.7 (Divisibility by nine)**.**
For all $n \in \mathbb{N}$, every integer of the form $10^n - 1$ is divisible by 9.

Proof. Let $P(n)$ be the statement:

$10^n - 1$ *is divisible by* 9.

Then $P(1)$ is (clearly) true. For the induction step consider:

$$\begin{aligned}10^{n+1} - 1 &= 10(10^n - 1) + 10 - 1 \\ &= 10(10^n - 1) + 9 \\ &= 10(9k) + 9,\end{aligned}$$

for some integer k, by the induction hypothesis. Hence $P(n+1)$ is true. The Principle of Induction therefore implies $P(n)$ is true for all $n \geqslant 1$. ■

Remarks *2.7.*

1) Proof by induction is by no means the only way to prove Proposition 2.7. It can also be proved directly using the formula for the sum of a geometric progression (which of course also requires proof!), and (even better) using modular arithmetic (see Proposition 3.5).

2) The numbers 10 and 9 in Proposition 2.7 are nothing special; the proof can be adapted for any pair of integers differing by 1. *(See whether you can do this; it's not difficult.)* Proposition 3.5 describes a different approach.

3) At the end of Set Piece 4 we will encounter "Fermat's Little Theorem" (Theorem **??**), which has a similar flavour. "Euler's Totient Theorem" (which we don't cover in these notes) is a further generalisation. ◊

In Remark 1.5 back in Chapter 1 we commented on the number of elements in the power set $\mathscr{P}(A)$ of a finite set A. If $|A| = n$ then by experimenting with small values of n it's not hard to come up with the "guess" that $|\mathscr{P}(A)| = 2^n$. Let's now prove this, by induction.

Proposition 2.8 (Cardinality of finite power sets)**.**
For all $n \in \mathbb{N}$, if $|A| = n$ then $|\mathscr{P}(A)| = 2^n$.

Proof. As advertised, we use induction on n. Although the proof of the induction step is in principle very simple, we have "sweated the detail" in order to tie it in with various ideas introduced in Chapter 1.

Let $P(n)$ be the statement:

If $|A| = n$ then $|\mathscr{P}(A)| = 2^n$.

If $n = 1$ then $A = \{a\}$ for some element a and we have:

$$\mathscr{P}(A) = \mathscr{P}\{a\} = \{\{a\}, \emptyset\},$$

so $|\mathscr{P}(A)| = 2 = 2^1$. Thus $P(1)$ is true.

Assume now that $P(n)$ is true, and suppose $|A| = n + 1$. Pick an element $a \in A$ and define:

$$\mathscr{P}' = \{B \subseteq A \mid a \notin B\}, \qquad \mathscr{P}'' = \{C \subseteq A \mid a \in C\}.$$

Since every subset of A either contains a or doesn't contain a, and these two possibilities are mutually exclusive, we have:

$$\mathscr{P}(A) = \mathscr{P}' \cup \mathscr{P}'', \qquad \mathscr{P}' \cap \mathscr{P}'' = \emptyset.$$

In other words, $\mathscr{P}(A)$ is the disjoint union of $\mathscr{P}'$ and $\mathscr{P}''$, hence:

$$|\mathscr{P}(A)| = |\mathscr{P}'| + |\mathscr{P}''|.$$

Now $\mathscr{P}' = \mathscr{P}(A')$ where $A' = A \smallsetminus \{a\}$, and since $|A'| = n$ it follows from the induction hypothesis that $|\mathscr{P}'| = 2^n$. To determine $|\mathscr{P}''|$, define functions $f\colon \mathscr{P}' \to \mathscr{P}''$ and $g\colon \mathscr{P}'' \to \mathscr{P}'$ by:

$$f(B) = B \cup \{a\}, \qquad g(C) = C \smallsetminus \{a\},$$

for all $B \in \mathscr{P}'$ and $C \in \mathscr{P}''$. Then $g \circ f(B) = B$ and $f \circ g(C) = C$, so f is invertible (Definition 1.5), hence a bijection (Proposition 1.2), which implies $|\mathscr{P}''| = |\mathscr{P}'|$ (Definition 1.9). Therefore:

$$\mathscr{P}(A) = 2^n + 2^n = 2.2^n = 2^{n+1},$$

which shows that $P(n+1)$ is true, completing the induction step. The result now follows from the Priniciple of Induction. ■

***Remark* 2.8.** Proposition 2.8 gives rise to the following alternative notation for power sets:

$$\mathscr{P}(A) = 2^A,$$

which is used regardless of whether or not A is finite. Although it's rather strange to see a set represented in this way, in the finite case it results in the rather nice formula:

$$|2^A| = 2^{|A|}. \qquad \lozenge$$

In order to prove the induction step it's occasionally necessary to "strengthen" the induction hypothesis by assuming all of $P(n_0), \ldots, P(n)$ are true. This doesn't affect the logical validity of the overall proof. We call this variant *strong mathematical induction.* As an example we prove a well-known result in elementary number theory. This requires the notion of prime numbers, for which the following definition serves as a reminder.

Definition 2.1 (Prime numbers)**.**
A positive integer p is *prime* if $p \geqslant 2$ and the only way of writing p as a product of natural numbers is: $p = 1 \times p$. Another way of saying this is that p has *no proper factors.* ⧫

***Remark* 2.9** (The natural number 1 isn't prime).
The natural number 1 is deliberately excluded from the family of primes. This is simply because when the prime numbers fulfil their primary (!) purpose as the building blocks for all other natural numbers (as we are about to show) then 1 is surplus to requirements; in fact, it would make the statements of many results unnecessarily complicated! ◊

Proposition 2.9 (Prime factorisation)**.**
Every natural number $2, 3, 4, \ldots$ is either prime, or can be written as a product of primes.

Proof. Whilst this is predominantly a proof by induction, the induction step contains a case-by-case argument (although, admittedly, only two cases).

Let $P(n)$ be the statement:

$n = p_1 \cdots p_r$ *where* $p_1, \ldots, p_r$ *are primes,* $r \geqslant 1$.

The base case is $P(2)$, which is true since 2 is prime; so we simply take $p_1 = 2$.

For the induction step, first note that if $n+1$ is prime then $P(n+1)$ holds with $p_1 = n+1$. (This was the first case!) Otherwise (and this is the second case), $n + 1 = ab$ where $2 \leqslant a, b \leqslant n$. By the strong induction hypothesis $P(a)$ and $P(b)$ are *both* true. So there exist primes $p_1, \dots, p_r$ and $q_1, \dots, q_s$ such that:

$$a = p_1 \cdots p_r, \qquad b = q_1 \cdots q_s.$$

Hence:

$$n + 1 = ab = p_1 \cdots p_r q_1 \cdots q_s,$$

a product of primes, which verifies $P(n + 1)$. The result now follows from the Principle of Induction. ■

Remark 2.10. Proposition 2.9 is the first part of the *Fundamental Theorem of Arithmetic,* which goes on to say that the prime factorisation of a natural number is unique. We will complete the proof of this theorem in Appendix A.3. ◊

The proof of Proposition 2.9 is remarkable in that it quickly and painlessly establishes a fundamental property of prime numbers, without requiring a complete list of all primes, and without any practical procedure for determining whether or not any given natural number is prime! In fact, these are both exceedingly difficult problems. Furthermore, the proof doesn't provide a practical method for factorising a natural number into primes! It's an example of what we would call a "non-constructive" proof. Many proofs by induction are like this.

Another feature of proof by induction is that, although very efficient, it requires a "prebaked" formula or statement to work with. As such, it may not provide much insight into the statement's "pedigree". Or, put another way:

If we have no idea why a statement is true, we can still prove it by induction![5]

By contrast, before Proposition 2.6 we mentioned that Gauss used a different argument to prove his summation formula, which unlike the inductive proof shows very clearly where the formula comes from. In fact Gauss' method is an illustration of our final proof technique.

2.4.6. Proof by counting

Sometimes we can prove a result by counting something in a particular way, or in different ways and combining the results; in fact we saw an example of this in our proof of Proposition 2.8. This is a very important idea, which is closely related to the mathematical discipline of "combinatorics", so we list it here as one of our methods. It often goes hand

[5] Gian-Carlo Rota (1932–1999): Italian-American mathematician, specialising in combinatorics.

in hand with an equivalence relation (Chapter 3); for example, it is used in the proof of Set Piece Theorem 4.

As an elementary example, we describe Gauss's derivation of his summation formula (cf. Proposition 2.6).

Proposition 2.10 (Gauss' summation formula; proof by counting)**.**
For all $n \in \mathbb{N}$ we have:

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

Note. For this proof it turns out to be better not to use the "sigma summation" notation!

Proof. Denote by $S(n)$ the sum we're after:

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

First, write this backwards:

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

Now add these two equations, taking care to pair up corresponding terms:

$$\begin{aligned} 2S(n) &= (1+n) + (2+n-1) + (3+n-2) + \cdots + (n-2+3) + (n-1+2) + (n+1) \\ &= (n+1) + (n+1) + \cdots + (n+1) \quad (n \text{ terms}) \\ &= n(n+1). \end{aligned}$$

Hence:

$$S(n) = \frac{1}{2}n(n+1),$$

as required. ■

There are similarly constructed counting proofs of the cardinality formula for finite power sets (Proposition 2.8), which we leave as an investigative exercise for anyone interested. The question of which type of proof—inductive or constructive—is "best" ultimately boils down to personal preference!

2.5. Proving logical equivalence; converse statements

A statement of the form $P \Leftrightarrow Q$ is actually asserting two things: it breaks up into $P \Rightarrow Q$ *and* $P \Leftarrow Q$. As noted earlier, the statement $P \Leftarrow Q$ is called the converse of $P \Rightarrow Q$, and is equivalent to the statement $Q \Rightarrow P$. The converse must not be confused with the contrapositive, or the negation, or any aspect of proof by contradiction. The contrapositive of the converse is in fact the statement:

$$\neg P \implies \neg Q,$$

which is sometimes called the *inverse* of the statement $P \Rightarrow Q$ (however we won't be using this terminology, to avoid any confusion with its use for functions, or later on when we meet groups in Chapter 4); it is logically equivalent to the converse.

When asked to prove a bi-implication $P \Leftrightarrow Q$, it is very important to remember to prove both directions. In favourable circumstances we may be able to prove the converse by simply reversing the direction of each implication in a direct proof of $P \Rightarrow Q$; we would usually present this as a sequence of bi-implications (recall Example 1.6). However this is not always possible, and a separate argument for the converse may be required.

It's also important to realise that the converse may not be true! Two important examples of this are given in Set Piece 4.

From this point in the course onwards we will indicate at the beginning of each proof what type it is, along with any further explanatory comments.

2.6. Set Piece 2: Different Sizes of Infinity

References.
Liebeck: Chapter 21, pp. 184–186.
Allenby: Chapter 9, Chapter 9, pp. 154–157.
Franklin and Daoud: Chapter 11, pp. 119–121.
Episodes 7 and 8 of video lectures.

In Set Piece 1 we considered countable sets, and showed that the infinite sets $\mathbb{Q}$, $\mathbb{Z}$ and $\mathbb{N}$ all have the same cardinality. Now we're going to consider our other favourite infinite set of numbers, namely the real numbers $\mathbb{R}$, and show that this set is fundamentally *bigger* than those three; in other words, the "infinity" describing how many whole numbers there are is *smaller* than the "infinity" describing how many real numbers there are. Even better (or worse, for anyone who doesn't like their mind to be messed with), we're going to prove Cantor's Theorem, which says that there are infinitely many infinities, each bigger than the previous one! The argument used in the proof of Cantor's Theorem is indeed mind bending, and we'll conclude by showing how a similar style of reasoning produces paradoxes in set theory. These precipitated a thorough examination of the foundations of set theory, and hence the foundations of mathematics, in the early twentieth century.

One of the themes throughout this Set Piece is the use of "proof by contradiction".

First, we need a bit more notation. Recall (Definition 1.9) that $|A| = |B|$ means there exists a bijection $f\colon A \to B$, in which case we say: "A and B have the same cardinality". So what should it mean for one set to have cardinality *less than* another? The following definition explains how to do this, using functions, and gives a few more familiar symbols a new meaning.

Definition 2.2 (Relative cardinality)**.**
Suppose A and B are sets.

- We write $|A| \leqslant |B|$ if there exists a one-to-one function $f\colon A \to B$. In this case we say: "the cardinality of A is less than or equal to that of B"; or: "the cardinality of B is at least that of A". We can also use the notation $|B| \geqslant |A|$, and say: "the cardinality of B is greater than or equal to that of A".
- We write $|A| \neq |B|$ if there doesn't exist any bijection $f\colon A \to B$. In this case we say: "the cardinality of A is not equal to that of B"; or: "A and B have different cardinalities".
- We write $|A| < |B|$ if $|A| \leqslant |B|$ and $|A| \neq |B|$. In this case we say: "the cardinality of A is strictly less than B". We can also use the notation $|B| > |A|$, and say: "B has strictly greater cardinality than A".

Putting together all these definitions, $|A| < |B|$ means precisely the following:

> *There exists a one-to-one function $f\colon A \to B$ but* there doesn't exist *a bijection.*

Intuitively, it is possible to pair up the elements of A with those of a subset of B (the image of f), but no matter how this is done there will always be some elements of B left over. ♦

Remarks 2.11.

1) Just like our original definition of cardinality (Definition 1.9), we're leaving the quantities $|A|$ and $|B|$ undefined (except when A, B are finite sets, in which case $|A|$ and $|B|$ are natural numbers); so in general the statements $|A| \leqslant |B|$ and $|A| < |B|$ have to be read simply as abbreviations for their defining conditions, and no more. (Of course if A, B are finite then the inequalities revert to their usual meaning.)

2) It's an interesting question whether or not the symbols "$\leqslant$" and "$<$" in these "inequalities" behave as we would want them to. (This was certainly the case for the "$=$" symbol in the "equation" $|A| = |B|$; recall Remark 1.21.) For example, it follows from Proposition 1.3 (which says that any composition of one-to-one functions is one-to-one) that:

   If $|A| \leqslant |B|$ *and* $|B| \leqslant |C|$ *then* $|A| \leqslant |C|$.

   However, the corresponding statement with "$<$":

   If $|A| < |B|$ *and* $|B| < |C|$ *then* $|A| < |C|$,

   whilst true is not quite so easy to prove. *(Can you see what the problem is?)*

   We'd also like to show in general that:

   If $|A| \leqslant |B|$ *and* $|B| \leqslant |A|$ *then* $|A| = |B|$.

Again, this isn't as obvious as the notation makes it look; writing out what each clause of the statement means should reveal why. The result that tells us it's in fact true is the "Schröder-Bernstein Theorem[6]". It's a little bit beyond the scope of our course (but could make an interesting group project). In what follows we will only be using results that we have proved. ♢

Here's a property of our new notation which is almost obvious but nevertheless useful to record.

Lemma 2.11 (Cardinality of subsets)**.**
If $A \subseteq B$ then $|A| \leqslant |B|$.

Proof. This is a direct proof.

There's a natural one-to-one function $f\colon A \to B$, namely the *inclusion map,* defined:

$$f(a) = a, \quad \text{for all } a \in A.$$

(This is rather similar to the identity function that we mentioned in Remark 1.17 (5); the rule is the same, but the codomain has changed!) It follows immediately from the definition (Definition 1.3) that f is one-to-one. ■

Remark. In contrast to finite sets, for infinite sets it is *not* the case that if $A \subset B$ (ie. A is a proper subset of B) then $|A| < |B|$; for example, $A = \mathbb{E}$ and $B = \mathbb{N}$. ♢

The following property generalises both Lemma 1.6 and Lemma 1.7. Recall (Definition 1.10) that a set is countable if it's either finite or in one-to-one correspondence with $\mathbb{N}$.

Proposition 2.12 (Cardinalities up to countability)**.**
Suppose A, B are sets, with B countable.

i) *If $|A| \leqslant |B|$ then A is countable.*

ii) *If $|A| < |B|$ then A is finite.*

Proof. The proof of (i) is direct, and we prove (ii) by contradiction!

The hypotheses of both (i) and (ii) include the existence of a one-to-one function $f\colon A \to B$. We denote by C the image of f (see Section 1.2.2); thus $C \subseteq B$. By restricting the codomain of f to C we obtain a bijection $f\colon A \to C$ (cf. the proof of Lemma 1.7; however, we haven't bothered renaming this "new" function!). Therefore $|A| = |C|$.

i) Being a subset of a countable set, C is countable by Lemma 1.7. Hence A is countable by Lemma 1.6.

[6] Named after German mathematicians Felix Bernstein (1878–1956) and Ernst Schröder (1841–1902), who published independent proofs. The result was originally stated without proof by Cantor.

ii) Suppose, for a contradiction, that A is infinite. Then so is C, hence so too is B by Proposition 1.5 (which would be contradicted if B were finite). So there exist bijections $g\colon A \to \mathbb{N}$ and $h\colon \mathbb{N} \to B$. By Proposition 1.3 the composition $h \circ g\colon A \to B$ is then also a bijection, contradicting the hypothesis $|A| \neq |B|$ (Definition 2.2). ■

Remark. By combining Proposition 2.12 (i) and Lemma 2.11 we recover Lemma 1.7: every subset of a countable set is countable. This is not surprising, since we used Lemma 1.7 during the proof! ◊

2.6.1. Uncountablity

Our study of cardinality in Set Piece 1 was confined to countable sets. However there is no reason to suppose that all infinite sets are countable, and in anticipation of this we make the following (rather obvious) definition.

Definition 2.3 (Uncountable sets).
A set A is *uncountable* if A is not countable; ie. A cannot be put into one-to-one correspondence with either $\mathbb{N}$ or $[n]$, for any $n \in \mathbb{N}$. ◆

Remark 2.12. If $|A| > |\mathbb{N}|$ then A is infinite (otherwise Proposition 1.5 would be contradicted, as in the proof of Proposition 2.12 (ii)), and A cannot be put into one-to-one correspondence with $\mathbb{N}$ (by Definition 2.2). Therefore A is uncountable. Although we won't need it, the converse is also true; ie. if A is uncountable then $|A| > |\mathbb{N}|$. However this requires some thought; ie. a proof! ◊

To show that Definition 2.3 is not vacuous we need an example of an uncountable set. Our next result provides us with one!

Theorem 2.13 (Uncountability of the reals).
The set $\mathbb{R}$ of real numbers is uncountable.

Proof. This is a rather famous proof by contradiction, featuring "Cantor's diagonal argument".

Suppose to the contrary that $\mathbb{R}$ is countable. Then $\mathbb{R}^+$ is also countable, by Proposition 2.12. This means there exists a bijection $f\colon \mathbb{N} \to \mathbb{R}^+$. Denoting $f(i) = r_i$, we therefore have a list $r_1, r_2, r_3, \ldots$ of all the positive real numbers. Let's write down the (start of) such a list, exhibiting each real number as a decimal expansion:

$$r_1 = m_1 \cdot a_{11}a_{12}a_{13}\ldots$$
$$r_2 = m_2 \cdot a_{21}a_{22}a_{23}\ldots$$
$$r_3 = m_2 \cdot a_{31}a_{32}a_{33}\ldots$$

$$\vdots$$

Here $m_i = 0, 1, 2, \dots$ is the integer part of r_i, followed by a decimal point, and then a digit $a_{ij} = 0, 1, \dots, 9$ in the j-th decimal place. To be unambiguous, we use the convention that the decimal expansion goes on forever, by replacing any finite decimal by one with recurring 9s (eg. $1 = 0{\cdot}9999999\cdots$).

We now produce a number $r \in \mathbb{R}^+$ *not* on the list, as follows:

$$r = 0 \cdot a_1 a_2 a_3 \dots$$

where $a_1 \neq a_{11}$, $a_2 \neq a_{22}$, $a_3 \neq a_{33}$ etc. We also choose each $a_i \neq 0$, to ensure r has a non-terminating decimal expansion. Then $r \neq r_i$ for each i, because they differ in their i-th decimal place. This implies that f is *not* onto, which is a contradiction. Hence $\mathbb{R}$ is uncountable. ■

Remark 2.13 (Continuum Hypothesis).
Theorem 2.13 shows that we have a new type of infinity, which is sometimes called the *cardinality of the continuum.* Furthermore Proposition 2.12 tells us there are no infinite sets X with cardinality $|X| < |\mathbb{N}|$. An interesting question is therefore whether there exist any (necessarily infinite) sets Y with cardinality $|\mathbb{N}| < |Y| < |\mathbb{R}|$. The *Continuum Hypothesis* asserts that there aren't. This was originally conjectured without proof by Cantor, and was first on the list of Hilbert's 23 problems[7]. Hilbert himself claimed to have found a proof in 1925 [22], but this was subsequently shown to be flawed. The conjecture still hasn't been proved; but neither has a counter-example been found! For a more recent twist in the story, which explains why this is not as surprising as it may sound, see Remark 2.16 below. ◊

Corollary 2.14 (Uncountability of the irrationals)**.**
The set of irrational numbers $\mathbb{Q}^c = \mathbb{R} \setminus \mathbb{Q}$ is uncountable.

Proof. A brief proof by contradiction.

The set $\mathbb{Q}$ is countable, by Set Piece Theorem 1. If $\mathbb{Q}^c$ were also countable then we could count the elements of $\mathbb{Q} \cup \mathbb{Q}^c$ using the same technique as Example 1.21 (which was also used in the final part of the proof of Set Piece Theorem 1). But $\mathbb{Q} \cup \mathbb{Q}^c = \mathbb{R}$, which is uncountable by Theorem 2.13. ■

We now move on to our main result, another theorem of Cantor, featuring the power set $\mathscr{P}(A)$ of a set A (see Section 1.1.6). We have seen how the cardinalities of $\mathscr{P}(A)$ and A are related if A is finite (Proposition 2.8). Cantor wanted to see what happens if A is

[7] David Hilbert (1862–1943): German mathematician who was pre-eminent at the cusp of the 19-th and 20-th centuries. He compiled 23 problems at the beginning of the new millennium, which set the course for much of the mathematical research during the twentieth century.

infinite, and ended up showing that there are an "infinitude of infinities". Interestingly, in some respects the proof is similar to Euclid's proof of the infinitude of primes (Theorem 2.4); sometimes: "The best tunes are played on the oldest fiddles.[8]"

Set Piece Theorem 2 (Cantor's Theorem)**.**
For any set A we have $|A| < |\mathscr{P}(A)|$. Hence we have an infinite sequence of infinities:

$$|\mathbb{N}| < |\mathscr{P}(\mathbb{N})| < |\mathscr{P}(\mathscr{P}(\mathbb{N}))| < \cdots$$

Note. Cantor's theorem is valid for *all* sets A. However when A is finite all it tells us is that $\mathscr{P}(A)$ has more elements than A, which is considerably less informative than Proposition 2.8. ♢

Proof. This is a case-by-case elimination of possibilities (proof by exhaustion), where each case is settled by a miniature proof by contradiction.

First off, there's a natural function $f\colon A \to \mathscr{P}(A)$ given by $f(a) = \{a\}$ for all $a \in A$, which is one-to-one by a simple application of the "standard routine" (Definition 1.3). So we just need to show there's no bijection. In fact, we will show that there's no surjection.

Suppose that $g\colon A \to \mathscr{P}(A)$ is any function. We show that g can't be onto, by finding an element of $\mathscr{P}(A)$ that doesn't lie in the image of g. Now, since elements of $\mathscr{P}(A)$ are subsets of A, for each $a \in A$ we have a subset $g(a) \subseteq A$. There are then two possibilities:

$$\begin{aligned} &a \in g(a), \text{ in which case we say } a \text{ is an "insider";} \\ &a \in g(a)^c, \text{ in which case we say } a \text{ is an "outsider".} \end{aligned}$$

Define a subset $B \subseteq A$, and hence an element of $\mathscr{P}(A)$, by:

$$B = \{a \in A : a \text{ is an outsider}\}.$$

To show that B is not in the image of g we argue by contradiction. Thus, suppose $B = g(b)$ for some $b \in A$.

Case 1. b is an insider. Thus $b \in g(b)$. But $g(b)$ contains only outsiders.

Case 2. b is an outsider. Thus $b \in g(b)^c$. But $g(b)^c$ contains only insiders.

Since both cases produce contradictions we're forced to conclude that no such element $b \in A$ exists, and therefore that g isn't onto. ■

Remarks 2.14.

1) Both Theorem 2.13 and Cantor's Theorem featured proof by contradiction, a technique that is particularly well-suited to "non-existence theorems" (cf. Theorem 2.5).

[8] Ralph Waldo Emerson (1803–1882): American philosopher, essayist, lecturer and anti-slavery campaigner, strongly motivated by "the infinitude of the private man".

2) It follows from Cantor's Theorem that $\mathscr{P}(\mathbb{N})$ is uncountable (see Remark 2.12). Then Theorem 2.13 invites the question:

$$|\mathscr{P}(\mathbb{N})| = |\mathbb{R}|?$$

This has a bearing on the Continuum Hypothesis (Remark 2.13) and is in fact true, although this is not immediately obvious; it requires a separate proof, which we won't pursue.

3) We will see an application of Cantor's Theorem in Chapter 4 (Example 4.18). $\Diamond$

2.6.2. Paradoxes of set theory

Our proof of Set Piece Theorem 2 established the following fact:

For all sets A there is no surjective function $g\colon A \to \mathscr{P}(A)$.

We can use this to settle a question that arises naturally in set theory.

Theorem 2.15 (Non-existence of a universal set)**.**
There is no "universal set"; ie. there is no set containing all sets.

Proof. Yet another proof by contradiction!

Suppose such a set exists; let's call it U. Now every subset of U, being a set, is an element of U. In other words, by definition of the power set, every element of $\mathscr{P}(U)$ is an element of U; so $\mathscr{P}(U) \subseteq U$ by definition of a subset.

We now construct a surjective function $g\colon U \to \mathscr{P}(U)$ as follows:

$$g(A) = \begin{cases} A, & A \in \mathscr{P}(U), \\ \emptyset, & A \in \mathscr{P}(U)^c. \end{cases}$$

(Recall that $\emptyset$ is always an element of the power set.) This contradicts the non-existence of a function from a set onto its power set; hence U cannot exist. ■

Remark 2.15 (Axiom of Choice).
The construction of the contradictory function g in the proof of Theorem 2.15 didn't require use of the "Axiom of Choice". This is a subtle point, relating to the axiomatisation of set theory that occurred historically as a direct result of "paradoxes" such as this. Without going into details, the Axiom of Choice is important and extremely useful because it permits the construction of sets without having to explicitly specify their elements. For example, it's needed to prove the elementary property that surjective functions are right-invertible (Remark 1.16)—actually, it's *equivalent* to that property. So the fact that it's not responsible for an inherent contradiction of naive set theory is good news! In fact, after

the axiomatisation of set theory it was shown by Gödel[9] [19] that the Axiom of Choice is consistent with the remaining axioms; ie. it doesn't introduce contradictions.

We will take a closer look at axiomatic mathematics (although not axiomatic set theory) in Chapter 4. $\Diamond$

Russell's paradox

The non-existence of a universal set:

$$U = \{A \mid A \text{ is a set}\}$$

shows that the "set builder" approach to defining sets described right at the beginning of Chapter 1:

$$\{x \mid P(x)\}$$

has limitations; some properties $P(x)$ give rise to internal contradictions, or "paradoxes". Bertrand Russell explored these limitations further, and in so doing ended up with what is now referred to as "Russell's paradox". In fact, he used an argument very similar to the proof of Set Piece Theorem 2.

Suppose we attempt to define a set B by:

$$B = \{A \mid A \text{ is a set and } A \notin A\}.$$

This is a reasonable proposal, because the condition $A \in A$ looks decidedly pathological, and something we could happily live without; in effect, since we'd have $U \in U$ if U were a set, we're restricting the definition of U in an attempt to make it more "civilised"! The argument now goes as follows:

> If B is a set, then either $B \in B$ or $B \notin B$ (*"to be or not to be"*).
>
> If $B \in B$ then $B \notin B$ (*"to be implies not to be"*).
>
> If $B \notin B$ then $\neg(B \notin B)$, which means $B \in B$ (*"not to be implies to be"*).

This contradiction can only be resolved by the conclusion that B is *not* a set.

What this means in practice is that when we push hard enough the "naive set theory" we've been using falls apart! The only way to put things right is to provide set theory with a proper foundation; certainly a set of rules that tell us precisely how we are allowed to build sets. These rules were laid out in "Zermelo-Fraenkel axiomatic set theory[10]"

[9] Kurt Gödel (1906–1978): Austrian mathematician, logician and philosopher, famous for (amongst other things) his "Incompleteness Theorem" which revolutionised our understanding of the nature of mathematical logic.

[10] Named after German mathematician Ernst Zermelo (1871–1953) and German/Israeli mathematician Abraham Fraenkel (1891–1965).

[12, 35], which was developed in response to Russell's paradox. Luckily all the "standard constructions" that we described in Section 1.1 are permitted, along with others that are slightly more subtle (such as those requiring the Axiom of Choice (Remark 2.15).

Remark 2.16 (Logical independence of the Continuum Hypothesis).
We might hope, and even expect, that any meaningful statement in axiomatic set theory can (ultimately) be either proved or disproved from the axioms. An example of such a statement is the Continuum Hypothesis (see Remark 2.13). It would therefore seem reasonable to assume that, given sufficient time and effort, it should be possible to decide whether it's true or false (in much the same way that a proof of Fermat's Last Theorem finally emerged after 350 years). However it was shown by Cohen[11] [10] that this is in fact impossible! More precisely, the Continuum Hypothesis is logically independent of the Zermelo-Fraenkel axioms for set theory; it is in fact an "undecidable" statement when working within the theory based on that particular system of axioms. ◊

[11] Paul Cohen (1934–2007): American mathematician, who was awarded the Fields Medal in 1966 for his work on the Continuum Hypothesis.

3. Equivalence Relations

References.
Liebeck: Chapter 18.
Episodes 9–11.5 of video lectures.

Informally, a relation on a set is a way of lumping together those of its elements that share some property. Equivalence relations have particular features that make them especially important and useful; they generalise the relation that lies at the heart of every area of mathematics: equality.

3.1. Relations

We saw in Set Piece Theorem 2 that sets can be very large. This makes it both natural and desirable to find ways of organising their elements. In fact, this applies even if a set is "only finite".

Take, for example, the set P of the world's current (human) inhabitants, a finite but very large set (approximately 7.8 billion elements). There are many organising principles for P: familial (eg. sharing a common ancestor); geopolitical (eg. citizenship); physical (eg. height); cultural (eg. musical taste), etc. Each of these gives rise to a "relation" on P. For example, we could say an individual $x \in P$ is "related" to another individual $y \in P$ if x and y have ever had a conversation. To record this information we could place x and y into the 2-element subset ("doubleton") $\{x, y\} \subset P$. The set of all these doubletons, which is a subset of the power set $\mathscr{P}(P)$, then gives us a complete "database" of who is "related" to whom.

However, there are some slightly more subtle relations on P that we may want to consider. For example, we could say that $x \in P$ is related to $y \in P$ if x has ever bought y a (not necessarily alcoholic) drink. In this case "x is related to y" is not the same as "y is related to x" (!), so to record this particular relationship between x and y we should use the *ordered pair* (x, y) rather than the unordered pair $\{x, y\}$. Now recall that the Cartesian product $P \times P$ is precisely the set of all ordered pairs of elements of P (Section 1.1.8). The "database" of who is related to whom is therefore a subset of $P \times P$, rather than $\mathscr{P}(P)$.

With all this in mind, the following formal definition of a relation should seem quite reasonable.

Definition 3.1 (Relation)**.**
Let A be a set.

- A *relation on A* is a subset $R \subseteq A \times A$; ie. a selection of ordered pairs of elements of A.
- If $x, y \in A$ with $(x, y) \in R$ we write $x \sim y$ and interpret this as: "x is related to y". ♦

Remarks 3.1.

1) Usually we'll just define a relation by specifying a rule for determining whether or not x is related to y, rather than explicitly giving the subset R of $A \times A$.
2) We will almost invariably refer to "the relation $\sim$" rather than the subset R.

(You may wonder why we introduced R at all! Really its only purpose is to allow us to make a precise definition.) ◊

There are a number of important properties that relations can have.

Definition 3.2 (Types of relation)**.**
Let $\sim$ be a relation on A. We say that $\sim$ is:

- *reflexive* if $x \sim x$ for all $x \in A$;
- *symmetric* if $x \sim y$ implies $y \sim x$, for all $x, y \in A$;
- *antisymmetric* if $x \sim y$ and $y \sim x$ implies $x = y$, for all $x, y \in A$;
- *transitive* if $x \sim y$ and $y \sim z$ implies $x \sim z$, for all $x, y, z \in A$. ♦

Remarks 3.2.

1) In order for any of the properties mentioned in Definition 3.2 to be valid they must hold for *all* elements of A. Thus, a property fails if it fails in just a single case, even though it may hold for all others. So, for example, the relation on P (the set of the world's current inhabitants) given by $x \sim y$ if "x has ever bought y a drink" in all likelihood has *none* of the properties of Definition 3.2.

2) The standard example of an antisymmetric relation is "$\leqslant$", defined on $\mathbb{N}$, $\mathbb{Z}$, $\mathbb{Q}$ or $\mathbb{R}$; ie. we define $x \sim y$ if and only if $x \leqslant y$. This relation is also reflexive and transitive, but not symmetric. Although antisymmetric relations are very important (for example, in the theory of partially ordered sets, or "posets"), they won't feature prominently in this course.

3) More generally, it's possible to define a relation between *two* sets A, B as a subset $R \subseteq A \times B$. It's the same idea: an element $a \in A$ is related to $b \in B$ if $(a, b) \in R$, and we even use the same notation $a \sim b$. We refer to R as a *relation from A to B*. In this case the properties of reflexivity, symmetry and transitivity no longer make sense (unless of course $B = A$), but there may be other properties of interest (see Remark 3.3). ◊

Examples 3.1 (Relations)**.**

1) The relation we all know and love, on any set A, is: $x \sim y$ if $x = y$. This has all the properties: reflexive, symmetric, transitive and antisymmetric.

2) We define a relation $\sim$ on $\mathbb{N}$ as follows:

   $n \sim m$ if and only if $n = 2^k m$ for some $k \in \mathbb{N}$.

   This relation is transitive. For, if $n \sim m$ and $m \sim p$ then $n = 2^k m$ and $m = 2^l p$, so $n = 2^k 2^l p = 2^{k+l} p$; hence $n \sim p$. However $\sim$ is neither reflexive, symmetric or antisymmetric.

3) We can "improve" the relation of (2) by making the following modification:

   $n \sim m$ if and only if $n = 2^k m$ for some $k \in \mathbb{N}_0$.

   Then $\sim$ picks up the reflexive property.

4) To further "improve" this relation we need to first change the set on which it is defined from $\mathbb{N}$ to $\mathbb{Q}$, and then make the following additional modification to the definition:

   $a \sim b$ for $a, b \in \mathbb{Q}$ if and only if $b = 2^k a$ for some $k \in \mathbb{Z}$.

   We end up with a relation that is reflexive, symmetric and transitive.

5) For any set A we define a relation $\sim$ on $\mathscr{P}(A)$ by:

   $X \sim Y$ if and only if $X \subseteq Y$.

   This has the same properties as "$\leqslant$" on $\mathbb{N}$ (see Remark 3.2 (2)). □

Remark 3.3 (Rigorous definition of a function).
Relations enable us to rigorously define functions; in particular, what we mean by the "rule" f. We define a function $f\colon A \to B$ to be a relation R from A to B (see Remark 3.2 (3)) with the following two properties:

- ○ $(\forall a \in A)\ (\exists b \in B)$ such that $(a, b) \in R$.
- ○ $(\forall a \in A)\ (\forall b, c \in B)\ (a, b) \in R$ and $(a, c) \in R$ implies $b = c$.

The use of quantifiers makes these statements look rather abstract. (Anyone going on to take Real Analysis will encounter similar-looking statements for the definitions of "limits", so this is a sort of "heads up" for what's coming!) However, after giving it a little bit of thought, all they're saying is that for every $a \in A$ there exists a unique $b \in B$ such that $(a, b) \in R$. We denote b by $f(a)$, and this gives us a precise definition of f. It captures the essence of our original, slightly less rigorous definition (Definition 1.1) of f as a "rule" that "unambiguously associates" an element of B to every element of A.

Another advantage of defining functions in this way is that it allows us to deduce the Principle for Equality of Functions. For, suppose we have functions $f, g\colon A \to B$, defined by relations R, S from A to B. Since $R, S \subseteq A \times B$ the Principle for Equality of Sets tells us precisely what is meant by $R = S$: the two sets have the same elements. Then, since $f(a)$ is the unique element of B such that $(a, f(a)) \in R$ we have $(a, f(a)) \in S$ also, hence by uniqueness $f(a) = g(a)$. Therefore $f(a) = g(a)$ for all $a \in A$, which is precisely the condition (1.2) for the rules f and g to be the same. ♢

3.2. Equivalence relations

An equivalence relation is a relation that has some of the special properties mentioned in Definition 3.2, that make it behave rather like the relation "$=$" of equality. Here's the definition:

Definition 3.3 (Equivalence relation)**.**
A relation $\sim$ on a set A is an *equivalence relation* if $\sim$ is reflexive, symmetric and transitive. Sometimes an equivalence relation is simply called an *equivalence.* ♦

An equivalence relation gives rise to "equivalence classes", which we can think of as its "families". Here's the definition:

Definition 3.4 (Equivalence classes)**.**
Given an equivalence relation $\sim$ on A and any $x \in A$, the *equivalence class* of x is the subset $[x] \subseteq A$ defined:

$$[x] = \{y \in A \mid x \sim y\}.$$

Reflexivity implies that for every $x \in A$ we have $x \in [x]$. The chosen element x is said to be a *representative* of the equivalence class $[x]$. ♦

Before looking at examples, we record a natural and very useful link between an equivalence relation and its equivalence classes: namely, that elements are related precisely when their equivalence classes are equal. Notice how each of the special properties of equivalence relations is used in the proof.

Proposition 3.1 (Characterisation of equivalence classes)**.**
Let $\sim$ be an equivalence relation on A. Then:

$$x \sim y \quad \textit{if and only if} \quad [x] = [y].$$

Proof. This is an "if and only if" statement, so we need to prove each direction. To prove the "only if" (forward) implication we use the Priniciple of Mutual Containment (Section 1.1.9) to show two sets are equal.

($\Rightarrow$) Suppose $x \sim y$. If $z \in [x]$ then by definition $x \sim z$, hence $z \sim x$ by symmetry. It now follows from transitivity that $z \sim y$, hence $y \sim z$ by symmetry again. Therefore $z \in [y]$ by definition. Hence $[x] \subseteq [y]$ by the definition of a subset. Since the relation is symmetric we can replace x by y in this argument to infer that $[y] \subseteq [x]$. Therefore $[x] = [y]$, by the Priniciple of Mutual Containment.

($\Leftarrow$) Conversely, suppose $[x] = [y]$. Then since $y \in [y]$ (as noted in Definition 3.4) we have $y \in [x]$; hence $x \sim y$. ■

Remark 3.4. The main purpose of an equivalence relation is to develop a slightly more generalised notion of equality than "true equality" (which is of course an equivalence relation; see Example 3.1 (1)). In fact, Proposition 3.1 says that an equivalence relation can indeed be regarded as true equality, provided the objects being equated are equivalence classes rather than individual elements. If the relation happens to be true equality, then its equivalence classes are singletons: $[a] = \{a\}$ for all $a \in A$; so the two notions coincide. However the greater flexibility permitted by an equivalence relation means that the equivalence classes can be larger; in some cases infinite! A "good" equivalence relation is one whose equivalence classes are reasonably large, but not so large that they become meaningless. ♢

In Section 1.2.7 we came across something that looks rather similar to an equivalence relation; let's take a closer look.

Example 3.2 (Equinumerosity)**.**
Let X be a fixed set, and define a relation $\sim$ between subsets $A, B \subseteq X$ by:

$$A \sim B \quad \text{if and only if} \quad |A| = |B|.$$

Thus, two subsets are related if they have the same cardinality; such sets are sometimes said to be *equinumerous.* Bearing in mind that $A \sim B$ means there exists a bijection between A and B (Definition 1.9), we showed in Remark 1.21 that $\sim$ is symmetric and transitive. Furthermore $\sim$ is also reflexive, because the identity function $1_A\colon A \to A$ is a bijection, so $A \sim A$. Therefore $\sim$ is an equivalence relation on the power set $\mathscr{P}(X)$.

The equivalence class $[A]$ contains all subsets of X that have the same cardinality as A. For example, if $X = \mathbb{R}$ then $[\mathbb{N}]$ contains $\mathbb{Z}$ and $\mathbb{Q}$ by Set Piece 1, and many more. The precise number of equivalence classes depends on the logical status of the Continuum Hypothesis (Remark 2.13). □

Remark 3.5. It is tempting to extend the equinumerosity relation of Example 3.2 to *all* sets. However, our definition of a relation (Definition 3.1) requires $\sim$ to be defined on a set; but we saw in Theorem 2.15 that the collection of all sets is not a set! This is an inherent limitation of Zermelo-Fraenkel axiomatic set theory. However it can be overcome by an extension of the theory due to von Neumann[1], Bernays[2] and Gödel, which allows

[1] John von Neumann (1903–1957): Hungarian/American mathematician, with panoramic mathematical interests.

[2] Paul Bernays (1888–1977): Swiss mathematician, who collaborated with David Hilbert.

"paradoxical sets" such as those mentioned in Set Piece 2 to exist as "classes" rather than sets, and then permits relations to be defined on classes. ◊

In the next Section we discuss a really important example, that in some sense is the prototype for Definitions 3.3 and 3.4.

3.3. Congruence of integers

This is a relation on $\mathbb{Z}$, the set of all integers (negative, positive, and of course zero). It originates with the early work of Gauss in number theory (circa 1800), before the general notion of an equivalence relation had been developed. Some of the notation and terminology reflects this, being of "classical" vintage.

Definition 3.5 (Congruence modulo n)**.**
Suppose n is a fixed natural number, and define a relation $\sim_n$ on $\mathbb{Z}$ by:

$$a \sim_n b \quad \text{if and only if} \quad a - b = kn \text{ for some } k \in \mathbb{Z}.$$

(It's important to allow $k \leqslant 0$, as we will shortly see.) This relation is called *congruence modulo n,* and n is known as the *modulus.* The alternative (classical) notation is also often used:

$$a \equiv b \pmod{n},$$

particularly in number theory. ♦

The idea of this relation is to consider two integers "the same" if one differs from the other by a multiple of n. So, for example, if $n = 12$ then $13 \sim_{12} 1$, $25 \sim_{12} 1$, $37 \sim_{12} 1$, etc. Thus the relation of congruence modulo 12 is the one we use on a daily basis to "clock" the hours of our lives!

Let's verify that $\sim_n$ is an equivalence relation:

Reflexive. We have $a \sim_n a$, because $a - a = 0 = 0.n$.

Symmetric. Suppose $a \sim_n b$, so $a - b = kn$. Then $b - a = (-k)n$, so $b \sim_n a$.

Transitive. Suppose $a \sim_n b$ and $b \sim_n c$, so $a - b = kn$ and $b - c = ln$. Then:

$$a - c = a - b + b - c = kn + ln = (k + l)n,$$

so $a \sim_n c$.

The equivalence classes of $\sim_n$ are denoted $[a]_n$, and called *congruence classes, modulo n.* To see what they are, we divide a by n as many times as possible to obtain an equation:

$$a = kn + r,$$

where the remainder $r = 0, 1, \dots, n-1$. Then:

$$a - r = kn,$$

so $a \sim_n r$, hence $[a]_n = [r]_n$ by Proposition 3.1. It follows that there are precisely n congruence classes, corresponding to each of the n possible remainders:

$$[0]_n = \{kn : k \in \mathbb{Z}\}, \quad [1]_n = \{1 + kn : k \in \mathbb{Z}\}, \dots, \quad [n-1]_n = \{n - 1 + kn : k \in \mathbb{Z}\}. \tag{3.1}$$

Notice that each class is an infinite subset of $\mathbb{Z}$. Because the remainder after division by n is also known (classically) as the *residue* modulo n, congruence classes are also known as *residue classes.*

A familiar example is $n = 10$, where the residue of a is nothing other than the first (ie. right-most) digit in its decimal representation. So, for example:

$$[157]_{10} = [7]_{10}.$$

It's therefore very easy to determine congruence classes modulo 10. But what about other moduli? We can think of the least residue of a modulo n as its first "digit" when expanded in base n (see also Remark 3.9 (2)). When $n = 2$ this is the first binary digit of a, which is 0 or 1 depending on whether a is even or odd; so:

$$[a]_2 = \begin{cases} [0]_2, & \text{if } a \text{ is even,} \\ [1]_2, & \text{if } a \text{ is odd.} \end{cases}$$

The method described in the following example, which dates back to the ancient Greeks, deals with modulus 9.

Example 3.3 (Casting out nines)**.**
Suppose $n = 9$. Assume for convenience that $a \in \mathbb{N}$ (ie. a positive integer) and write a in decimal form:

$$a = a_n \cdots a_2 a_1,$$

where the digits $a_i = 0, 1, \dots, 9$. We first compute the *digit sum:*

$$a' = a_1 + \cdots + a_n.$$

We now compute the digit sum a'' of a', and keep going. After a finite number, say m, of iterations we reach a digit sum $a^{(m)} = r$ where r is a single digit, called the *digital root* of A. We claim that:

$$a \equiv r \pmod 9;$$

so if $r = 1, \dots, 8$ then the digital root is the residue of a modulo 9, whereas if $r = 9$ then the residue is 0.

Before giving a proof, suppose for example that $a = 2159483$. Then:

$$a' = 2 + 1 + 5 + 9 + 4 + 8 + 3 = 32, \qquad a'' = 5.$$

So the digital root of a is 5, which means that:

$$[2159483]_9 = [5]_9.$$

Now let's prove this. Our notation for the decimal digits of a means precisely that:

$$a = a_n 10^{n-1} + \cdots + a_2 10 + a_1.$$

Therefore:

$$a - a' = a_n(10^{n-1} - 1) + \cdots + a_2(10 - 1).$$

Each term of this sum is divisible by 9 (see Proposition 2.7), hence by Definition 3.5:

$$a \sim_9 a'. \tag{$*$}$$

By successive application of $(*)$ we obtain a chain of equivalences:

$$a \sim_9 a' \sim_9 a'' \sim_9 \cdots \sim_9 a^{(m-1)} \sim_9 a^{(m)} = r,$$

which, since $\sim_9$ is transitive, implies $a \sim_9 r$, and therefore $[a]_9 = [r]_9$ by Proposition 3.1.

Although "casting out nines" is already remarkably efficient, the relation $(*)$ allows us to do the job even more quickly, by ignoring digits that are either 9 or sum to 9. So, for $a = 2159483$ we can ignore the digits 9, 5, 4, 8 and 1 and arrive at the digital root $2 + 3 = 5$ immediately! □

Being able to calculate congruence classes modulo 10, 2 and 9 allows us to determine classes with respect to a range of other moduli. There are two main techniques for doing this, which we describe in the next pair of examples.

Example 3.4 (New congruence classes from old: divisors of the modulus)**.**
If $a \sim_n b$ and m is a divisor of n then it follows immediately from Definition 3.5 that $a \sim_m b$. Therefore, if r is the least residue of a modulo n, then the least residue of r modulo m is the least residue of a modulo m.

This probably sounds more complicated than it really is! For example, if $a = 2159483$ and $n = 9$ then having determined in Example 3.3 that $[a]_9 = [5]_9$ we can easily determine the congruence class of a mod 3 by:

$$[a]_3 = [5]_3 = [2]_3,$$

where the second equation is obtained by simply taking the remainder of 5 when divided by 3.

The same principle applies to determining residue classes mod 5 from residue classes mod 10. For example, since $[157]_{10} = [7]_{10}$ we have:

$$[157]_5 = [7]_5 = [2]_5.$$

For this pair of moduli we have simply formalised a calculation that we would otherwise probably consider to be "obvious". □

Example 3.5 (New congruence classes from old: products of moduli)**.**
It's also possible to determine the congruence class of an integer a modulo $n = n_1 n_2$ if we already know the congruence classes $[a]_{n_1}$ and $[a]_{n_2}$. In effect, this is the opposite of what we did in Example 3.4, and we therefore approach the problem by using what we learned there and "thinking backwards".

As an illustration, let's calculate the least residue r of $a = 2159483$ modulo $n = 18$. We express $n = n_1 n_2$ where $n_1 = 2$ and $n_2 = 9$, knowing that $[a]_9 = [5]_9$ from Example 3.3 and $[a]_2 = [1]_2$ since a is odd. From Example 3.4 we also know that:

$$[r]_9 = [a]_9 = [5]_9.$$

So, since r lies in the range 0 to 17 there are two possibilities: $r = 5$ or $r = 14$. However Example 3.4 also tells us that:

$$[r]_2 = [a]_2 = [1]_2,$$

so r is odd. There is now only one possibility: $r = 5$. Therefore $[a]_{18} = [5]_{18}$. By applying Example 3.4 one more time, since 6 is also a factor of 18 we get "for free" that $[a]_6 = [5]_6$.

This all seems very satisfactory, and may tempt us to believe that the method will work for the product of any pair n_1, n_2 of moduli. However, in mathematics it's always advisable to check several examples before formulating a conjecture! So, in the spirit of discovery, for the same integer $a = 2159483$ let's attempt to determine $[a]_{18}$ from the known congruence classes $[a]_6 = [5]_6$ and $[a]_3 = [2]_3$. There is already the hint of a fly in the soup, since Example 3.4 tells us that $[a]_3$ can be determined from $[a]_6$ and therefore brings nothing new to the table. Now, since $[r]_6 = [5]_6$ there are three possibilities: $r = 5$, $r = 11$ or $r = 17$. But $[r]_3 = [2]_3$ gives: $r = 2, 5, 8, 11, 14, 17$, which has *lengthened* the list of possibilities! So the method fails for $n_1 = 3$ and $n_2 = 6$; indeed, this counter-example shows that in general the congruence class of an integer mod 18 cannot be determined from its congruence classes mod 3 and mod 6. □

Example 3.5 leaves us with an intriguing question, the answer to which is: the Chinese Remainder Theorem. To state and prove this requires the following simple idea, which crops up on numerous occasions in elementary number theory.

Definition 3.6 (Coprime integers)**.**
Two natural numbers are said to be *coprime,* or *relatively prime,* if their only common positive divisor is 1. ♦

It's clear that if $m, n \in \mathbb{N}$ are both prime then they're coprime. However, there are other possibilities: for example, $m = 3$ and $n = 4$, where just one integer is prime; and $m = 4$, $n = 9$, where neither is prime.

Closely related to this are the following two quantities.

Definition 3.7 (Greatest common divisor and least common multiple)**.**
Suppose $m, n \in \mathbb{N}$.

1) The *greatest common divisor* (or *highest common factor*) of m and n is the greatest integer d satisfying:
$$m = ad \quad \text{and} \quad n = bd,$$
for integers a, b. We write $d = \gcd(m, n) = \operatorname{hcf}(m, n)$.

2) The *least common multiple* of m and n is the least positive integer M satisfying:
$$M = am = bn,$$
for integers a, b. We write $M = \operatorname{lcm}(m, n)$. ♦

It follows immediately from Definitions 3.6 and 3.7 that m, n are coprime if and only if $\gcd(m, n) = 1$, which is its minimum possible value. However, for this to be of any practical use we need an efficient method for calculating greatest common divisors. Luckily there is one: "Euclid's algorithm" (see Appendix A.2). (Although this algorithm is remarkably effective and widely used, it is not crucial to what follows; furthermore, most of our examples are simple enough to be done "by inspection".) Having calculated the greatest common divisor, we can then easily determine the least common multiple from the following equation:
$$mn = \gcd(m, n) \operatorname{lcm}(m, n). \tag{3.2}$$
(We will prove equation (3.2) in Appendix A.3; see Proposition A.7.) In particular, this tells us that m, n are coprime if and only if $\operatorname{lcm}(m, n) = mn$, which is its maximum possible value.

The following result is named in honour of the Chinese mathematician Sun Zi[3].

Theorem 3.2 (Chinese Remainder Theorem)**.**
Suppose natural numbers n_1, n_2 are coprime, and $n = n_1 n_2$. Then for all $a \in \mathbb{Z}$ the congruence class $[a]_n$ is determined by the classes $[a]_{n_1}$ and $[a]_{n_2}$.

Proof. This is a proof by contradiction.

Let r_1 and r_2 denote the residues of a modulo n_1 and n_2, respectively. The theorem states that if n_1 and n_2 are coprime then there exists a unique integer r in the range 0 to $n - 1$ such that:
$$[r]_{n_1} = [r_1]_{n_1} \quad \text{and} \quad [r]_{n_2} = [r_2]_{n_2}.$$
So assume for a contradiction that there exist distinct integers r, s in the range 0 to $n - 1$ such that:
$$[r]_{n_1} = [r_1]_{n_1} = [s]_{n_1} \quad \text{and} \quad [r]_{n_2} = [r_2]_{n_2} = [s]_{n_2}.$$

[3] Sun Zi (400–460): Chinese mathematician of the late Jin and early North-South dynasties, and the eponymous author of "Sunzi Suanjing" ("Sun Zi's Mathematical Manual"), which contains the earliest known written record of the Chinese Remainder Theorem.

By Definition 3.5 this means that there exist integers k_1, k_2, l_1, l_2 such that:

$$r = r_1 + k_1 n_1 = r_2 + k_2 n_2, \qquad s = r_1 + l_1 n_1 = r_2 + l_2 n_2.$$

Therefore:

$$r - s = (k_1 - l_1) n_1 = (k_2 - l_2) n_2.$$

Assuming for argument's sake that $r > s$, this implies that the positive integer $r - s$ is a common multiple of n_1 and n_2 that is less than n. Hence $\mathrm{lcm}(n_1, n_2) < n$, contradicting that n_1 and n_2 are coprime. ■

Reviewing Example 3.5 in the light of Theorem 3.2, our method "worked" for the moduli $n_1 = 2$ and $n_2 = 9$, which are coprime, but failed for $n_1 = 3$ and $n_2 = 6$, which aren't.

The Chinese Remainder Theorem, as stated, is purely an existence theorem; it doesn't provide any details about how to calculate $[a]_n$ from $[a]_{n_1}$ and $[a]_{n_2}$. It's possible to construct a general formula that does this; however in practice the "sieving method" from Example 3.5 will usually do the trick, with the additional virtue of not having to memorise anything! We give another example to show how easy this is.

Example 3.6 (Chinese Remainder Theorem)**.**
Suppose once again that $a = 2159483$. We saw in Example 3.4 that $[a]_3 = [2]_3$, and it is easy to see that $[a]_5 = [3]_5$. Since the moduli 3 and 5 are coprime (by "inspection"), the Chinese Remainder Theorem ensures that these two classes will allow us to determine $[a]_{15}$.

Since $[a]_3 = [2]_3$ the possible residues of a modulo 15 are: $2, 5, 8, 11, 14$. But from $[a]_5 = [3]_5$ the only possibilities are: $3, 8, 13$. These two lists have only one entry in common; therefore $[a]_{15} = [8]_{15}$. □

3.4. Partitions and quotients

Looking back at Section 3.3 we see that every integer belongs to one and only one congruence class. We could illustrate this geometrically by "colour-coding" the points of $\mathbb{Z}$ according to their congruence class (Figure 3.1). Then every integer receives a colour, and no integer receives more than one.

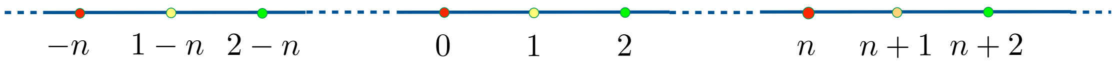


Figure 3.1.: Colour-coded congruence classes

We say that the congruence classes "partition" $\mathbb{Z}$. This property of congruence classes is in fact shared by equivalence classes in general. Here's the formal definition.

Definition 3.8 (Partition)**.**
A *partition* of a set A is a collection of subsets $(S_i)_{i\in I}$ for some index set I, with the following two properties:

P1) $A = \bigcup_{i\in I} S_i$

P2) $S_i \cap S_j = \emptyset$ for all $i \neq j$.

We also say that the subsets $(S_i)_{i\in I}$ *partition* A. ♦

Remarks 3.6.

1) Technically speaking, a partition of A is a subset of the power set $\mathscr{P}(A)$.
2) Condition (P1) says that every element of A belongs to some S_i. We say that the S_i *cover* A.
3) Condition (P2) says that no element of A belongs to more than one S_i. We say that the S_i are *pairwise disjoint.*
4) In practice, it's often more convenient to work with the contrapositive version of (P2) (see Section 2.4.2), which is the statement:

   $$\text{if } S_i \cap S_j \neq \emptyset \text{ then } S_i = S_j.$$

   (Can you see why this might be preferable?) We will use this in our proof of Proposition 3.3 below. ♢

Although Definition 3.8 looks quite abstract, a moment's thought should be enough to see that it's encoding exactly the same property of congruence classes: every element of A belongs to a subset S_i, and no element belongs to more than one. A good analogy is that a partition is a "jigsaw puzzle" for the set A (see Figure 3.2).

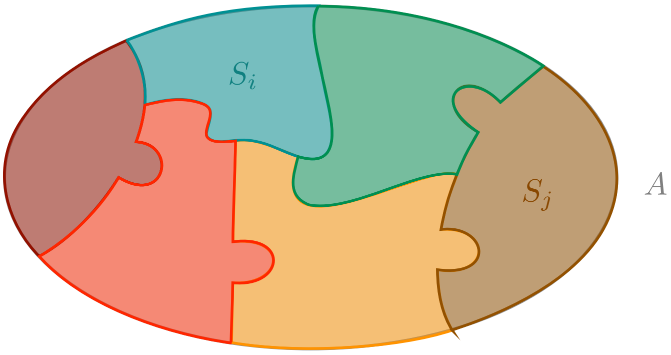


Figure 3.2.: Partition of a set

Example 3.7 (Partition of the plane by circles)**.**
In Example 1.18 we encountered a partition of the plane $\mathbb{R}^2$ into a family of circles, together with the origin. The index set in this case is:

$$I = \mathbb{R}^+ \cup \{0\},$$

which is infinite, indicating that the number of sets in the partition is infinite, and for each index $i \in I$ the set S_i is:

$$S_i = \{(x, y) \in \mathbb{R}^2 \mid x^2 + y^2 = i^2\}.$$

Thus if $i > 0$ then S_i is a circle centred at $(0, 0)$ and the index i is simply its radius, whereas $S_0 = \{(0, 0)\}$ is the "degenerate circle" of "radius" 0. We can visualise properties (P1) and (P2) geometrically (see Figure 1.5): the circles fill up the entire plane, and every point lies on precisely one circle. □

Now here's the main result, showing that equivalence relations and partitions are essentially the same thing.

Proposition 3.3 (Equivalence relations and partitions)**.**
Let A be any set.

i) *If $\sim$ is an equivalence relation on A, then the equivalence classes of $\sim$ partition A.*

ii) *Conversely, if $(S_i)_{i \in I}$ is a partition of A, define a relation $\sim$ on A by:*

$$a \sim b \quad \textit{if and only if} \quad a, b \in S_i \ \textit{ for some } i \in I.$$

Then $\sim$ is an equivalence relation, whose equivalence classes are precisely the subsets S_i.

Proof. This proof is an "if and only if", with each part corresponding to one direction; furthermore, the first part uses the contrapositive.

i) Every element $a \in A$ belongs to an equivalence class, namely $[a]$ (by the reflexive property). So the equivalence classes cover A, which is (P1).

We now prove the contrapositive version of (P2) (see Remark 3.6 (4)). So, suppose $[a]$ and $[b]$ are equivalence classes with $[a] \cap [b] \neq \emptyset$. Hence there exists an element $x \in [a] \cap [b]$. Then $x \in [a]$ and $x \in [b]$, so $a \sim x$ and $b \sim x$, which by symmetry and transitivity implies $a \sim b$. Hence $[a] = [b]$ by Proposition 3.1.

ii) Conversely, if we define $\sim$ on A as stated then it follows immediately that $\sim$ is reflexive and symmetric. To show it's transitive, suppose $a \sim b$ and $b \sim c$. Then $a, b \in S_i$ and $b, c \in S_j$ for some $i, j \in I$. But then $S_i \cap S_j$ is non-empty (it contains b) so $S_i = S_j$ by the contrapositive version of (P2), and this implies $a \sim c$. So $\sim$ is an equivalence relation.

To determine the equivalence classes, suppose $a \in S_i$. Then:

$$\begin{aligned} x \in [a] &\iff a \sim x, && \text{by Definition 3.4} \\ &\iff x \in S_i && \text{by definition of } \sim. \end{aligned}$$

Therefore $[a] = S_i$. ■

Note. We didn't use the property (P1) of partitions in the proof of (ii). However it was used (implicitly) to define $\sim$ in the first place. $\Diamond$

Example 3.8 (Equivalence relation for partition by circles)**.**
It follows from Proposition 3.3 that the circles in Example 3.7 must be the equivalence classes of an equivalence relation $\sim$ on $\mathbb{R}^2$. To see what this relation could possibly be, recall that this example was originally introduced (Example 1.18) to illustrate the level sets of the function:

$$f\colon \mathbb{R}^2 \to \mathbb{R};\, f(x, y) = x^2 + y^2.$$

In fact $S_i = f^{-1}(i)$, for $i \in \mathbb{R}_0^+$ (see Remark 1.19 for a reminder of this notation). Now two points $(x, y), (u, v) \in \mathbb{R}^2$ are related by $\sim$ precisely when they lie on the same circle S_i, hence in the same level set of f. Hence:

$$(x, y) \sim (u, v) \text{ if and only if } f(x, y) = f(u, v).$$

Since we already know that the circles partition $\mathbb{R}^2$ it follows from Proposition 3.3 that there is no need to check that $\sim$ is an equivalence relation. $\square$

Example 3.8 points towards the following much more general result, which provides some useful information about the configuration of the fibres of a function.

Proposition 3.4 (Partition by fibres of a function)**.**
Suppose $f\colon X \to Y$ is any function, and define a relation $\sim$ on X by:

$$a \sim b \textit{ if and only if } f(a) = f(b),$$

for all $a, b \in X$. Then $\sim$ is an equivalence relation, whose equivalence classes are the fibres of f. Therefore the fibres of f partition X.

Proof. This is a straightforward verification of properties.

The relation $\sim$ is clearly reflexive, symmetric and transitive, and is therefore an equivalence relation (Definition 3.3). For any $a \in A$ the equivalence class $[a]$ is by definition (Definition 3.4):

$$[a] = \{x \in X \mid f(x) = f(a)\}.$$

On the other hand, since the fibre of f over $y \in Y$ is by definition (see Remark 1.19):

$$f^{-1}(y) = \{x \in X \mid f(x) = y\},$$

it follows that:

$$[a] = f^{-1}(f(a)).$$

That the fibres of f partition X now follows from Proposition 3.3. ■

One of the reasons for introducing equivalence relations is to help increase our mathematical efficiency, by working with equivalence classes rather than individual elements of the original set. It is therefore desirable to produce a new set whose elements are precisely the equivalence classes themselves.

Definition 3.9 (Quotient set)**.**
Suppose $\sim$ is an equivalence relation on a set A. The *quotient set* (also called the *factor set,* or *identification set*) of A by $\sim$ is defined to be the new set:

$$A/{\sim} = \{[a] : a \in A\}.$$

There is also the *quotient function* $\pi\colon A \to A/{\sim}$ defined:

$$\pi(a) = [a],$$

for all a in A. (Use of "π" to denote this function comes from the alternative terminology "projection" map; it has nothing whatsoever to do with the number π.) ♦

Remarks 3.7.

1) The reason for the "quotient" terminology will become apparent in Section 3.7.
2) The "identification" terminology highlights an important mathematical phrase and concept, that pops up on many occasions throughout the subject. We're using the word "identify" in the sense "identify with", as in identifying with a particular group, cause or idea. So, when two elements are related by an equivalence relation they are considered to identify with each other. Then in the "identification set" they have literally been "identified", or "regarded as the same"; in effect, their individual identities have been lost (or disregarded)!
3) The quotient construction is the sixth important way of constructing new sets from old, following the five mentioned in Chapter 1: intersection, union, difference (and complement), power set, Cartesian product.
4) The elements of $A/{\sim}$ are subsets of A. Recalling that the power set $\mathscr{P}(A)$ is the set of *all* subsets of A, it follows that $A/{\sim} \subseteq \mathscr{P}(A)$. However, that's not usually the way we think about the quotient set!
5) The quotient function $\pi\colon A \to A/{\sim}$ is onto, by definition. What about one-to-one? By Proposition 3.1 we have:

   $$\pi(a) = \pi(b) \implies [a] = [b] \implies a \sim b.$$

   So π is one-to-one only if "$\sim$" is "$=$".
6) A nice theoretical aspect of the quotient set and quotient map is that the construction of equivalence relations from functions described in Proposition 3.4 becomes universal. In other words, for every equivalence relation $\sim$ on a set X there's a function $f\colon X \to Y$ such that:

   $$a \sim b \text{ if and only if } f(a) = f(b).$$

The set Y is the quotient set $X/{\sim}$, and f is the quotient function π. This is a direct consequence of Definition 3.9 when placed alongside Proposition 3.1:

$$a \sim b \text{ if and only if } [a] = [b]. \qquad \Diamond$$

If we're lucky, the quotient set $A/{\sim}$ will share some of the properties of A. (If we're very lucky $A/{\sim}$ may have additional properties that A *doesn't* have; see Section 4.3.3). In the next Section we will see this in action for the example of congruence of integers (Section 3.3); and further directions of travel are mentioned briefly in Remark 4.28 (5). It shows why congruence is such an important equivalence relation.

3.5. The modular integers and modular arithmetic

In Section 3.3 we defined the relation $\sim_n$ of congruence modulo n on $\mathbb{Z}$, and showed it is an equivalence relation. The quotient set for this relation is:

$$\mathbb{Z}_n = \mathbb{Z}/{\sim_n} = \{[a]_n : a \in \mathbb{Z}\} = \{\bar{a} : a \in \mathbb{Z}\}.$$

The set $\mathbb{Z}_n$ is usually simply called "zed en", or more formally *$\mathbb{Z}$ mod n,* or *the integers modulo n.* As we have seen, its cardinality is:

$$|\mathbb{Z}_n| = n.$$

What makes $\mathbb{Z}_n$ so useful is that we can do arithmetic in it! Furthermore, this arithmetic is closely related to standard arithmetic on $\mathbb{Z}$, so any calculations made in $\mathbb{Z}_n$ tell us something about corresponding calculations in $\mathbb{Z}$ (although there will inevitably be some loss of information). Even better, we are already familiar (to some extent) with how this works. Here are the definitions.

Modular addition. We define:

$$[a]_n + [b]_n = [a+b]_n. \tag{3.3}$$

Modular multiplication. We define:

$$[a]_n [b]_n = [ab]_n. \tag{3.4}$$

In other words, we simply add or multiply integers as usual, then divide through by n and extract the remainder, or "residue". However, giving this a moment's thought, we see there is a potential ambiguity: congruence classes have infinitely many different representatives, which means there are infinitely many ways of computing the sum $a+b$ or product ab in $\mathbb{Z}$ before taking the residue. (This kind of problem almost always arises when we work with equivalence classes.) We need to show that the result is independent of the representatives we choose to work with; in other words, that modular addition and multiplication are "well-defined". So let's do that.

Modular addition is well-defined. Suppose $[a]_n = [c]_n$ and $[b]_n = [d]_n$. By Proposition 3.1 this means: $a \sim_n c$ and $b \sim_n d$; thus $a - c = kn$ and $b - d = ln$ for integers $k, l \in \mathbb{Z}$. Therefore:

$$a + b = c + kn + d + ln = c + d + (k + l)n,$$

hence:

$$a + b \sim_n c + d.$$

So by Proposition 3.1 again:

$$[a + b]_n = [c + d]_n,$$

which resolves the potential ambiguity in equation (3.3).

Modular multiplication is well-defined. Under the same assumptions on a, b, c, d we calculate:

$$ab = (c + kn)(d + ln) = cd + cln + knd + kln^2 = cd + (cl + kd + kln)n,$$

hence:

$$ab \sim_n cd.$$

Therefore:

$$[ab]_n - [cd]_n,$$

and this resolves the potential ambiguity in equation (3.4).

We have just defined the basic operations of *modular arithmetic.* In practice, when treating congruence classes as "numbers" in this way it is often convenient to adopt a more streamlined notation. So, if the modulus that we're using is clearly understood, and it therefore really isn't necessary to keep writing it down, then we commonly abbreviate:

$$[a]_n = [a] = \bar{a}.$$

The "bar notation" removes the "notational clutter" of the subscript and square brackets ("Principle of Notational Simplicity"). It enables us to write modular addition and multiplication more smoothly, as follows:

$$\bar{a} + \bar{b} = \overline{a + b}, \qquad \bar{a}\bar{b} = \overline{ab}.$$

Remarks 3.8.

1) Modular arithmetic obeys many of the laws of ordinary arithmetic, such as the associative, commutative and distributive laws:

$$(\bar{a} + \bar{b}) + \bar{c} = \bar{a} + (\bar{b} + \bar{c}), \qquad (\bar{a}\bar{b})\bar{c} = \bar{a}(\bar{b}\bar{c}),$$
$$\bar{a} + \bar{b} = \bar{b} + \bar{a}, \qquad \bar{a}\bar{b} = \bar{b}\bar{a},$$

$$\bar{a}(\bar{b}+\bar{c}) = \bar{a}\bar{b}+\bar{a}\bar{c}, \quad \text{etc.}$$

These are consequences of the corresponding laws in $\mathbb{Z}$. For example:

$$(\bar{a}+\bar{b})+\bar{c} = \overline{(a+b)+c} = \overline{a+(b+c)} = \bar{a}+(\bar{b}+\bar{c}).$$

We will take a more detailed look at some of these in Chapter 4 (Section 4.3.2).

2) In addition to modular addition and multiplication we can also define *modular subtraction* as follows:

$$[a]-[b] = [a-b].$$

This is well-defined, because it can be expressed in terms of modular addition and multiplication:

$$[a-b] = [a]+[-b] = [a]+[-1][b].$$

3) It's also possible to define *modular exponentiation* for any exponent $m \in \mathbb{N}$ by:

$$[a]^m = [a^m].$$

This is simply iterated modular multiplication:

$$[a^m] = [a][a^{m-1}] = \cdots = [a]\cdots[a], \quad m \text{ times},$$

and is therefore also well-defined. ♢

Modular arithmetic is something we're already familiar with, perhaps unwittingly, in at least a couple of cases. If $n = 12$ then modular addition is simply "clock addition": so for example, we know that 8 hours on from 5 o-clock is 1 o-clock, which is the same answer we get using modular addition:

$$\bar{5}+\bar{8} = \overline{5+8} = \overline{13} = \bar{1}.$$

If $n = 10$ then we know from "long addition" and "long multiplication" that the last decimal digit of $a+b$ or ab is the last digit of the sum or product of the last digits of a and b. This is exactly what happens when we perform addition or multipication modulo 10. So, for example, the last digit of $759+618$ is:

$$\overline{759+618} = \overline{759}+\overline{618} = \bar{9}+\bar{8} = \overline{9+8} = \overline{17} = \bar{7},$$

and that of $(759)(618)$ is:

$$\overline{(759)(618)} = \overline{759}\,\overline{618} = \bar{9}\,\bar{8} = \overline{9.8} = \overline{72} = \bar{2}.$$

A major application of modular arithmetic, and the reason why it was originally introduced, is to quickly prove nice things in number theory. Here's a very simple example, generalising Proposition 2.7. Not only is the proof quicker, but it also gives a much better idea why the result is true.

Proposition 3.5 (Divisibility in base b)**.**
Suppose b and n are positive integers, with $b > 1$. Then $b^n - 1$ is divisible by $b - 1$.

Proof. This is a simple direct proof.

The statement of the result can be expressed rather neatly in terms of congruence classes modulo $b - 1$ as follows:

$$[b^n - 1] = [0],$$

where we have simplified the notation by omitting the subscripts $b - 1$. We can verify this equation by manipulating the left hand side using the laws of modular arithmetic:

$$\begin{aligned}[b^n - 1] &= [b^n] - [1] = [b]^n - [1] \\ &= [1]^n - [1], \quad \text{since } b \equiv 1 \pmod{b-1} \\ &= [1^n] - [1] \\ &= [1] - [1] = [1 - 1] = [0],\end{aligned}$$

as required. ■

Remarks 3.9.

1) One way to “see” Proposition 3.5 is by switching from decimal to base b representations of integers, whose “digits” are multiples of powers of b. Then $b^n - 1$ has $n - 1$ “digits”, each of which is $b - 1$.

2) The most familiar example of a non-decimal number system is of course the *binary* system ($b = 2$), which also happens to be the one case when Proposition 3.5 has nothing to say! Other examples of note are *duodecimal* ($b = 12$), *hexadecimal* ($b = 16$) and *undecimal*[4] ($b = 11$), which we will encounter again in Example 3.10. There is a famous mathematical construction[5] that uses the *ternary* system ($b = 3$).

3) It follows immediately from Proposition 3.5 that $b^n - 1$ is divisible by all factors of $b - 1$. For example, $11^n - 1$ is always even (divisible by 2) and divisible by 5 as well as 10; and $13^n - 1$ is divisible by 2, 3, 4 and 6, as well as 12.

4) Pursuing the theme of the previous remark, if b is odd then $b - 1$ is even, so $b^n - 1$ is even by Proposition 3.5, therefore b^n is odd. So we've quickly proved that any (positive integer) power of an odd integer is odd.

5) There are other ways to prove Proposition 3.5. Proof by induction will certainly work (cf. Proposition 2.7), but perhaps the most direct method is to observe the factorisation:

$$b^n - 1 = (b - 1)(1 + b + b^2 + \cdots + b^{n-1}).$$

This is familiar, because when rearranged it gives the well-known formula for the sum of a geometric progression! Since b is an integer so is $1 + b + \cdots + b^{n-1}$; hence $b - 1$ divides $b^n - 1$. ◊

[4] An undecimal number system was briefly proposed in revolutionary France, as a compromise between those wanting to introduce a duodecimal system and those wanting to retain decimal numbers!

[5] The Cantor set.

The final section of Chapter 3 before we reach Set Piece 3 is an application of modular arithmetic to an area of much more recent significance: error detection.

3.6. Error detection

When manipulating numerical data there is always the possibility of errors arising; so much so, in fact, that rather than attempt to prevent errors it is often better to detect (and then correct) them. Clever use of modular arithmetic helps us to do this. We will describe a couple of scenarios that illustrate the ideas.

Our first scenario is based on a very simple idea. Suppose we have proposed values for a sum $a + b$ or product ab of two integers, and want to check them. This might be simply because we want to check our own calculations. Or it could be because we have received these values after their transmission along a "noisy" data pipe. For example, transmitting the value of a product of two (large) primes is an important ingredient of "public key cryptography".

From the laws (3.3) and (3.4) of modular arithmetic, for any modulus the residue class of $a+b$ is the sum of those of a and b, and the residue class of ab is the product of those of a and b. So, computing the residue classes of $a + b$ and ab and comparing them with the sum/product of the individual classes provides us with a check.

Example 3.9 (Error detection for arithmetic)**.**
Suppose $a = 769$ and $b = 887$ (which are in fact both prime, although rather small for cryptographic purposes). We will work with the modulus 9, and use "casting out nines" (Example 3.3) to quickly compute residues.

- We first compute the residues (mod 9) of a and b by computing their digit sums (ignoring digits that are either 9 or sum to 9):

$$a' = 7 + 6 = 13, \quad a'' = 4, \quad \text{so} \ \ a \sim_9 4;$$
$$b' = 8 + 8 + 7 = 23, \quad b'' = 5, \quad \text{so} \ \ b \sim_9 5.$$

- Now suppose we have received the following values for the sum and product:

$$a + b = 1746, \quad ab = 682203.$$

Casting out nines yields their residues (mod 9):

$$a + b \sim_9 1 + 7 + 4 + 6 = 18 \sim_9 1 + 8 = 9 \sim_9 0,$$
$$ab \sim_9 6 + 8 + 2 + 2 + 3 = 21 \sim_9 2 + 1 = 3.$$

- Suppose we have also been sent the above residues of a and b (mod 9), as an extra pair of digits at the end of the values of $a+b$ and ab; so 165645 and 68220345, respectively. (Note that the order of the extra digits is unimportant.) We can then use modular arithmetic to compute:

$$\bar{a} + \bar{b} = \bar{4} + \bar{5} = \overline{4+5} = \bar{9} = \bar{0};$$
$$\bar{a}\,\bar{b} = \bar{4}\,\bar{5} = \overline{4.5} = \overline{20} = \bar{2}.$$

We conclude immediately that the received value of ab is incorrect; it may have been incorrectly entered, or corrupted during transmission, and we would probably ask for it to be re-sent. (The true value of ab is in fact 682103.) However, we can only conclude that the received value of $a + b$ is "consistent with being correct", because passing our test is only a necessary condition for being correct, not sufficient. We may be happy to take our chances and proceed. Or we may ask the sender to compute the residues of a and b with respect to another modulus and send those to us (along with the new modulus), so we can perform another check. (The true value of $a + b$ is in fact 1656.) □

Our second scenario is more sophisticated, and arises when we have a large database whose information is being transcribed: either entered, extracted, or cross referenced. Many transcriptional errors can be detected using "checksums". The following example illustrates the idea.

Example 3.10 (International Standard Book Numbers)**.**
Every book now comes with an International Standard Book Number, or ISBN, which serves to identify it uniquely. Originally, when first introduced back in 1970, these were 10 digit "numbers". For example:

0-471-00006-X

was the ISBN of [4]. Ignoring the hyphens, this is of the form:

$$x_{10}x_9 \cdots x_2x_1,$$

where $x_2, \ldots, x_{10}$ are ordinary decimal digits (ie. $0, 1, \ldots, 9$), and x_1 is an *undecimal* digit (ie. $0, 1, \ldots, 9,$ X); that is, a digit in base 11 (see Remark 3.9 (2)). The decimal digits constitute the book catalogue number $x_{10} \cdots x_2$; since there are nine of them, the database can accommodate up to a billion books. The undecimal digit x_1 is a "checksum", and is related to the catalogue number by the following equation in $\mathbb{Z}_{11}$:

$$[x_1 + 2x_2 + \cdots + 10x_{10}]_{11} = [0]_{11}. \tag{3.5}$$

The term on the left hand side is a *weighted sum* of all the digits; the coefficients, or *weights,* are what are responsible for the error detecting property of the ISBN (see Proposition 3.6 below).

Switching to "bar notation" (mod 11) and using modular arithmetic allows us to rewrite (3.5) as:

$$\overline{x_1} = -\bar{2}\,\overline{x_2} - \cdots - \overline{10}\,\overline{x_{10}},$$

which shows that the checksum digit is uniquely determined by the catalogue number. For example, the catalogue number of [4] has:

$$\begin{aligned}\bar{2}\,\overline{x_2} + \cdots + \overline{10}\,\overline{x_{10}} &= \bar{2}\bar{6} + \bar{7}\bar{1} + \bar{8}\bar{7} + \bar{9}\bar{4} \\ &= \overline{12} + \bar{7} + \overline{56} + \overline{36} \\ &= \bar{1} + \bar{7} + \bar{1} + \bar{3} \\ &= \overline{12} = \bar{1}.\end{aligned}$$

Therefore $\overline{x_1} = -\bar{1} = \overline{10}$, so $x_1 = $ X, exactly as in the ISBN. □

We now come to the main practical purpose of the checksum, which is to eliminate one of the most common sources of transcriptional error: transpositions. The following result shows how this works for ISBNs (Example 3.10).

Proposition 3.6 (ISBN error detection)**.**
The ISBN checksum detects transpositional errors; ie. catalogue numbers with a pair of flipped digits have different checksums.

Proof. We prove the contrapositive. The proof culminates with a mini exhaustion.

Suppose we have ISBNs:

$$X = x_{10} \cdots x_1, \qquad Y = y_{10} \cdots y_1,$$

such that $y_r = x_s$ and $y_s = x_r$ for some pair r, s with $2 \leqslant r < s \leqslant 10$, and $x_i = y_i$ for all $i \neq r, s$. In particular $x_1 = y_1$, so X and Y have the same checksum. We want to show that this can only happen if $x_r = x_s$; ie. $X = Y$. Taking the contrapositive gives the result as stated.

We know that X and Y both satisfy (3.5), so subtracting one equation from the other and applying modular arithmetic gives us:

$$\begin{aligned}[0]_{11} &= [x_1 - y_1 + 2(x_2 - y_2) + \cdots + 10(x_{10} - y_{10})]_{11} \\ &= [(r - s)(x_r - x_s)]_{11}.\end{aligned}$$

This means that 11 divides $(r - s)(x_r - x_s)$. Since 11 is prime, this can only happen if 11 divides either $r - s$ or $x_r - x_s$. The former is impossible since $1 \leqslant s - r \leqslant 8$. Since $-9 \leqslant x_r - x_s \leqslant 9$ the latter is only possible if $x_r - x_s = 0$; ie. $x_r = x_s$. ■

Remarks 3.10.

1) A similar but slightly simpler argument shows that two catalogue numbers that differ in just one digit also have different checksums. (In fact, this argument doesn't require the sum (3.5) to be weighted.) So ISBNs sucessfully detect the two most common types of transcription errors: "flips" and (one digit) "slips".

2) Since 2007 ISBNs have had 13 digits: 12 catalogue digits and a checksum digit, allowing a catalogue of up to a trillion books. Although the precise mechanism is slightly different, the underlying principle is the same as for 10 digit ISBNs.

3) International Bank Account Numbers (IBANs) work in a similar way, as do International Standard Serial Numbers (ISSNs), which are used to catalogue magazines and newspapers, International Standard Music Numbers (ISMNs) for cataloguing sheet music, and many others. ◊

There are lots more examples of equivalence relations and their corresponding partitions in Problem Set 3. Our next Set Piece gives a more involved example.

3.7. Set Piece 3: Construction of the rational numbers

References.
Episodes 11 and 11.5 of video lectures.

We tend to think of the rational numbers $\mathbb{Q}$ as being a subset of the real numbers $\mathbb{R}$, which of course they are. However, this assumes that we have already constructed $\mathbb{R}$, a construction which in fact requires $\mathbb{Q}$ as a starting point! (We won't go into the details of this construction, since it requires some knowledge of real analysis.) So, to avoid a circular definition, at some point an independent construction of $\mathbb{Q}$ is required.

Using the notion of an equivalence relation, we can now indeed give a rigorous construction of $\mathbb{Q}$, starting with the integers $\mathbb{Z}$. (Of course, this immediately invites the question: how do we construct $\mathbb{Z}$? For this, see Section 3.7.2.) Although this procedure doesn't tell us anything startlingly new, it does put some of the earlier material in the course on an even more rigorous footing; and it gives another concrete example of how equivalence relations are used in modern algebra.

Starting from $\mathbb{Z}$ means that we *don't* have the operation of division available to us—we can only add, subtract or multiply—so we can't simply "divide a by b" when trying to create the rational number a/b from the integers a, b. Even if we treat a/b purely symbolically we still have the problem that different pairs of integers a, b produce the same (symbolic) rational number a/b; for example, $1/2$ and $(-3)/(-6)$. The solution is to lump together these seemingly different "fractions" using an equivalence relation; in other words, to construct $\mathbb{Q}$ as a quotient set. We then need to check that the laws of addition and multiplication make sense once this lumping together has been achieved. All this is analogous to defining the relation of congruence modulo n on $\mathbb{Z}$, and setting up modular arithmetic on the quotient set $\mathbb{Z}_n$ (Sections 3.3 and 3.5).

First off, since rational numbers are fractions, which involve a numerator and a denominator, we're really working with ordered pairs of integers; we'll think of the first "coordinate" as the numerator, and the second as the denominator. So we start with the set $S = \mathbb{Z} \times \mathbb{Z}^*$, where $\mathbb{Z}^*$ denotes the set of nonzero integers (since we don't permit division by 0; we discuss the reason for this in Section 3.7.1). To see what relation we should impose on S we look ahead to what we're trying to achieve, and notice that if ordered pairs (a, b) and (c, d) correspond to the same rational number then we would have the following equation in $\mathbb{Q}$ (once constructed!):

$$\frac{a}{b} = \frac{c}{d}.$$

By "clearing the denominators" this can be rearranged to:

$$ad = bc,$$

which is an equation that makes sense in $\mathbb{Z}$. We can therefore use it to define a relation $\sim$ on S by:

$$(a, b) \sim (c, d) \quad \text{if and only if} \quad ad = bc, \tag{3.6}$$

for all $a, b, c, d \in \mathbb{Z}$ with $b, d \neq 0$. We now need to check that $\sim$ does everything that we hope it will.

Set Piece Theorem 3 (Construction of the rationals)**.**
With the above in hand, we have:

i) *The relation $\sim$ is an equivalence relation on S.*

ii) *Define $\mathbb{Q} = S/{\sim}$, the quotient set, and denote the equivalence class of (a, b) by a/b. Then $a/b = c/d$ if and only if $ad = bc$.*

iii) *There are well-defined operations of addition and multiplication on $\mathbb{Q}$ given by:*

$$\frac{a}{b} + \frac{c}{d} = \frac{ad + bc}{bd}, \qquad \frac{a}{b} \cdot \frac{c}{d} = \frac{ac}{bd}.$$

Note. The operations of arithmetic on $\mathbb{Q}$ agree with the rules we learnt at school! ◊

Proof. This is a direct proof; the statement of the theorem tells us what to do, and our task is to go through each step required. The routines (showing something is an equivalence relation and then defining some structure on the quotient set) are themselves fairly standard, and will become increasingly familiar.

i) We show that $\sim$ has the three required properties (Definition 3.3), bearing in mind its definition (3.6).

Reflexive. We have $(a, b) \sim (a, b)$ for all $(a, b) \in S$, because $ab = ba$. (Notice the use of commutativity of multiplication in $\mathbb{Z}$.)

Symmetric. Suppose $(a, b) \sim (c, d)$ for $(a, b), (c, d) \in S$. Then $ad = bc$. But this is the same as $cb = da$, which is the condition for $(c, d) \sim (a, b)$. (Again, we have used the commutativity of multiplication in $\mathbb{Z}$.)

Transitive. Suppose $(a, b) \sim (c, d) \sim (e, f)$ for $(a, b), (c, d), (e, f) \in S$. Then $ad = bc$ and $cf = de$. We multiply the first equation by f and the second by b:

$$adf = bcf, \qquad bcf = bde,$$

and then eliminate the common term to obtain:

$$adf = bde.$$

Since $d \neq 0$ we can cancel it (see Proposition 2.1 and Remark 2.1) to get:

$$af = be,$$

which is the condition for $(a, b) \sim (e, f)$. (Yet again, we have used commutativity of multiplication in $\mathbb{Z}$, along with associativity, both of which we have taken for granted.)

ii) This follows directly from Proposition 3.1 and the definition (3.6) of $\sim$.

iii) Now, applying our "fractional notation" for the equivalence classes of $\sim$, suppose:

$$\frac{a}{b} = \frac{\alpha}{\beta} \quad \text{and} \quad \frac{c}{d} = \frac{\gamma}{\delta}.$$

Thus by (ii):

$$a\beta = b\alpha \quad \text{and} \quad c\delta = d\gamma. \tag{†}$$

Our task is to show:

$$\frac{a}{b} + \frac{c}{d} = \frac{\alpha}{\beta} + \frac{\gamma}{\delta} \quad \text{and} \quad \frac{a}{b} \cdot \frac{c}{d} = \frac{\alpha}{\beta} \cdot \frac{\gamma}{\delta},$$

which, using the proposed definitions of addition and multiplication, along with (ii), means showing:

$$(ad + bc)\beta\delta = bd(\alpha\delta + \beta\gamma), \tag{$*$}$$

$$ac\beta\delta = bd\alpha\gamma. \tag{$**$}$$

Addition. Consider the left hand side of $(*)$:

$$\begin{aligned}(ad + bc)\beta\delta &= ad\beta\delta + bc\beta\delta \\ &= b\alpha d\delta + d\gamma b\beta, \quad \text{by (†)} \\ &= bd(\alpha\delta + \beta\gamma),\end{aligned}$$

which is what we wanted.

Multiplication. Consider the left hand side of $(**)$:

$$\begin{aligned}ac\beta\delta &= b\alpha d\gamma, \quad \text{by (†)} \\ &= bd\alpha\gamma,\end{aligned}$$

as required.

Notice that in these final calculations we have used very nearly all the laws of arithmetic in $\mathbb{Z}$. ■

Remarks 3.11.

1) The equivalence classes (colour-coded) are illustrated in Figure 3.3. In this picture we visualise $S = \mathbb{Z} \times \mathbb{Z}^*$ as a subset of the Cartesian plane, comprising all points whose coordinates are integers (ie. the entire "integer lattice"; cf. Example 1.22), with those on the x-axis removed. Two points of S are then related if they lie on the same straight line through the origin. The fact that the origin does *not* belong to S ensures the classes are pairwise disjoint, as they must be by Proposition 3.3.

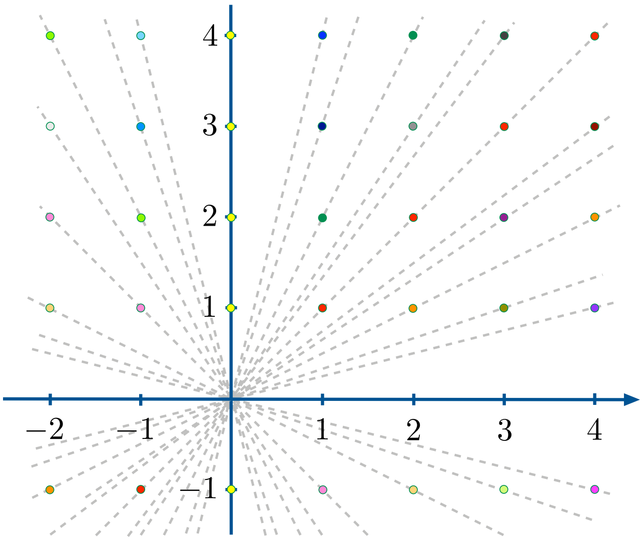


Figure 3.3.: Colour-coded equivalence classes for the rationals

2) We apply familiar terminology to the elements of $\mathbb{Q}$; thus we refer to a/b as a *fraction,* with *numerator* a and *denominator* b. (Nevertheless, we should't forget that a/b is really an equivalence class!)

3) It follows from part (ii) of the Theorem that common factors in the numerator and denominator of fractions can be cancelled in the usual way; that is, if $k \in \mathbb{Z}$ is nonzero, then:

$$\frac{ak}{bk} = \frac{a}{b},$$

for all $a, b \in \mathbb{Z}$ with $b \neq 0$. (We used this in our proof of Set Piece Theorem 1, and the famous proof of Theorem 2.5.)

4) There is a natural copy of $\mathbb{Z}$ inside $\mathbb{Q}$. More precisely, we define a function $f\colon \mathbb{Z} \to \mathbb{Q}$ by $f(a) = a/1$. Then f is one-to-one; for if $f(a) = f(b)$ then $a/1 = b/1$ hence $a = b$, by part (ii) of the Theorem. It follows that $f\colon \mathbb{Z} \to f(\mathbb{Z})$ is a bijection; so $f(\mathbb{Z}) \subset \mathbb{Q}$ is a copy of $\mathbb{Z}$. (Notice that we used the "trick" of creating a new (onto) function by simply reducing the size of the codomain; we already encountered this in our proofs of Set Piece Theorem 1 and Proposition 2.12 in Set Piece 2.)

It is conventional to blur this distinction slightly, writing $\mathbb{Z} \subset \mathbb{Q}$ and denoting $a/1$ by a. For the time being, we refer to these as the "integer elements" of $\mathbb{Q}$.

5) The operations of arithmetic in $\mathbb{Q}$ are consistent with those of $\mathbb{Z}$. More precisely, using the function $f\colon \mathbb{Z} \to \mathbb{Q}$ defined in (4):

$$f(a+b) = \frac{a+b}{1} = \frac{a}{1} + \frac{b}{1} = f(a) + f(b),$$
$$f(ab) = \frac{ab}{1} = \frac{a}{1}\frac{b}{1} = f(a)f(b),$$

where addition and multiplication on the left hand side of these equations are those of $\mathbb{Z}$, whereas on the right hand side they are the newly defined operations of $\mathbb{Q}$. (To an algebraist, these properties say that f is a *ring homomorphism;* more about this in Year 2.) It's important to note this, because in principle the sum and product of two integer elements of $\mathbb{Q}$ needn't have turned out to be an integer!

6) Suppose $p, q, r \in \mathbb{Q}$. Then it's not hard to check the *associative laws, commutative laws* and *distributive law* all hold:

$$(p+q)+r = p+(q+r), \qquad (pq)r = p(qr),$$
$$p+q = q+p, \qquad pq = qp,$$
$$(p+q)r = pr + qr.$$

Again, it's important to note this because, although by (5) these laws hold for the integer elements of $\mathbb{Q}$ (since they hold in $\mathbb{Z}$), there is no particular reason why they should hold for all the newly-constructed elements.

7) The integer elements $0 = 0/1$ and $1 = 1/1$ interact with the other elements of $\mathbb{Q}$ as they should:

$$0 + \frac{a}{b} = \frac{0b+1a}{1b} = \frac{a}{b}, \qquad 0 \cdot \frac{a}{b} = \frac{0a}{1b} = 0, \qquad 1 \cdot \frac{a}{b} = \frac{1a}{1b} = \frac{a}{b},$$

for all $a/b \in \mathbb{Q}$. The first and third of these equations say that 0 is an *additive identity* and 1 is a *multiplicative identity* in $\mathbb{Q}$, respectively. (We will have more to say about identity elements in Chapter 4.)

8) Define:

$$-\frac{a}{b} = \frac{-a}{b},$$

which by part (ii) of the Theorem is a well-defined element of $\mathbb{Q}$; in fact:

$$-\frac{a}{b} = (-1) \cdot \frac{a}{b}.$$

It satisfies:

$$\frac{a}{b} + \left(-\frac{a}{b}\right) = 0;$$

otherwise said, $-a/b$ is an *additive inverse* for a/b. (We will have more to say about inverse elements in Chapter 4.) We now extend the operation of subtraction from $\mathbb{Z}$ to $\mathbb{Q}$ in a natural way by defining:

$$\frac{a}{b} - \frac{c}{d} = \frac{a}{b} + \Big(-\frac{c}{d}\Big).$$

9) Everything we have said so far has been to show that the familiar arithmetic of $\mathbb{Z}$ (addition, multiplication and subtraction) extends to $\mathbb{Q}$. However, $\mathbb{Q}$ has an "ace up its sleeve". If $a \neq 0$ we define:

$$\Big(\frac{a}{b}\Big)^{-1} = \frac{b}{a},$$

which by part (ii) of the Theorem is a well-defined element of $\mathbb{Q}$. Then:

$$\frac{a}{b} \cdot \Big(\frac{a}{b}\Big)^{-1} = \frac{a}{b} \cdot \frac{b}{a} = \frac{1}{1} = 1.$$

This shows that every nonzero element of $\mathbb{Q}$ has a *multiplicative inverse.*

We now define *division* in $\mathbb{Q}$ by:

$$\frac{a}{b}\Big/\frac{c}{d} = \frac{a}{b} \cdot \Big(\frac{c}{d}\Big)^{-1} = \frac{a}{b} \cdot \frac{d}{c}.$$

In particular, the rational number a/b is then indeed the integer a divided by the integer b, for:

$$a \cdot b^{-1} = \frac{a}{1} \cdot \Big(\frac{b}{1}\Big)^{-1} = \frac{a}{1} \cdot \frac{1}{b} = \frac{a}{b}.$$

Furthermore for all $q \in \mathbb{Q}$ we have the *rule of reciprocation:*

$$q^{-1} = 1 \cdot q^{-1} = \frac{1}{q}.$$

10) These properties show that $\mathbb{Q}$ is a *field;* it's called the *field of fractions* of $\mathbb{Z}$. This construction is so successful that it can be replicated with virtually no changes in other more abstract settings; but this is really a topic for a more advanced course in abstract algebra. ◊

3.7.1. Division by zero: a glimpse of projective geometry

A frequently asked question concerning the arithmetic of fractions is: "Why can't we divide by zero?" Our construction of the rationals deliberately prohibited this, by defining the equivalence relation $\sim$ on the set $S = \mathbb{Z} \times \mathbb{Z}^*$ of all ordered pairs (a, b) with $b \neq 0$, thereby excluding the possibility of producing fractions of the form $a/0$. So the short answer is: "Because there aren't enough rational numbers!" In fact, there is a very good reason for this prohibition: it enables us to use the cancellation law (Proposition 2.1) in the proof of part (i) of Set Piece Theorem 3 to show that $\sim$ is transitive. This is important because $\sim$ defines equality in the quotient set $\mathbb{Q} = S/{\sim}$ (Proposition 3.1); so without transitivity we could encounter situations where rational numbers p, q, r satisfy $p = q$ and $q = r$ but $p \neq r$. Amongst other things, this would make algebra in $\mathbb{Q}$ impossible!

However, the definition (3.6) of $\sim$ certainly makes sense on the Cartesian product $\mathbb{Z}\times\mathbb{Z}$, and careful consideration will show that in fact it *is* an equivalence relation provided we remove just the single point $(0, 0)$. Thus, we define $S' = (\mathbb{Z} \times \mathbb{Z}) \smallsetminus \{(0, 0)\}$ and let $\sim'$ be the relation on S' defined by:

$$(a, b) \sim' (c, d) \quad \text{if and only if} \quad ad = bc.$$

This is essentially the relation $\sim$ but with the restriction $b, d \neq 0$ removed. Now, referring back to the proof of part (i) of Set Piece Theorem 3 we see that $\sim'$ is reflexive and symmetric, since these properties only depend on the commutativity of multiplication in $\mathbb{Z}$. However, we need to revisit the argument for transitivity. So let $(a, b), (c, d), (e, f) \in S'$ with $(a, b) \sim' (c, d) \sim' (e, f)$; thus:

$$ad = bc \quad \text{and} \quad cf = ed.$$

If $d \neq 0$ then we can proceed exactly as in the proof, cancelling d to obtain $(a, b) \sim' (e, f)$. If $d = 0$ then from $ad = bc$ we obtain $bc = 0$. Now $c \neq 0$, since $(c, d) \neq (0, 0)$ and $d = 0$; hence $b = 0$. Similarly, from $cf = ed$ we obtain $cf = 0$; hence $f = 0$. Therefore:

$$af = 0 = eb,$$

which shows that $(a, b) \sim' (e, f)$, albeit in a rather trivial way. Thus $\sim'$ is transitive, and therefore an equivalence relation on S'; the argument for transitivity only required $(c, d) \neq (0, 0)$ rather than $d \neq 0$.

We can now go ahead and form the quotient set $S'/{\sim'}$, which we denote rather provocatively by $\mathbb{Q}_\infty$ and refer to as the set of *projectively extended rational numbers,* or more concisely the *projectivised rationals,* or the *rational projective line.* The appearance of the "∞" symbol, and meaning of some of the terminology, will be explained in due course. The question now is: "What do the elements of $\mathbb{Q}_\infty$ look like, and how do they interact with the elements of $\mathbb{Q}$?" Elements of $\mathbb{Q}_\infty$ are of course the equivalence classes of $\sim'$, which we will denote by $[(a, b)]'$ for $(a, b) \in S'$. We will continue to denote the equivalence classes of $\sim$ by $[(a, b)]$, bearing in mind that $\sim$ is only defined on S so $b \neq 0$.

Our first observation is that $\mathbb{Q} \subseteq \mathbb{Q}_\infty$. This is essentially because $S \subset S'$ and $\sim'$ agrees with $\sim$ on S. More precisely, if $[(a, b)] \in \mathbb{Q}$ then by the definition of equivalence classes (Definition 3.4) we have:

$$\begin{aligned}[(a, b)] &= \{(c, d) \in S \mid (a, b) \sim (c, d)\}\\ &\subseteq \{(c, d) \in S' \mid (a, b) \sim' (c, d)\}\\ &= [(a, b)]'.\end{aligned}$$

On the other hand, if $(a,b) \sim' (c,d)$ for $(c,d) \in S'$ then since $ad = bc$, $b \neq 0$ and $(c,d) \neq (0,0)$ we have $d \neq 0$. Hence $(c,d) \in S$ and $(a,b) \sim (c,d)$. Therefore:

$$\begin{aligned}[(a,b)]' &= \{(c,d) \in S' \mid (a,b) \sim' (c,d)\} \\ &\subseteq \{(c,d) \in S \mid (a,b) \sim (c,d)\} \\ &= [(a,b)].\end{aligned}$$

It follows (from the Principle of Mutual Containment) that:

$$[(a,b)] = [(a,b)]', \tag{3.7}$$

and therefore that $[(a,b)] \in \mathbb{Q}_\infty$. Hence $\mathbb{Q} \subseteq \mathbb{Q}_\infty$; so $\mathbb{Q}_\infty$ "extends" $\mathbb{Q}$. (This argument turned out to be quite a good "work out" for set theory, which is perhaps not surprising given that we are dealing with infinite sets whose elements are themselves infinite sets!)

We now extend the "fractional notation" from $\mathbb{Q}$ to $\mathbb{Q}_\infty$ by simply defining:

$$[(a,b)]' = \frac{a}{b}.$$

which by (3.7) is consistent with its use to denote elements of $\mathbb{Q}$. The fractions a/b with $b \neq 0$ are precisely the elements of $\mathbb{Q}$, and fractions $a/0$ with $a \neq 0$ are now also permitted! However, the fraction $0/0$ is not allowed, since $(0,0) \notin S'$. Thus:

$$\mathbb{Q}_\infty = \left\{\frac{a}{b} \mid a,b \in \mathbb{Z},\ a,b \text{ not both zero}\right\}.$$

Since $\sim$ and $\sim'$ are defined by the same equation, the rule for equality in $\mathbb{Q}_\infty$ stays the same:

$$\frac{a}{b} = \frac{c}{d} \quad \text{if and only if} \quad ad = bc.$$

In particular:

$$\frac{a}{0} = \frac{c}{0},$$

for all $a, c \neq 0$. So there is only one "new" fraction, for which it is irresistibly tempting to introduce the notation:

$$\infty = \frac{1}{0}.$$

We conclude that $\mathbb{Q} \subset \mathbb{Q}_\infty$ and $\mathbb{Q}_\infty = \mathbb{Q} \cup \{\infty\}$.

Having justified the introduction of the "∞" symbol from a purely set theoretic point of view, we can also get a nice geometric picture of $\mathbb{Q}_\infty$ from Figure 3.3. If $b \neq 0$ then a/b may be visualised as the set of all non-zero integer lattice points in the Cartesian plane lying on the line through the origin with slope b/a. So every rational number corresponds to a line through $(0,0)$ with non-zero rational slope. There is one glaring omission from this family of lines—the x-axis—and the extended rational number ∞ is simply the set of all non-zero integer points on this line.

So far we have seen that parts (i) and (ii) of Set Piece Theorem 3 generalise rather nicely to our new set $\mathbb{Q}_\infty$. But what about (iii)? In other words, can we do arithmetic in $\mathbb{Q}_\infty$? The answer is: “To a certain extent.” The operation of addition in $\mathbb{Q}$, defined in the statement of Set Piece Theorem 3, partially extends to $\mathbb{Q}_\infty$, giving us:

$$p + \infty = \infty = \infty + p, \tag{3.8}$$

for all non-zero $p \in \mathbb{Q}$. However, when used to formulate the sums $0 + \infty$ and $\infty + \infty$ it produces the “forbidden fraction” $0/0$. We can sidestep this by decreeing:

$$0 + \infty = \infty = \infty + 0, \qquad \infty + \infty = \infty, \tag{3.9}$$

on the grounds that (3.8) holds for p with very large denominator, as well as p with very large numerator. The operation of multiplication in $\mathbb{Q}$, defined in the statement of Set Piece Theorem 3, also partially extends to $\mathbb{Q}_\infty$, giving us:

$$p.\infty = \infty = \infty.p, \tag{3.10}$$

for all non-zero $p \in \mathbb{Q}$. Furthermore, it also gives:

$$\infty.\infty = \infty. \tag{3.11}$$

However when used to formulate the product $0.\infty$ it produces $0/0$, and this time no sidestepping is possible. This is because $0.q = 0$ whereas $p.\infty = \infty$ for all non-zero $p, q \in \mathbb{Q}$, so there’s no longer a “limiting value” that we can assign to $0.\infty$.

It follows from equations (3.8)–(3.11) that there are no solutions $x, y \in \mathbb{Q}_\infty$ to the equations $x + \infty = 0$ and $y.\infty = 1$. This means that ∞ has neither an additive nor a multiplicative inverse. Furthermore, equation (3.10) has the curious consequence that:

$$(-1).\infty = \infty.$$

This means that subtraction doesn’t extend to $\mathbb{Q}_\infty$. However, the rule of reciprocation does extend, resulting in:

$$\frac{1}{\infty} = \frac{0}{1} = 0,$$

and this allows division by ∞:

$$\frac{p}{\infty} = p.0 = 0,$$

for all $p \in \mathbb{Q}$. However ∞/∞ cannot be defined.

So, returning to our original question: “Why can’t we divide by zero?” the simple answer is: “We can!”. To make this possible we need to extend the set $\mathbb{Q}$, by introducing a single new point “∞”, which allows us to perform division $p/0$ for any non-zero rational number p. However it is not possible to formulate $0/0$, ∞/∞ or $0.\infty$, and we can’t fully integrate ∞ with the familiar arithmetic of the rational numbers. This means that ∞ isn’t “just another rational number”; it has a special status, and is usually referred to as the *point at infinity.*

Remarks 3.12.

1) The use of the ∞ symbol in this construction is slightly different to its use in calculus and real analysis, and care must be taken not to confuse the two. For example, in calculus ∞ and $-\infty$ are *not* the same, and we typically see them used in expressions involving limits, such as:

$$\lim_{x\to\infty} e^x = \infty, \qquad \lim_{x\to 0} \ln|x| = -\infty.$$

2) The construction of the rational projective line $\mathbb{Q}_\infty$ is a clever way to add a point at infinity to $\mathbb{Q}$, and a similar method can be used to add a point at infinity to $\mathbb{R}$, resulting in the *real projective line* $\mathbb{R}_\infty$. (The *complex projective line* $\mathbb{C}_\infty$ can be similarly constructed from $\mathbb{C}$.) It is also possible to step up one (or more) dimension(s), and construct the *real projective plane* from the Cartesian plane $\mathbb{R}^2$ (or *real projective n-space* from $\mathbb{R}^n$). We won't go into the details of this, except to say that *every* line gets its own point at infinity, which is shared with all parallel lines. Practically speaking, what this means is that every pair of parallel lines in the Cartesian plane will meet once and only once in the projective plane, at their common point at infinity, and therefore every pair of lines now has a unique point of intersection. So, moving out of the Cartesian plane into the projective plane changes the rules of Euclidean geometry, in a rather spectacular way. This new non-Euclidean geometry is called *projective geometry.* ◊

3.7.2. Construction of the integers

The starting point for our construction of $\mathbb{Q}$ was the set $\mathbb{Z}$ of all integers. These in turn may be constructed, using a similar strategy, starting from the natural numbers $\mathbb{N}$.

This time the challenge is to:

- obtain the negative integers along with 0;
- extend addition and multiplication from $\mathbb{N}$ to this larger set;
- define the "new" operation of subtraction.

If we "reverse engineer" the problem, as we did in our contstruction of $\mathbb{Q}$, we could obtain these "new" numbers as *differences* of natural numbers (see Example 1.5). This suggests taking $S = \mathbb{N} \times \mathbb{N}$, and using the ordered pair (a, b) as a model for the integer $a - b$; so the ordered pairs (a, b) with $a \leqslant b$ give us the "new" (ie. non-positive) integers that we're looking for. The problem now is that different pairs of natural numbers (a, b), (c, d) can have the same difference:

$$a - b = c - d. \tag{3.12}$$

The solution is to impose an equivalence relation on S that "lumps together" all such pairs of natural numbers. We do this by noting that equation (3.12) rearranges to:

$$a + d = b + c,$$

which is now a valid equation in $\mathbb{N}$. We therefore define a relation $\sim$ on S by:

$$(a, b) \sim (c, d) \quad \text{if and only if} \quad a + d = b + c. \tag{3.13}$$

Notice that (3.13) is the additive version of (3.6).

The routine is now rather similar to that of Set Piece Theorem 3 and the long check list of Remarks 3.11. We give a summary, without going into all the details. *(We invite you to supply these for yourself.)*

- First we need to check that $\sim$ is an equivalence relation. We then define $\mathbb{Z}$ to be the quotient set $S/\sim$. An integer is therefore an equivalence class of $\sim$.

- The equivalence classes can be visualised geometrically as subsets of the positive integer lattice in the Cartesian plane, shown (colour-coded) in Figure 3.4. Each class is a subset of a diagonal line, whose x-intercept is the integer point in $\mathbb{R}$ (when $\mathbb{R}$ is identified with the x-axis) that the class defines! This in effect allows us to view $\mathbb{Z}$ as a subset of $\mathbb{R}$. The diagonal lines are all parallel, which is a geometric manifestation of the classes being pairwise disjoint

Remark 3.13. There is a similar geometric correspondence that allows us to view the rational numbers $\mathbb{Q}$ as a subset of the real line $\mathbb{R}$. The rational number a/b is an equivalence class that may be visualised as a subset of the Cartesian plane; the elements of this class are points on the line through (a, b) and the origin (Figure 3.3). We simply take the intersection of this line with the horizontal line Y with equation $y = 1$, and then identify Y with $\mathbb{R}$ via its x-coordinate. It's easy to see that this gives us the *real* number a/b. ♢

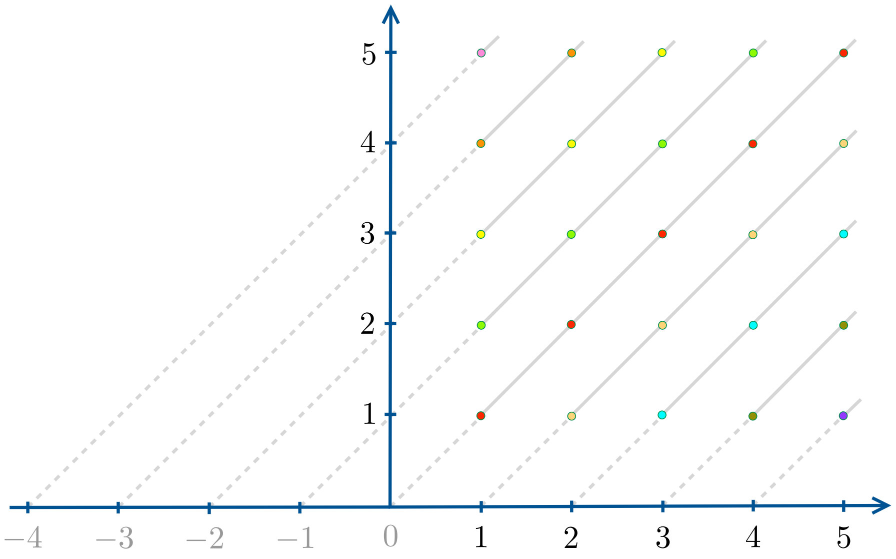


Figure 3.4.: Colour-coded equivalence classes for the integers

- Now, denoting the equivalence class of (a, b) by $[(a, b)]$, we define operations of addition and multiplication on $\mathbb{Z}$ by:

  $$[(a, b)] + [(c, d)] = [(a + c, b + d)], \qquad [(a, b)] \cdot [(c, d)] = [(ac + bd, ad + bc)].$$

  We need to check that these operations are well-defined. (The rule for multiplication may look rather strange; however it's what we would get if we interpret $[(a, b)]$ as $a - b$ and $[(c, d)]$ as $c - d$, and then multiply out the brackets.)

- We also need to check that these operations are associative, commutative and distributive.

- There is a (natural!) copy of $\mathbb{N}$ inside $\mathbb{Z}$, as the image of the one-to-one function $f : \mathbb{N} \to \mathbb{Z}$ defined:

  $$f(n) = [(1 + n, 1)], \quad \text{for all } n \in \mathbb{N}.$$

  We henceforward abbreviate $[(1 + n, 1)]$ to n, and refer to this as a *positive integer.*

- We must now show that the operations of addition and multiplication in $\mathbb{Z}$ are consistent with those of $\mathbb{N}$.

- We define $0 = [(1, 1)]$, and check that 0 behaves as it should relative to addition and multiplication in $\mathbb{Z}$.

- We should also check that 1 is a multiplicative identity.

- For any $n \in \mathbb{N}$ we define $-n = [(1, 1 + n)]$, and refer to this as a *negative integer.*

- We extend this to all elements of $\mathbb{Z}$ by defining:

  $$-[(a, b)] = (-1) \cdot [(a, b)] = [(a + 2b, 2a + b)].$$

  We then check that:

  $$[(a, b)] + (-[(a, b)]) = 0.$$

- Finally, we define subtraction in $\mathbb{Z}$ by:

  $$p - q = p + (-q),$$

  for all $p, q \in \mathbb{Z}$.

Remark 3.14. One final remark: construction of sets like $\mathbb{Q}$ and $\mathbb{Z}$ is important! Although it would be possible to axiomatise $\mathbb{Q}$ as: "A set of symbols a/b where $a, b \in \mathbb{Z}$ with $b \neq 0$, satisfying certain properties ...", there is no immediate guarantee that such a set exists.

In fact a similar problem arose with the complex numbers $\mathbb{C}$, which were first introduced in the 16-th century by Cardano[6] in a purely functional way as a formal means

[6] Girolamo Cardano (1501–1576): Italian mathematician and Renaissance man, whose book *Ars Magna* (pub. 1545) featured the use of complex numbers to solve cubic and quartic equations.

of solving cubic equations. Over the following years (centuries) the essential algebraic properties of complex numbers were written down, and used with impunity by some mathematicians, including Euler[7]. However many mathematicians remained sceptical, and it wasn't until the early nineteenth century that the now familiar geometric model of the complex plane emerged, courtesy of Wessel[8] and Argand[9]. This settled the question of their existence, and finally enabled complex numbers to take their place in mainstream mathematics. ♢

Following the construction of $\mathbb{Z}$, it would appear that the next (or first) step in the quest to rigorously define the special sets $\mathbb{N}, \mathbb{Z}, \mathbb{Q}, \mathbb{R}, \ldots$ is to construct the natural numbers $\mathbb{N}$. Kronecker[10] thought that this could be left to the Almighty (!), reputedly having said:

> "God made the integers, all else is the work of man."

Unsurprisingly, perhaps, most mathematicians, were not happy with a "deus ex machina" approach to the existence of such an important and fundamental mathematical object, and there are now several ways of dealing with $\mathbb{N}$. One of these, proposed by Peano[11], was to define $\mathbb{N}$ axiomatically. We will take a brief look at how he did this at the beginning of Chapter 4.

[7] Leonhard Euler (1707–1783): Switzerland's greatest mathematician, and one of the greatest mathematicians of all time.

[8] Caspar Wessel (1745–1818): Norwegian/Danish mathematician and cartographer, whose work in surveying and map making led him to explore the geometric properties of complex numbers. His ideas were published in 1797, but this paper went unnoticed until the late nineteenth century.

[9] Jean-Robert Argand (1768–1822): Swiss/French mathematician, whose geometric representation of the complex numbers appeared in an essay written in 1806.

[10] Leopold Kronecker (1823–1891): German mathematician, primarily interested in number theory and algebra, but well-known throughout mathematics for the "Kronecker delta" function.

[11] Giuseppe Peano (1858–1932): Italian mathematician, with interests in logic and set theory.

4. Axiomatic Mathematics (mainly Group Theory)

References.
Liebeck: Chapter 20 (for permutations and the symmetric group)
Franklin and Daoud: Chapter 15 (for axioms, Peano axioms).
Episodes 11.5–18 of video lectures.

In this Chapter we're going to explore the idea of developing a mathematical theory from a set of *axioms:* elementary statements that are considered to be "self evidently" true (and therefore don't require proof), from which all other true statements (theorems, propositions, corollaries, etc.) in the theory can be proved. This point of view is very important in mathematics (especially pure maths), and has a pedigree going right back to the ancient Greeks: Euclid's "Elements" [15] is famous for being an axiomatic development of (Euclidean) geometry. We've already referred to the axiomatic approach on several occasions (notably, set theory). However our main example here will be something new: "group theory". But before getting into that we'll touch on another significant example.

4.1. The Peano axioms

The Peano axioms (which were in fact attributed by Peano to Dedekind[1]) provide an axiomatic definition of the natural numbers $\mathbb{N}$. There are many ways to express Peano's axioms, but we're going to formulate them in terms which fit in with how we've been working during the course; then after the formal statement of each one we'll give an informal paraphrase.

There is also the thorny issue of whether or not $\mathbb{N}$ should include 0. The Peano axioms can easily be "tweaked" to allow this, but we're not going to do so, which is the convention we've been observing throughout the course so far, and is in line with most current practice. However there are many mathematicians who consider 0 to be "natural" (!), so this is something we need to watch out for. Of course we need to have 0 as soon as

[1] Richard Dedekind (1861–1916): German mathematician, best known for his construction of the real numbers from the rationals.

we want to do arithmetic, and it duly appears when we construct the integers $\mathbb{Z}$ from $\mathbb{N} = \{1, 2, 3, \dots\}$, as we saw in Section 3.7.2.

Definition 4.1 (Peano axioms)**.**
The set $\mathbb{N}$ of *natural numbers* is characterised by the following five properties:

PA1) $1 \in \mathbb{N}$.

"1 is a natural number."

PA2) There exists a function $S\colon \mathbb{N} \to \mathbb{N}$.

"Every natural number n has a successor $S(n)$."

PA3) $1 \notin S(\mathbb{N})$.

"1 is not the successor of any natural number."

PA4) S is one-to-one.

"No two natural numbers have the same successor."

PA5) Suppose A is set satisfying

- $1 \in A$;
- for all $n \in \mathbb{N}$, if $n \in A$ then $S(n) \in A$.

Then $\mathbb{N} \subseteq A$.

"Every inductive set contains all the natural numbers." ♦

Remarks 4.1 (Peano axioms).

1) Axiom (PA1) says nothing more than the set $\mathbb{N}$ is non-empty. The element is denoted "1" because it will ultimately play the rôle of the natural number 1; but we could have chosen any symbol we like for it!

2) Axioms (PA2) and (PA3) say that $\mathbb{N}$ contains another element: $S(1)$. Axiom (PA2) says that $S(1)$ is an element of $\mathbb{N}$, and (PA3) says that $S(1) \neq 1$. We could decide to denote $S(1)$ by the symbol "2".

3) Axiom (PA4) says that we can repeat this process indefinitely. So, for example, the element $S(2)$ of $\mathbb{N}$ is different from $2 = S(1)$, because $2 \neq 1$ and different elements have different successors. Furthermore $S(2) \neq 1$ by (PA3). It would be natural (!) to denote $S(2)$ by the symbol "3".

4) Axiom (PA5), which is usually called the *Principle of Induction,* says: “That’s all folks!” More precisely, if we denote by $A \subseteq \mathbb{N}$ the subset containing all elements of the form $S \circ \cdots \circ S(1)$ (ie. all finite compositions of S applied to 1) then A is clearly an inductive set, so $\mathbb{N} \subseteq A$ by (PA5). Therefore $A = \mathbb{N}$ by the Principle of Mutual Containment.

5) In summary, the Peano axioms are simply a precise way of saying: “If we want to get from 1 to any other natural number then we can do so by applying the successor function suitably many times”. In particular, this endows $\mathbb{N}$ with the fundamental properties of being infinite yet countable, which of course we can then transfer to other sets (Chapter 1, Section 1.3).

6) The Principle of Induction also tells us that $\mathbb{N}$ is the *smallest* inductive set. It explains why we always find a copy of $\mathbb{N}$ inside our other special sets $\mathbb{Z}, \mathbb{Q}, \mathbb{R}$ etc. which are also inductive. It’s also a strong indication of its uniqueness.

7) Indeed, it can be shown (although we won’t) that there is “essentially” only one set satisfying the Peano axioms; thus $\mathbb{N}$ is unique.

8) A set satisfying the Peano axioms (in other words, a “model” for this axiomatic system) can be constructed using set theory. What has become the “standard model” was first presented by von Neumann in the early twentieth century.

9) The Peano axioms can also be used to introduce the usual operations of arithmetic. For example, addition is defined using the successor function via the basic stipulation that for all $n \in \mathbb{N}$:

$$\begin{aligned} n + 1 &= S(n), \\ n + 2 &= S(S(n)) = S \circ S(n), \text{ etc.} \end{aligned}$$

With patience, all the laws of additive and multiplicative arithmetic can be established, along with the “well-ordering” relation $<$. $\Diamond$

4.1.1. Proof by induction, revisited

With the Peano axioms and Remarks 4.1 in hand, we can “prove” that proof by induction “works” (Chapter 2, Section 2.4.5), as follows.

Let $P(n)$ be a statement, for each $n \in \mathbb{N}$. Suppose we want to show that $P(n)$ is true for all $n \geqslant n_0$. Define a set:

$$A = \{n \in \mathbb{N} : P(n + n_0 - 1) \text{ is true}\}.$$

Clearly $A \subseteq \mathbb{N}$. Referring to the statement of Peano axiom (PA5), the condition “$1 \in A$” then means:

$P(n_0)$ is true,

(aka. the "base case"), whereas the statement "if $n \in A$ then $S(n) \in A$" translates to:

$P(n)$ is true implies $P(n+1)$ is true, for all $n \geqslant n_0$,

(aka. the "induction step"). Therefore if both the base case and the induction step are true then A is an inductive set, and the Principle of Induction tells us that $\mathbb{N} \subseteq A$. Combining these two inclusions via the Principle of Mutual Containment, we conclude that $A = \mathbb{N}$, which translates to:

$P(n)$ is true for all $n \geqslant n_0$.

Remark 4.2. The Principle of Induction can also be used to prove the well-ordering principle (see Remark 1.24); we won't give the details, but it's an inductive proof by contradiction! ◊

4.2. The group axioms

Groups encode the idea of "symmetry" in mathematical terms. Symmetry is perhaps most familiar to us (and everyone else) as a geometric property. However, the underlying idea is in fact much more general and appears in many different guises, across maths (both pure and applied) and science, so it is useful to make the notion *abstract* in order to study it in its own right. This is the essence of "group theory". The things we prove in this abstract setting are then applicable *whenever* a group is present, in *whatever* setting. This saves us the effort of concocting a plethora of different proofs for what is essentially the same thing, whilst at the same time giving us a clearer picture of what's really going on.

We will launch straight in to the fully abstract axiomatic definition of a "group". This will probably seem rather strange at first sight, but as we consider various examples (from Section 4.3 onwards) should begin to make sense.

Definition 4.2 (Axioms of a group)**.**
A *group* $(G, \cdot)$ consists of a set G together with a *multiplication function:*

$$\begin{aligned} \mu\colon G \times G &\to G; \\ (x, y) &\mapsto \mu(x, y) = x \cdot y, \end{aligned}$$

satisfying the following three axioms:

G1) **Associativity:** for all $x, y, z \in G$ we have:

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

G2) **Identity:** there exists an element $e \in G$ such that for all $x \in G$:

$$e \cdot x = x \cdot e = x.$$

G3) **Inverses:** for every $x \in G$ there exists an element $y \in G$ satisfying:

$$x \cdot y = e = y \cdot x. \qquad ♦$$

Remarks 4.3 (Group axioms).

1) Notice that the third axiom (G3) doesn't make sense without the second (G2), and therefore has to come after it!

2) We have used the word "multiplication" to describe the rule "μ" for combining pairs of elements to get a new element, because it bears similarities to the familiar operation of multiplication of real numbers. Indeed, this is a basic example of a group; see Section 4.3.4 below. However, sometimes group multiplication can be something completely different, and calling it "multiplication" can be confusing, possibly even deceptive; for example, see Sections 4.3.1 and 4.3.2. For this reason the alternative terminology *binary operation* is sometimes used.

3) We gave the multiplication rule a (phonetically engineered) symbol "μ", but we'll seldom use it! We'll usually just use the "$\cdot$" symbol (or similar) to denote the multiplication on a group. Once we get used to dealing with groups, it is very common to even drop this multiplication symbol too, and just write xy for $x \cdot y$. Since we're not quite that sophisticated yet, we'll keep the "$\cdot$" for now.

4) Axiom (G3) should *not* be taken to mean that group multiplication is commutative:

$$x \cdot y = y \cdot x, \quad \text{for all } x, y \in G.$$

   This is one property of ordinary multiplication that we *don't* want to replicate in general: if we asked for all groups to be commutative not only would there be very few examples, but the mathematical theory would become virtually trivial! Nevertheless, commutative groups have an important rôle to play, and are often referred to as *abelian*[2] *groups.*

5) The axiomatic definition of a group didn't emerge until the end of the nineteenth century, although various examples were known well before then. It is usually attributed to von Dyck[3] [39]. ◊

We can immediately prove a surprising number of elementary things about groups using the axioms. Some of these resemble the kind of properties we're familiar with from elementary algebra and arithmetic.

[2] Named after Norwegian mathematician Niels Abel (1802–1829), whose most famous work was to show that the familiar "quadratic formula" cannot be generalised to find the roots of quintic polynomials.

[3] Walther von Dyck (1856–1934): German mathematician, who worked in group theory, topology and geometry.

Proposition 4.1 (Elementary properties of groups)**.**
Let $(G, \cdot)$ be a group.

i) *The identity element is unique.*

ii) *Each $x \in G$ has a unique inverse, which we write as x^{-1} for any $x \in G$.*

iii) *For each $x \in G$ we have $(x^{-1})^{-1} = x$.*

iv) *For all $x, y \in G$ we have $(x \cdot y)^{-1} = y^{-1} \cdot x^{-1}$.*

v) *If $x \cdot y = x \cdot z$ for $x, y, z \in G$, then $y = z$ (left cancellation). Similarly, if $y \cdot x = z \cdot x$, then $y = z$ (right cancellation).*

Proof. To prove the first two properties, which are uniqueness statements, we use the contrapositive (cf. Proposition 2.3). The remaining properties will be proved directly.

i) If $e' \in G$ is another element satisfying the identity axiom then we have:

$$e = e \cdot e' = e'.$$

ii) If y and z are two inverses for $x \in G$ then we can write:

$$y = y \cdot e = y \cdot (x \cdot z) = (y \cdot x) \cdot z = e \cdot z = z,$$

using all three axioms along the way.

iii) By the definition (G3) of x^{-1} we have:

$$x \cdot x^{-1} = e = x^{-1} \cdot x.$$

But this is also the condition for x to be inverse to x^{-1}. Hence by uniqueness of the inverse we deduce that $(x^{-1})^{-1} = x$.

iv) Applying all three axioms, we have:

$$\begin{aligned}(y^{-1} \cdot x^{-1}) \cdot (x \cdot y) &= ((y^{-1} \cdot x^{-1}) \cdot x) \cdot y \\ &= (y^{-1} \cdot (x^{-1} \cdot x)) \cdot y \\ &= (y^{-1} \cdot e) \cdot y \\ &= y^{-1} \cdot y = e.\end{aligned}$$

A similar argument shows that:

$$(x \cdot y) \cdot (y^{-1} \cdot x^{-1}) = e.$$

We have shown that $y^{-1} \cdot x^{-1}$ satisfies the condition to be inverse to $x \cdot y$. Hence by uniqueness of the inverse we deduce that $(x \cdot y)^{-1} = y^{-1} \cdot x^{-1}$.

v) If $x \cdot y = x \cdot z$ then "multiplying" this equation on the left by x^{-1} and using all three axioms gives:

$$z = e \cdot z = (x^{-1} \cdot x) \cdot z = x^{-1} \cdot (x \cdot z) = x^{-1} \cdot (x \cdot y) = (x^{-1} \cdot x) \cdot y = e \cdot y = y.$$

The proof of right cancellation is identical. ■

Remarks 4.4 (Elementary properties of groups).

1) Having established its uniqueness in Proposition 4.1 (i), we refer to e as *the identity element* of G.

2) It follows from axiom (G2) that $e \cdot e = e$. Hence by axiom (G3) and Proposition 4.1 (ii) we deduce that $e^{-1} = e$, as we might have expected.

3) Proposition 4.1 (iv) may be paraphrased:

   The inverse of a product is the product of the inverses, reversed.

   The reversal of order is important, since group multiplication is not in general commutative; see Remark 4.3 (4).

4) The cancellation laws look as if we're performing "division by x". However this viewpoint is unhelpful, since groups do not have a separate operation of division. Instead, it is better, and more accurate, to think of "multiplication by x^{-1}".

5) If we have established that $(G, \cdot)$ is a group, then the practical task of finding inverse elements becomes slightly easier, because we then need only check *one* of the equations from axiom (G3). For example, if $x \cdot y = e$ then by writing $e = x \cdot x^{-1}$ and applying left cancellation we get $y = x^{-1}$. Similarly, if $y \cdot x = e$ then $y = x^{-1}$ by right cancellation. ♢

4.3. Examples of groups

We hinted at the beginning of Section 4.2 that there are many different mathematical structures that satisfy the group axioms, in contrast to the Peano axioms where there is essentially only one. Here are a few examples.

4.3.1. Integers

Let $G = \mathbb{Z}$, and let the multiplication operation be addition (!). (It's clear why "binary operation" would be more appropriate here!) Thus:

$$\mu(a, b) = a + b,$$

for all $a, b \in \mathbb{Z}$. It follows that the identity element e must be 0, and the inverse of a is therefore $-a$. (Using the general x^{-1} notation for inverses would clearly be inappropriate!) Furthermore addition satisfies the associative law:

$$(a + b) + c = a + (b + c), \quad \text{for all } a, b, c \in \mathbb{Z}.$$

So the group axioms (G1)–(G3) are satisfied, therefore $(\mathbb{Z}, +)$ is a group.

In this case the binary operation is also commutative:

$$a + b = b + a, \quad \text{for all } a, b \in \mathbb{Z}.$$

So the group $(\mathbb{Z}, +)$ is abelian.

4.3.2. Modular integers

Let $G = \mathbb{Z}_n$, the integers modulo n (Section 3.5). Then G is a group under the binary operation of addition modulo n, defined by equation (3.3):

$$\mu([a]_n, [b]_n) = [a]_n + [b]_n = [a + b]_n.$$

We'll check this carefully. Suppose $a, b, c \in \mathbb{Z}$.

Associativity. We have:

$$\begin{aligned} ([a]_n + [b]_n) + [c]_n &= [a + b]_n + [c]_n \\ &= [(a + b) + c]_n \\ &= [a + (b + c)]_n, \quad \text{by associativity of addition in } \mathbb{Z} \\ &= [a]_n + [b + c]_n \\ &= [a]_n + ([b]_n + [c]_n). \end{aligned}$$

Identity. We have:

$$[0]_n + [a]_n = [0 + a]_n = [a]_n = [a + 0]_n = [a]_n + [0]_n.$$

Therefore the identity element is $[0]_n$.

Inverses. We have:

$$[a]_n + [-a]_n = [a + (-a)]_n = [0]_n = [(-a) + a]_n = [-a]_n + [a]_n.$$

Therefore the inverse of $[a]_n$ is $[-a]_n$. (Again, the $*^{-1}$ notation would be inappropriate in this case.)

Notice that the group properties for $\mathbb{Z}_n$ are all inherited from those of $\mathbb{Z}$. In particular, the binary operation is again commutative, so $(\mathbb{Z}_n, +)$ is also an abelian group.

4.3.3. Non-zero modular integers with prime modulus

Suppose p is a positive prime number and $G = \mathbb{Z}_p^*$, the set of non-zero congruence classes modulo p:

$$\mathbb{Z}_p^* = \{\bar{1}, \bar{2}, \dots, \overline{p - 1}\},$$

where we have switched to the "bar" notation, for simplicity. Then, perhaps surprisingly, G is a group under the binary operation of multiplication modulo p, defined by equation (3.4):

$$\mu(\bar{a}, \bar{b}) = \bar{a}\bar{b} = \overline{ab}.$$

Again, we'll check this carefully. Suppose $a, b, c \in \mathbb{Z}$, and none is a multiple of p (which ensures $\bar{a}, \bar{b}, \bar{c} \in G$).

Associativity. We have:

$$\begin{aligned}(\bar{a}\bar{b})\bar{c} &= \overline{ab}\,\bar{c} = \overline{(ab)c} \\ &= \overline{a(bc)}, \quad \text{by associativity of multiplication in } \mathbb{Z} \\ &= \bar{a}\,\overline{bc} \\ &= \bar{a}(\bar{b}\bar{c}).\end{aligned}$$

Identity. We have:

$$\bar{1}\bar{a} = \overline{1a} = \bar{a} = \overline{a1} = \bar{a}\bar{1}.$$

Therefore the identity element is $\bar{1}$.

Inverses. Since p is prime (a fact we haven't used so far) and a is *not* a multiple of p, the highest common factor of a and p is 1 (ie. a and p are *coprime,* or *relatively prime*). We now appeal to a fact from elementary number theory, which is that it's possible to express the highest common factor of two integers as a "linear combination" of those integers (see Appendix A.2), which in this case means there exist integers $b, c \in \mathbb{Z}$ such that the following equation holds:

$$ab + pc = 1.$$

This is sometimes referred to as *Bézout's identity*[4]. Taking congruence classes modulo p, and applying the definitions of modular addition and multiplication:

$$\bar{1} = \overline{ab + pc} = \overline{ab} + \overline{pc} = \bar{a}\bar{b} + \bar{p}\bar{c} = \bar{a}\bar{b} + \bar{0}\bar{c} = \bar{a}\bar{b}.$$

Similarly $\bar{b}\bar{a} = \bar{1}$. Therefore the inverse of $\bar{a}$ is $\bar{b}$:

$$(\bar{a})^{-1} = \bar{b}.$$

Remarks 4.5.

1) Notice that the first two group axioms are inherited from multiplication in $\mathbb{Z}$. However the final property is novel, since it has no counterpart in $\mathbb{Z}$ (or $\mathbb{Z}^*$). It's another illustration of how quotient sets can often bring something new to the table (cf. Set Piece Theorem 3).

[4] Étienne Bézout (1730–1783): French mathematician, whose primary interest was the solution of simultaneous algebraic equations.

2) It's not hard to check that multiplication in $\mathbb{Z}_p^*$ is commutative (which again follows from the commutativity of multiplication in $\mathbb{Z}$); so $(\mathbb{Z}_p^*, \cdot)$ is an abelian group. In fact, by combining this group with the additive group $(\mathbb{Z}_p, +)$ we've created another example of a *field,* like $\mathbb{Q}$ but with only finitely many elements! (More about this in Year 2.) $\Diamond$

The mechanism for constructing inverses in $\mathbb{Z}_p^*$ relied on Bézout's identity. This may be derived algorithmically, using *Euclidean algorithm,* which we describe in detail in Appendix A.2. In the following example we show how this works in practice.

Example 4.1 (Inverses in the group $\mathbb{Z}_{11}^*$)**.**
We consider the multiplicative group $\mathbb{Z}_{11}^*$, which has 10 elements:

$$\bar{1}, \quad \bar{2}, \quad \bar{3}, \quad \bar{4}, \quad \bar{5}, \quad \bar{6}, \quad \bar{7}, \quad \bar{8}, \quad \bar{9}, \quad \overline{10}.$$

Apart from $\bar{1}$, whose inverse is $\bar{1}$ (see Remark 4.4 (2)), it's not clear at first sight which of the remaining nine elements is inverse to which! Intriguingly, having an odd number of elements to account for means that it's not possible to pair off one with another. The logical conclusion, although inexorable, is nevertheless quite surprising: there must be (at least) one element that is its own inverse! To see how everything works out we'll have to analyse each case separately, bearing in mind Proposition 4.1 (iii) which allows us to reduce our workload by (up to) half.

First, to find the inverse of $\bar{2}$ we first divide 11 by 2:

$$11 = 5.2 + 1.$$

The remainder is 1, which is what we're after, so we rearrange the equation, take residues modulo 11 and do a little modular arithmetic:

$$\bar{1} = -\bar{5}\,\bar{2} + \overline{11} = \overline{-5}\,\bar{2} + \bar{0} = \bar{6}\,\bar{2}.$$

Therefore $\bar{2}^{-1} = \bar{6}$. Proposition 4.1 (iii) then implies $\bar{6}^{-1} = \bar{2}$, a.

Moving on to $\bar{3}$, we first divide 11 by 3 which gives:

$$11 = 3.3 + 2.$$

This time the remainder is not 1, so we have to perform another division:

$$3 = 1.2 + 1.$$

We rearrange this and use the first equation to eliminate the unwanted "2":

$$1 = 3 - 2 = 4.3 - 11.$$

This is Bézout's identity. Taking residues modulo 11 gives us:

$$\bar{1} = \bar{4}\,\bar{3} - \overline{11} = \bar{4}\,\bar{3} - \bar{0} = \bar{4}\,\bar{3}.$$

Therefore $\bar{3}^{-1} = \bar{4}$, hence $\bar{4}^{-1} = \bar{3}$.

We now consider $\bar{5}$, for which we have the identity:

$$11 = 2.5 + 1,$$

hence:

$$\bar{1} = \overline{11} - \bar{2}\,\bar{5} = \bar{9}\,\bar{5}.$$

Therefore $\bar{5}^{-1} = \bar{9}$, and $\bar{9}^{-1} = \bar{5}$.

We can skip $\bar{6}$ and move on to $\bar{7}$, for which we need to perform three divisions:

$$11 = 1.7 + 4, \qquad 7 = 1.4 + 3, \qquad 4 = 1.3 + 1.$$

Rearranging the last of these equations, and substituting into it the other two yields Bézout's identity:

$$1 = 4 - 3 = 2.4 - 7 = 2.11 - 3.7.$$

Therefore:

$$\bar{1} = \bar{2}\,\overline{11} - \bar{3}\,\bar{7} = \bar{8}\,\bar{7}.$$

Hence $\bar{7}^{-1} = \bar{8}$, and $\bar{8}^{-1} = \bar{7}$.

There is now only one further inverse to determine, $\overline{10}^{-1}$, which by process of elimination can only be $\overline{10}$. At first sight this appears slightly crazy, but is easy enough to check using modular arithmetic:

$$\overline{10}\,\overline{10} = \overline{100} = \overline{99} + \bar{1} = \bar{1}.$$

Therefore $\overline{10}^{-1} = \overline{10}$.

To summarise:

$\bar{a}$	$\bar{1}$	$\bar{2}$	$\bar{3}$	$\bar{4}$	$\bar{5}$	$\bar{6}$	$\bar{7}$	$\bar{8}$	$\bar{9}$	$\overline{10}$
$\bar{a}^{-1}$	$\bar{1}$	$\bar{6}$	$\bar{4}$	$\bar{3}$	$\bar{9}$	$\bar{2}$	$\bar{8}$	$\bar{7}$	$\bar{5}$	$\overline{10}$

Table 4.1.: Inverses in $\mathbb{Z}_{11}^*$

In Example 4.3 we revisit these results from a slightly different point of view. □

4.3.4. Non-zero rational or real numbers

Let $G = \mathbb{Q}^*$ or $\mathbb{R}^*$, the set of *non-zero* rational or real numbers, and let the binary operation be (true) multiplication:

$$\mu(x, y) = xy,$$

for all rational or real numbers $x, y \neq 0$. It's easy to check that the group axioms are satisfied, this time with $e = 1$ and $x^{-1} = 1/x$ (which of course is also denoted x^{-1}). Again, this group is abelian.

4.3.5. The general linear group

The previous example can be generalised in the following way. Let G be the set of all $n \times n$ matrices with non-zero determinant, which we usually denote by $GL(n)$:

$$GL(n) = \{A : A \text{ is a } n \times n \text{ matrix}, \ \det(A) \neq 0\}.$$

(If you haven't met matrices and determinants before don't worry. It's the only time we will be mentioning them in the main part of the course, so you can just skip this example, and maybe return to it at some point in the future if you want to. There is an in-depth discussion of determinants in Appendix A.6.) Let the binary operation be matrix multiplication:

$$\mu(A, B) = AB.$$

For this to be a valid binary operation on $GL(n)$ we need to show that $\det(AB) \neq 0$. This follows from the "multiplicative property" of determinants:

$$\det(AB) = \det(A)\det(B),$$

which holds for *all* $n \times n$ matrices. (This is a non-trivial equation, requiring proof, which we'll leave to the Algebra course; alternatively, see Appendix A.6.)

Matrix multiplication is associative (another non-trivial fact, which again we'll leave to the Algebra course; alternatively, see Appendix A.4). Also, it's easy to see that the $n \times n$ *identity matrix* $\mathbb{I}_n$ defined:

$$\mathbb{I}_n = \begin{pmatrix} 1 & 0 & \dots & 0 \\ 0 & 1 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & 1 \end{pmatrix} \tag{4.1}$$

satisfies:

$$A\mathbb{I}_n = A = \mathbb{I}_n A,$$

which is precisely the property of the group identity element e. Furthermore, since $\det(A) \neq 0$ there exists a unique *inverse matrix* A^{-1} satisfying:

$$AA^{-1} = \mathbb{I}_n = A^{-1}A.$$

(Again, we'll leave the justification of this to the Algebra course.) This is precisely the property that A^{-1} is the group inverse element of A. Thus $GL(n)$ is a group; it's called the *general linear group.*

If $n = 1$ then we recover our previous example: $GL(1) = \mathbb{R}^*$. However, it's interesting to note that for $n > 1$ matrix multiplication is *not* commutative; so here we have a "natural" example of a non-abelian group.

4.3.6. Unit modular integers

We saw in Section 4.3.3 that the set $\mathbb{Z}_p^*$ of non-zero residue classes modulo a prime p is a group under modular multiplication. More generally, we consider the set $\mathbb{Z}_n^*$ for a general natural number n:

$$\mathbb{Z}_n^* = \{\bar{1}, \bar{2}, \ldots, \overline{n-1}\}.$$

If n isn't prime then $\mathbb{Z}_n^*$ isn't a group under modular multiplication. To see why, consider for example $n = 8$, and notice that:

$$\bar{6}\bar{4} = \overline{24} = \bar{3}\bar{8} = \bar{0}.$$

Elements with this extraordinary property are called *zero divisors,* and their presence shows that multiplication mod 8 is *not* a binary operation on $\mathbb{Z}_8^*$ (since $\bar{0}$ is not an element of $\mathbb{Z}_8^*$). On the other hand:

$$\bar{5}\bar{5} = \overline{25} = \bar{3}\bar{8} + \bar{1} = \bar{1},$$

so $\bar{5}$ has a multiplicative inverse. In general, a modular integer $\bar{a} \in \mathbb{Z}_n$ is said to be a *unit* if there exists another modular integer $\bar{b}$ such that $\bar{a}\bar{b} = \bar{1}$. We would like to know which classes are zero divisors, and which are units.

Proposition 4.2 (Units of $\mathbb{Z}_n$)**.**
The units of $\mathbb{Z}_n$ are the elements $\bar{a} \in \mathbb{Z}_n$ where a is coprime to n.

Proof. Direct and straightforward, using Bézout's identity for the greatest common divisor.

We have:

$$\begin{aligned} \bar{a}\bar{b} = \bar{1} &\iff \overline{ab} = \bar{1} \\ &\iff ab - 1 = kn, \quad \text{for some } k \in \mathbb{Z}, \end{aligned}$$

by definition of congruence mod n (Section 3.3)

$$\iff ab - kn = 1 \iff \gcd(a, n) = 1. \qquad \blacksquare$$

To stand any chance of producing a group using modular multiplication (in particular, the existence of inverses) we restrict attention to the subset of units:

$$\begin{aligned} \mathbb{Z}_n^\times &= \{\text{units of } \mathbb{Z}_n\} \\ &= \{\bar{a} : a \text{ coprime to } n\}, \end{aligned}$$

by Proposition 4.2. Notice the subtle change of notation! If $n = p$ is prime, then $\mathbb{Z}_p^\times = \mathbb{Z}_p^*$ since *all* non-zero residues are coprime to p; otherwise $\mathbb{Z}_n^\times \subset \mathbb{Z}_n^*$ is a proper subset.

We claim that $\mathbb{Z}_n^\times$ is a group under multiplication modulo n. However, first we need to check that this is a valid binary operation; in other words, multiplying two units of $\mathbb{Z}_n$ produces another unit. Now, if $\bar{a}, \bar{x} \in \mathbb{Z}_n^\times$ then by Bézout's identity we can write:

$$ab + cn = 1 = xy + zn,$$

for $b, c, y, z \in \mathbb{Z}$. Hence:

$$1 = (ab + nc)(xy + nz) = (by)ax + (abz + cxy + czn)n,$$

which shows that ax and n are coprime (because any common factor of ax and n must divide 1); hence $\bar{a}\bar{x} \in \mathbb{Z}_n^\times$. The three group axioms can now be checked in exactly the same way as for $\mathbb{Z}_p^*$ in Section 4.3.3. The group $\mathbb{Z}_n^\times$ is (once again) abelian.

Example 4.2 (The group $\mathbb{Z}_8^\times$)**.**
As an example, take $n = 8$. Then:

$$\mathbb{Z}_8^\times = \{\bar{1}, \bar{3}, \bar{5}, \bar{7}\}.$$

We can record the results of group multiplication in tabular form, the "group multiplication table":

	$\bar{1}$	$\bar{3}$	$\bar{5}$	$\bar{7}$
$\bar{1}$	$\bar{1}$	$\bar{3}$	$\bar{5}$	$\bar{7}$
$\bar{3}$	$\bar{3}$	$\bar{1}$	$\bar{7}$	$\bar{5}$
$\bar{5}$	$\bar{5}$	$\bar{7}$	$\bar{1}$	$\bar{3}$
$\bar{7}$	$\bar{7}$	$\bar{5}$	$\bar{3}$	$\bar{1}$

Table 4.2.: Multiplication table for $\mathbb{Z}_8^\times$

Notice that each element appears precisely once in every row and column of the table. *(This may remind you of Sudoku!)* The table is also symmetric about the main diagonal, which indicates that the group is abelian. *(Can you see why?)* We can read off inverses by simply looking for the appearances of the identity element $\bar{1}$ in the table. In this example these are all on the main diagonal, which tells us that $\bar{a}^2 = \bar{1}$ so $\bar{a}^{-1} = \bar{a}$ for all elements $\bar{a}$. Group elements with this property are called *involutions,* and are not typical! This particular 4-element group is called the *Klein*[5] *4-group.* □

[5] Felix Klein (1849–1925): German mathematician, primarily interested in the relationship between group theory and geometry, expressed in his influential "Erlangen Program" [25].

Remark 4.6 (Euler's totient function).
Whilst working on number theory, Euler introduced in [16] what is now known as the *Euler totient*[6] *function* $\varphi\colon \mathbb{N} \to \mathbb{N}$ defined:

$$\varphi(n) = |\{n \in \mathbb{N} : m < n,\ \gcd(m,n) = 1\}|;$$

in other words, $\varphi(n)$ is the number of natural numbers less than n (including 1) that are coprime to n. By Proposition 4.2 this is precisely the cardinality of $\mathbb{Z}_n^\times$:

$$|\mathbb{Z}_n^\times| = \varphi(n).$$

Euler showed that φ has certain nice properties, notably:

$$\varphi(ab) = \varphi(a)\varphi(b), \quad \text{for all coprime } a, b,$$

and:

$$\varphi(p^m) = p^{m-1}(p-1), \quad \text{for all primes } p.$$

(We will leave the proofs of these to the Introduction to Number Theory module in Year 2.) These help us to evaluate $\varphi(n)$ without having to laboriously calculate a multitude of greatest common divisors. For example:

$$\varphi(8) = \varphi(2^3) = 2^2(2-1) = 4,$$

which tells us that $\mathbb{Z}_8^\times$ has 4 elements, without having to list them! ♢

The only non-commutative group so far to appear in our catalogue of examples happens to be infinite: it's the general linear group of matrices from Section 4.3.5. However, there are many examples of *finite* non-abelian groups; so many, in fact, that group theorists still don't quite know how to classify them all! Indeed, the groups that really started the group theory ball rolling are non-abelian and finite. We'll meet them in Section 4.5 below, after a brief interlude to catch up with some general group theory.

4.4. Orders of groups and their elements

The word "order" is used in group theory in two seemingly quite different ways. Here's the first.

Definition 4.3 (Order of a group)**.**
A group $(G,\cdot)$ has *finite order* if G is a finite set. The *order of G* is then defined to be $|G|$; ie. the number of elements in G (see Definition 1.8). ♦

For the second we'll need to define exponentiation in groups.

[6] From the Latin *tot*, meaning "thus many"; hence the phrase "tot up". The terminology was not used by Euler, but introduced later by the English mathemetician James Sylvester (1814–1897).

Definition 4.4 (Exponentiation in groups)**.**
Suppose $(G, \cdot)$ is a group, and $a \in G$.

- For any $n \in \mathbb{N}$ we denote by a^n the n-fold product of a with itself:

  $$a^n = a \cdot a \cdots a \quad (n \text{ times}),$$

  and call this the *n-th power* of a. Then for all $m, n \in \mathbb{N}$ we have the *elementary group exponentiation laws:*

  $$a^{m+n} = a^m \cdot a^n \quad \text{and} \quad a^{mn} = (a^m)^n = (a^n)^m. \tag{4.2}$$

- We also define $a^0 = e$, and $a^{-n} = (a^{-1})^n$. Then we have the *extended group exponentiation laws:*

  $$a^{k+\ell} = a^k \cdot a^\ell \quad \text{and} \quad a^{k\ell} = (a^k)^\ell = (a^\ell)^k, \tag{4.3}$$

  for all $k, \ell \in \mathbb{Z}$. ♦

Remarks 4.7.

1) The elementary laws of exponentiation (4.2) follow directly from the definition of positive powers of group elements; they are simply ways of inserting brackets into a string of a's. However the extended laws (4.3) are more subtle; for example, they include as a special case:

   $$(a^n)^{-1} = (a^{-1})^n.$$

   These therefore deserve and require some proof. *(Can you put together a proof? This will require some case-by-case analysis: a proof by exhaustion!)*

2) We're familiar with these rules of exponentiation if a is a rational or real number. The beauty of working with groups is that we see the precise algebraic properties from which the rules are derived. ♢

With this under our belt, we now come to the second use of the word "order".

Definition 4.5 (Order of a group element)**.**
An element $a \in G$ has *finite order* if $a^r = e$ for some $r \in \mathbb{N}$. The *order of a* is then the *least* such r, and is denoted $o(a)$. ♦

Remarks 4.8 (Orders of group elements).

1) If $o(a) = 1$ then it follows immediately from Definition 4.5 that $a = e$ (by the uniqueness of the identity element).

2) If $o(a) = 2$ then rearranging the equation $a^2 = e$ (by multiplying both sides by a^{-1}) yields $a^{-1} = a$. We say that a is *self-inverse,* or an *involution.* We have already seen examples of this, such as all the non-identity elements of $\mathbb{Z}_8^\times$. It is nevertheless still rather surprising that group elements can have this property!

3) More generally, and perhaps only marginally less surprising: if a group element a has finite order then the inverse element a^{-1} is a power of a. For, if $o(a) = n$ then we simply reorganise the equation $a^n = e$ to get (for example):

$$a \cdot a^{n-1} = e,$$

from which (by "multiplying" both sides of the equation on the left by a^{-1}, or writing $e = a \cdot a^{-1}$ and applying left cancellation) it follows that:

$$a^{n-1} = a^{-1}.$$

We will put this to use in Example 4.3 below.

4) We say a has *infinite order* if a doesn't have finite order. This simply means that $a^r \neq e$ for *all* $r \in \mathbb{N}$. In this case we are allowed to write $o(a) = \infty$; the use of the "∞" symbol here is purely symbolic! ◊

Definitions 4.3 and 4.5 lack any immediately obvious similarity; indeed Definition 4.3 doesn't even involve the group operation! Nevertheless, the orders of groups and their elements are closely related. We'll see precisely what this relation is in Section 4.6.1 (Proposition 4.11), but there is already a hint of it in part (iii) of our next result.

Proposition 4.3 (Orders of group elements)**.**
Suppose $(G, \cdot)$ is a group, and $a \in G$ is any element.

i) *We have $a^m = e$ for $m \in \mathbb{N}$ if and only if a has finite order and m is a multiple of $o(a)$.*

ii) *We have $a^m = a^p$ for distinct $m, p \geqslant 0$ if and only if a has finite order and $m \equiv p \pmod{n}$ where $n = o(a)$.*

iii) *If a has finite order then so does a^{-1}, and $o(a^{-1}) = o(a)$.*

iv) *If G has finite order then a has finite order.*

Note. In (ii) we are using the classical notation of congruence modulo m (see Section 3.3) simply as shorthand for the statement: "m and p differ by a multiple of n". Notice that (i) is the special case $p = 0$ of (ii). ◊

Proof. This is essentially a direct proof, with a "mini exhaustion" in (i), and a "mini contradiction" in (iii). Both (i) and (ii) are "if and only if" statements, so we need to prove both forward and reverse implications.

(i) ($\Rightarrow$) If $a^m = e$ then it follows from Definition 4.5 that a has finite order; say $o(a) = n$. It also follows from Definition 4.5 that $m \geqslant n$, so we can write $m = qn + r$ where $q \in \mathbb{N}$ and $r = 0, \ldots, n - 1$. Then by the elementary exponentiation laws (4.2):

$$e = a^m = a^{qn+r} = a^{qn} \cdot a^r = (a^n)^q \cdot a^r = e^q \cdot a^r = e \cdot a^r = a^r.$$

Since n is the *least* natural number satisfying this equation we must have $r = 0$, hence $m = qn$.

($\Leftarrow$) Conversely, if a has finite order n, and $m = qn$ for some $q \in \mathbb{N}$, then:

$$a^m = a^{qn} = (a^n)^q = e^q = e.$$

(ii) ($\Rightarrow$) If $a^m = a^p$ then, assuming for the sake of argument that $m > p$, "multiplying" both sides of the equation by a^{-p} and applying the extended exponentiation laws (4.3) gives:

$$e = a^m \cdot a^{-p} = a^{m-p}.$$

Since $m - p \in \mathbb{N}$ it follows from (i) that a has finite order and $m - p$ is a multiple of $o(a)$.

($\Leftarrow$) Conversely, if a has finite order n, and $m \equiv p \pmod{n}$, then assuming for the sake of argument that $m - p$ is a positive multiple of n, using (4.2) and (i) gives us:

$$a^m = a^{m-p} \cdot a^p = e \cdot a^p = a^p.$$

(iii) Suppose $o(a) = n$. We noted in Remark 4.8 (3) that $a^{-1} = a^{n-1}$, hence by the elementary laws of exponentiation (4.2):

$$(a^{-1})^n = (a^{n-1})^n = a^{n(n-1)} = (a^n)^{n-1} = e^n = e.$$

It follows from (i) that a^{-1} has finite order and $n = o(a)$ is a multiple of $o(a^{-1})$. Now, applying the same argument to a^{-1} and using the identity $(a^{-1})^{-1} = a$ (see Proposition 4.1 (iii)), we also have that $o(a^{-1})$ is a multiple of $o(a)$. Therefore $o(a) = o(a^{-1})$.

(iv) Let $f \colon \mathbb{N} \to G$ be the function with rule $f(r) = a^r$ for all $r \in \mathbb{N}$. Since G is finite, f cannot be one-to-one (otherwise the subset $f(\mathbb{N}) \subseteq G$ is infinite, contradicting Proposition 1.5). So there exist distinct $m, p \in \mathbb{N}$ such that $f(m) = f(p)$, meaning that $a^m = a^p$. Therefore a has finite order, by (ii). ■

In a finite group, the order of every element turns out to be no greater than the order of the group itself; we will prove this in Section 4.6.1 (Corollary 4.12). Meanwhile, the following example provides us with some evidence, and suggests that there is something even more intriguing going on.

Example 4.3 (Orders of the elements of $\mathbb{Z}_{11}^*$).
Let $(G, \cdot)$ be the group $\mathbb{Z}_p^*$ with the binary operation of multiplication modulo a prime number p (Section 4.3.3). This is a group of order $p - 1$, whose identity element is $\bar{1}$. To calculate the order of any element $\bar{a}$ we simply keep multiplying $\bar{a}$ by itself until we reach $\bar{1}$, which Proposition 4.3 (iv) tells us will happen eventually. The trick is to make sure that we take the least residue every time we perform a multiplication, which saves us having to work with very large numbers.

For example, suppose $p = 11$ and $a = 7$. Then we have:

$$\begin{aligned} \bar{7}^2 &= \bar{7}\,\bar{7} = \overline{49} = \bar{5}, & \bar{7}^3 &= \bar{7}^2\,\bar{7} = \bar{5}\,\bar{7} = \overline{35} = \bar{2}, \\ \bar{7}^4 &= \bar{7}^3\,\bar{7} = \bar{2}\,\bar{7} = \overline{14} = \bar{3}, & \bar{7}^5 &= \bar{7}^4\,\bar{7} = \bar{3}\,\bar{7} = \overline{21} = \overline{10}, \\ \bar{7}^6 &= \bar{7}^5\,\bar{7} = \overline{10}\,\bar{7} = \overline{70} = \bar{4}, & \bar{7}^7 &= \bar{7}^6\,\bar{7} = \bar{4}\,\bar{7} = \overline{28} = \bar{6}, \\ \bar{7}^8 &= \bar{7}^7\,\bar{7} = \bar{6}\,\bar{7} = \overline{42} = \bar{9}, & \bar{7}^9 &= \bar{7}^8\,\bar{7} = \bar{9}\,\bar{7} = \overline{63} = \bar{8}, \\ \bar{7}^{10} &= \bar{7}^9\,\bar{7} = \bar{8}\,\bar{7} = \overline{56} = \bar{1}. \end{aligned}$$

So $\bar{7}$ has order 10: $o(\bar{7}) = 10$.

Having determined the order of $\bar{7}$, we can apply Remark 4.8 (3) to easily find its inverse:

$$\bar{7}^{-1} = \bar{7}^9 = \bar{8}.$$

This agrees with Example 4.1, where we calculated $\bar{7}^{-1}$ using Bézout's identity. We then get "for free" that $\bar{8}^{-1} = \bar{7}$ by Proposition 4.1 (iii), just as in Example 4.1, and $o(\bar{8}) = 10$ from Proposition 4.3 (iv).

Let's now consider $\bar{5}$. Using the same procedure, and suppressing some of the steps to speed things up now that we know what we're doing:

$$\bar{5}^2 = \overline{25} = \bar{3}, \qquad \bar{5}^3 = \bar{3}\,\bar{5} = \overline{15} = \bar{4}, \qquad \bar{5}^4 = \bar{4}\,\bar{5} = \overline{20} = \bar{9}, \qquad \bar{5}^5 = \bar{9}\,\bar{5} = \overline{45} = \bar{1}.$$

So $o(\bar{5}) = 5$ and $\bar{5}^{-1} = \bar{5}^4 = \bar{9}$. Therefore $o(\bar{9}) = 5$.

We now consider $\bar{2}$:

$$\begin{aligned} \bar{2}^2 &= \bar{4}, & \bar{2}^3 &= \bar{4}\,\bar{2} = \bar{8}, & \bar{2}^4 &= \bar{8}\,\bar{2} = \overline{16} = \bar{5}, \\ \bar{2}^5 &= \bar{5}\,\bar{2} = \overline{10}, & \bar{2}^6 &= \overline{10}\,\bar{2} = \overline{20} = \bar{9}, & \bar{2}^7 &= \bar{9}\,\bar{2} = \overline{18} = \bar{7}, \\ \bar{2}^8 &= \bar{7}\,\bar{2} = \overline{14} = \bar{3}, & \bar{2}^9 &= \bar{3}\,\bar{2} = \bar{6}, & \bar{2}^{10} &= \bar{6}\,\bar{2} = \overline{12} = \bar{1}. \end{aligned}$$

Therefore $o(\bar{2}) = 10$ and $\bar{2}^{-1} = \bar{2}^9 = \bar{6}$. Hence $o(\bar{6}) = 10$ too.

The calculation for $\bar{3}$ goes as follows:

$$\bar{3}^2 = \bar{9}, \qquad \bar{3}^3 = \bar{9}\,\bar{3} = \overline{27} = \bar{5}, \qquad \bar{3}^4 = \bar{5}\,\bar{3} = \overline{15} = \bar{4}, \qquad \bar{3}^5 = \bar{4}\,\bar{3} = \overline{12} = \bar{1}.$$

Hence $o(\bar{3}) = 5$ and $\bar{3}^{-1} = \bar{4}$. Therefore $o(\bar{4}) = 5$.

Finally, here's $\overline{10}$:

$$\overline{10}^2 = \overline{100} = \bar{1}.$$

Hence $\overline{10}$ is an involution (ie. it has order 2), and is therefore self-inverse: $\overline{10}^{-1} = \overline{10}$.

To summarise:

$\bar{a}$	$\bar{1}$	$\bar{2}$	$\bar{3}$	$\bar{4}$	$\bar{5}$	$\bar{6}$	$\bar{7}$	$\bar{8}$	$\bar{9}$	$\overline{10}$
$o(\bar{a})$	1	10	5	5	5	10	10	10	5	2

Table 4.3.: Orders in $\mathbb{Z}_{11}^*$

There is something interesting emerging here: the orders of all elements happen to be factors of 10, which is the order of the group. We will take a closer look at what's going on in Set Piece 4 (Section 4.6.1). □

Remarks 4.9.

1) The methodology of Example 4.3 can be applied, in exactly the same way, to the groups $\mathbb{Z}_n^\times$.

2) It's interesting to take a look at the least residues of successive powers of an element $\bar{a}$ of $\mathbb{Z}_n^\times$. For example, in $\mathbb{Z}_{11}^*$ we have:

$$7, 5, 2, 3, 10, 4, 6, 9, 8, 1 \quad \cdots \text{ the least residues of successive powers of } \bar{7},$$
$$2, 4, 8, 5, 10, 9, 7, 3, 6, 1 \quad \cdots \text{ the least residues of successive powers of } \bar{2}.$$

We might expect to observe some sort of pattern in these sequences of integers; after all, pattern-spotting is something mathematicians are good at! However, in many cases, particularly for elements of high order, it is impossible to predict the progression from one integer to the next, without knowledge of n and a. To all intents and purposes, these sequences are random! In fact, this is the basis of the *Lehmer*[7] *random number generator,* one of the earliest algorithms for machine generation of sequences of random numbers [28]. ◊

4.4.1. Diffie-Hellman protocol

The Diffie-Hellman[8] protocol is a procedure that allows two correspondents, who are using a non-secure communication channel, to securely share a secret password. This can then be used as a "private key" to encrypt their correspondence, keeping it away from eavesdroppers. We will describe the basic idea, which is very simple.

Suppose two computer nerds (let's call them Ada and Bill) want to communicate with each other privately, but all their messages are intercepted and read by a third nerd (let's call this one Hal[9]). They can nevertheless exchange a piece of information, known only to themselves, by taking the following steps.

Step 1. Ada and Bill agree on a group G. By eavesdropping, this will also be known to Hal.

Step 2. Ada and Bill now agree on an element $g \in G$. This will also be known to Hal.

[7] Dick Lehmer (1905–1991): American mathematician, who pioneered the use of computers in number theory.

[8] Whitfield Diffie (b. 1944) and Martin Hellman (b. 1945): American cryptologists, credited with the invention of public key cryptography, through their influential paper [14].

[9] In "honour" of HAL 9000, the rogue computer in Stanley Kubrick's film: "2001: A Space Odyssey".

Step 3. Ada picks an exponent $a \in \mathbb{N}$ and works out the group element g^a. She then transmits this to Bill. Hal therefore knows g^a. However, Hal does not know a.

Step 4. Bill also picks an exponent $b \in \mathbb{N}$, works out g^b and transmits it to Ada. Hal is therefore aware of g^b, but not b.

Step 5. On receiving g^a from Ada, Bill raises it to his chosen power, calculating $(g^a)^b$. He keeps this to himself.

Step 6. On receiving g^b from Bill, Ada raises it to her chosen power, calculating $(g^b)^a$. She also keeps this to herself.

Step 7. By the elementary laws of exponentiation in groups (4.2):

$$(g^a)^b = g^{ab} = (g^b)^a.$$

So Ada and Bill are in possession of the same group element.

Step 8. Hal knows g^a and g^b, and can therefore calculate:

$$g^a\, g^b = g^{a+b}.$$

But in general:

$$g^{a+b} \neq g^{ab},$$

so without knowing a and b Hal can't construct Ada and Bill's shared group element from the eavesdropped information. Ada and Bill have therefore established their "private key"!

The success of this protocol depends on the practical impossibility of Hal being able to "reverse engineer" Ada's and Bill's chosen exponents a and b from his knowledge of the group elements g^a and g^b. This depends entirely on the chosen group G and element g. As observed in Remark 4.9 (2), good candidates are $G = \mathbb{Z}_p^*$ for some large prime p, and g an element of high order.

Example 4.4 (Diffie-Hellman in $\mathbb{Z}_{11}^*$**).**
For simplicity, we illustrate the Diffie-Hellman protocol in the group $G = \mathbb{Z}_p^*$ with $p = 11$ (see Example 4.3).

Suppose Ada and Bill have chosen the group element $g = \bar{7}$. Let's say Ada now chooses the exponent $a = 2$ and Bill chooses $b = 4$. Ada transmits $g^a = (\bar{7})^2 = \bar{5}$ to Bill, who then calculates:

$$(g^a)^b = (\bar{5})^4 = \bar{3}\,(\bar{5})^2 = \bar{4}\,\bar{5} = \bar{9}.$$

Bill now transmits $g^b = (\bar{7})^4 = \bar{3}$ to Ada, who then calculates:

$$(g^b)^a = (\bar{3})^2 = \bar{9}.$$

So Ada and Bill are now in possession of the same key. However, Hal knows only $g = \bar{7}$, $g^a = \bar{5}$ and $g^b = \bar{3}$. Because of the randomness of the sequence of powers of $\bar{7}$, the only way Hal can obtain the exponents a and b is by systematically calculating all powers of $\bar{7}$ until he reaches $\bar{5}$ and $\bar{3}$. In this case that's an easy task, meaning that Hal will very quickly be able to discover Ada and Bill's secret. □

Practical applications of the Diffie-Hellman protocol use groups $G = \mathbb{Z}_p^*$ with primes p that are much larger than 11, typically around 2^{11} binary digits ("binary digits"—or "bits"—because of course all calculations are computerised). To get a feeling for the size of such numbers, if we approximate $2^{10} = 1024 \approx 10^3$ then using the familiar numerical laws of exponentiation (of which (4.2) are an algebraic generalisation) gives us the following approximation:

$$2^{2048} = 2^{(10\times 204)+8} = (2^{10})^{204}\, 2^8 \approx 256 \times 10^{612} = 2.56 \times 10^{614}.$$

So we are talking about primes p with around 614 decimal digits. For comparison, it is currently estimated that there are up to 10^{82} atoms in the universe! Although these primes p are large, they pale in comparison to the largest known prime number, which has 24,862,048 decimal digits; see Remark 2.5. By choosing elements $g \in G$ of high order it is possible to generate sequences that appear to be random and are effectively infinite in length, making it virtually impossible for Hal to either guess or calculate Ada and Bill's private key.

4.5. Symmetric groups and permutations

Symmetric groups play a prominent rôle in group theory; for example, they were used by Galois[10] in his revolutionary theory about the roots of polynomials. (More about this in Year 3.) Having been around long before the abstract concept of a group emerged (in fact, they are the prototype example that motivated the axiomatic definition), they have acquired a notation and theory all of their own, which we explore from a modern perspective in this Section, before resuming our investigation of general group theory in Section 4.6.

There are some striking similarities between the basic ideas and terminology of group theory introduced in Section 4.2 and aspects of our discussion of bijections in Section 1.2.4. This is not a coincidence: the bijections from a given set to itself form a group! Here it is.

[10] Evariste Galois (1811–1832): French mathematician, who revolutionised the theory of polynomial equations, but got caught up in revolutionary politics and died in a duel!

Definition 4.6 (Symmetric group)**.**
Let X be a set. A bijection $\sigma\colon X \to X$ is (sometimes) called a *symmetry* of X. The *symmetric group on X* is then the set of all symmetries:

$$\mathrm{Sym}(X) = S_X = \{\sigma : \sigma \text{ a symmetry of } X\},$$

whose "multiplication" is composition of functions (see Section 1.2.5):

$$\mu(\sigma, \tau) = \sigma \circ \tau, \quad \forall \sigma, \tau \in S_X.$$

Recall that the composition of two bijections is a bijection (Proposition 1.3), so μ is a well-defined binary operation on S_X. ♦

Note. The elements of symmetric groups are functions; be careful! Note also the specific notation σ, τ etc. (rather than the usual f, g etc.), which is part of the heritage of these groups. ♢

Let's check that $(S_X, \circ)$ is indeed a group, by verifying the axioms.

Associative. Composition of functions is always associative (see Remarks 1.17 (2)).

Identity. The identity function $1_X\colon X \to X$ defined $1_X(x) = x$ for all $x \in X$ (see Remarks 1.17 (5)) is a bijection, hence an element of S_X, and satisfies:

$$(\sigma \circ 1_X)(x) = \sigma(1_X(x)) = \sigma(x) = 1_X(\sigma(x)) = (1_X \circ \sigma)(x).$$

So by the Principle for Equality of Functions we can write:

$$\sigma \circ 1_X = \sigma = 1_X \circ \sigma, \quad \forall \sigma \in S_X,$$

which is precisely axiom (G2) in this context.

Inverses. Since σ is a bijection it has an inverse function $\sigma^{-1}\colon X \to X$ (Definition 1.4), which is also a bijection (Proposition 1.1), and by Remark 1.17 (5) satisfies:

$$\sigma^{-1} \circ \sigma = 1_X = \sigma \circ \sigma^{-1}.$$

This is precisely axiom (G3) in this context.

Having checked the group axioms it follows from Proposition 4.1 that the identity element of S_X *is* the identity function of X, and the inverse element of $\sigma \in S_X$ *is* the inverse function σ^{-1}. So the terminology of groups matches up perfectly with that of functions!

Of particular importance are the symmetric groups on *finite* sets. These are the "permutation groups".

Definition 4.7 (Permutation group)**.**
If $X = \{1, \ldots, n\} = [n]$, then we write S_n for the symmetric group S_X. A bijection $[n] \to [n]$ is nothing other than a *permutation* of the integers $1, \ldots, n$, and S_n is therefore often referred to as a *permutation group.* Since there are precisely $n!$ permutations of n objects, S_n is a group of order $n!$ (ie. $|S_n| = n!$). ♦

Remark 4.10. The order of S_n gets very large, very quickly; for example, the number of shuffles of a standard pack of cards is:

$$|S_{52}| = 52! > 8 \times 10^{67},$$

whereas throwing in the jokers, and those aces up our sleeve, increases this to:

$$|S_{58}| > 2 \times 10^{78},$$

which is of the same order of magnitude as current lower estimates of the number of atoms in the universe! It is clear that there is absolutely no way in which group multiplication tables (cf. Table 4.2) are a sensible way to study groups of this size; we need to be much cleverer. ◊

Remarks 4.11 (Permutations; elementary concepts).

1) Elements of the permutation group S_n can be written in *2-line notation:*

$$\sigma = \begin{pmatrix} 1 & 2 & \cdots & n \\ a_1 & a_2 & \cdots & a_n \end{pmatrix},$$

where $\{a_1, \ldots, a_n\} = \{1, \ldots, n\}$. This is the permutation:

$$\sigma(1) = a_1, \quad \sigma(2) = a_2, \ \ldots, \ \sigma(n) = a_n.$$

2) There is also *cycle notation:*

$$\sigma = \begin{pmatrix} c_1 & c_2 & \cdots & c_r \end{pmatrix},$$

where $\{c_1, \ldots, c_r\} \subseteq \{1, \ldots, n\}$. This indicates that σ permutes the integers $c_1, \ldots, c_r$ in the following way:

$$c_1 \mapsto c_2, \quad c_2 \mapsto c_3, \ \ldots, \ c_{r-1} \mapsto c_r, \quad c_r \mapsto c_1.$$

Thus σ simply "cycles" through $c_1, \ldots, c_r$ in the specified order, and leaves all other integers unchanged. A permutation of this type is called an *r-cycle.*

The cycle notation isn't unique, because:

$$\begin{pmatrix} c_1 & c_2 & \cdots & c_r \end{pmatrix} = \begin{pmatrix} c_2 & \cdots & c_r & c_1 \end{pmatrix} = \cdots$$

3) If $\sigma, \tau \in S_n$ then the composition $\sigma \circ \tau$ is often abbreviated to $\sigma\tau$ and referred to as the *product* of σ and τ. This is consistent with our general inclination to drop the group multiplication symbol whenever possible (see Remark 4.3 (3)). However, we need to remember when dealing with "products" of permutations that the order is important, and that $\sigma\tau$ indicates that τ is applied *before* σ. ◊

4.5.1. Cycle decomposition

Not every permutation is a cycle; however every permutation can be expressed as a composition/product of "disjoint" cycles. In this sense, permutations may be "factorised" by cycles. Rather than give a formal definition and description at the outset, we illustrate this with the following sequence of examples.

Example 4.5 (Multiplication)**.**
First, to show how group multiplication works using 2-line notation, suppose $n = 5$ and σ, τ are the following elements of S_5:

$$\sigma = \begin{pmatrix} 1 & 2 & 3 & 4 & 5 \\ 5 & 4 & 2 & 1 & 3 \end{pmatrix}, \qquad \tau = \begin{pmatrix} 1 & 2 & 3 & 4 & 5 \\ 3 & 1 & 5 & 4 & 2 \end{pmatrix}.$$

Then bearing in mind that "multiplication" is composition, which operates from right to left:

$$\sigma\tau = \begin{pmatrix} 1 & 2 & 3 & 4 & 5 \\ 2 & 5 & 3 & 1 & 4 \end{pmatrix}, \qquad \tau\sigma = \begin{pmatrix} 1 & 2 & 3 & 4 & 5 \\ 2 & 4 & 1 & 3 & 5 \end{pmatrix}.$$

Notice that $\sigma\tau \neq \tau\sigma$; so S_5 is a non-abelian (ie. non-commutative) group. It is easy to construct similar examples which show that S_n is non-abelian for all $n \geqslant 3$. □

Example 4.6 (Inverses)**.**
To find the inverse of a permutation expressed in two-line notation we simply swap the two lines; then re-arrange the columns so the numbers in the top row are correctly ordered. For example, inverting σ from Example 4.5:

$$\sigma^{-1} = \begin{pmatrix} 1 & 2 & 3 & 4 & 5 \\ 5 & 4 & 2 & 1 & 3 \end{pmatrix}^{-1} = \begin{pmatrix} 5 & 4 & 2 & 1 & 3 \\ 1 & 2 & 3 & 4 & 5 \end{pmatrix} = \begin{pmatrix} 1 & 2 & 3 & 4 & 5 \\ 4 & 3 & 5 & 2 & 1 \end{pmatrix}.$$

(You could/should check that $\sigma\sigma^{-1} = e$.) □

Example 4.7 (Cycle decomposition)**.**
Now let $n = 10$ and let $\sigma \in S_{10}$ be the permutation:

$$\sigma = \begin{pmatrix} 1 & 2 & 3 & 4 & 5 & 6 & 7 & 8 & 9 & 10 \\ 7 & 5 & 8 & 10 & 4 & 2 & 9 & 3 & 1 & 6 \end{pmatrix}.$$

We will write σ as a product of disjoint cycles. For this, we pick a starting point, say 1, and track its journey under σ:

$$1 \mapsto 7 \mapsto 9 \mapsto 1.$$

This is the 3-cycle $(1\ 7\ 9)$. Now start at a number not in this cycle, say 2, and track its journey:

$$2 \mapsto 5 \mapsto 4 \mapsto 10 \mapsto 6 \mapsto 2.$$

This is the 5-cycle $(2\ 5\ 4\ 10\ 6)$. Finally, we note that $3 \mapsto 8 \mapsto 3$, which is the 2-cycle $(3\ 8)$.

We have now recorded what happens to all ten integers, and can therefore express σ as the product of these cycles:

$$\sigma = \big(3\ \ 8\big)\big(2\ \ 5\ \ 4\ \ 10\ \ 6\big)\big(1\ \ 7\ \ 9\big),$$

which means:

> *First do* $(1\ \ 7\ \ 9)$*, then do* $(2\ \ 5\ \ 4\ \ 10\ \ 6)$ *and finally do* $(3\ \ 8)$.

We could equally well perform the cycles in reverse (or any other) order:

$$\sigma = (1\ \ 7\ \ 9)(2\ \ 5\ \ 4\ \ 10\ \ 6)(3\ \ 8).$$

This is because the numbers being permuted by each cycle live in disjoint subsets of $\{1, 2, \ldots, 10\}$, which is what is meant by the phrase *disjoint cycles.* In other words: "disjoint cycles commute". Notice also that in writing these products of cycles we have (again) omitted the composition symbol "$\circ$", which makes the notation much tidier. □

Example 4.8 (Inverse of a product of disjoint cycles)**.**
It's very easy to find the inverse of a cycle: just write it down backwards! For example:

$$(1\ \ 7\ \ 9)^{-1} = (9\ \ 7\ \ 1) = (1\ \ 9\ \ 7).$$

Then, if a permutation σ has been factorised into disjoint cycles, it follows from Proposition 4.1 (iv) that we can write down σ^{-1} immediately (ie. without reverting to 2-line notation) by simply inverting each cycle and taking their product. Furthermore, since the inverted cycles remain disjoint, we can arrange this product however we like.

For example, if σ is the permutation of Example 4.7 then:

$$\sigma^{-1} = (3\ \ 8)^{-1}(2\ \ 5\ \ 4\ \ 10\ \ 6)^{-1}(1\ \ 7\ \ 9)^{-1} = (3\ \ 8)(2\ \ 6\ \ 10\ \ 4\ \ 5)(1\ \ 9\ \ 7).$$

Not only have we found the inverse of σ, we have also factorised it into disjoint cycles! □

Example 4.9 (Non-disjoint cycles)**.**
When multiplying cycles that are *not* disjoint we have to be a little more careful. For example, suppose $\sigma, \tau \in S_5$ are the 5-cycles:

$$\sigma = (5\ \ 4\ \ 2\ \ 1\ \ 3), \qquad \tau = (3\ \ 1\ \ 5\ \ 4\ \ 2).$$

(These look similar to the permutations in Example 4.5, but we're using different notation, so they're different!) Then:

$$\sigma\tau = \big(2\ \ 5\big)\big(1\ \ 4\big), \qquad \tau\sigma = \big(3\ \ 4\big)\big(2\ \ 5\big).$$

Notice that these cycles no longer commute. □

Note. We emphasise again, as seen in all the above examples, that composition of functions operates right-to-left (ie. backwards!); we always need to remember this when "multiplying" in the symmetric group. ◊

Having seen cycles at work, we now formalise the definitions made in Remark 4.11 (2) and Example 4.7, and prove that the techniques work in general. We also show how to find the order of any permutation.

Definition 4.8 (Cycle; formal definition)**.**
A permutation $\sigma \in S_n$ is an *r-cycle* if there exists a subset of integers:

$$C = \{c_1, \dots, c_r\} \subseteq \{1, \dots, n\}$$

such that:

$$\sigma(c_1) = c_2,\ \sigma(c_2) = c_3,\ \dots,\ \sigma(c_{r-1}) = c_r,\ \sigma(c_r) = c_1,$$
$$\sigma(m) = m, \quad \text{for all } m \notin C.$$

Two cycles σ, σ' are *disjoint* if $C \cap C' = \emptyset$. ♦

More generally, we define the notion of "disjoint permutations". For this, we recall the notion of a partition of a set (Definition 3.8).

Definition 4.9 (Disjoint permutations)**.**
Permutations $\sigma, \tau \in S_n$ are *disjoint* if there exists a partition $[n] = S \cup T$ (which in this case simply means $S \cap T = \emptyset$) such that:

$$\sigma(t) = t,\ \text{for all } t \in T; \quad \tau(s) = s,\ \text{for all } s \in S.$$

Since σ and τ are bijections, this implies that $\sigma(s) \in S$ and $\tau(t) \in T$ for all $s \in S$ and $t \in T$. In other words, σ permutes the elements of S and leaves the elements of T untouched, whereas τ does exactly the opposite. ♦

Remark 4.12 (Disjoint permutations commute).
Like disjoint cycles, disjoint permutations commute. Because, bearing in mind that "multiplication" in the symmetric group is composition of functions, for all $s \in S$ and $t \in T$ we have:

$$\sigma\tau(s) = \sigma(\tau(s)) = \sigma(s) = \tau(\sigma(s)) = \tau\sigma(s),$$
$$\sigma\tau(t) = \sigma(\tau(t)) = \tau(t) = \tau(\sigma(t)) = \tau\sigma(t).$$

Therefore $\sigma\tau = \tau\sigma$, by the Principle for Equality of Functions. ◊

Proposition 4.4 (Order of a product of disjoint permutations)**.**
Suppose $\sigma, \tau \in S_n$ are disjoint permutations, with $o(\sigma) = p$ and $o(\tau) = q$. Then:

$$o(\sigma\tau) = \mathrm{lcm}(p, q),$$

where lcm stands for "least common multiple".

Proof. This is a direct proof.

Since σ and τ commute, for all $m \in \mathbb{N}$ we have:

$$(\sigma\tau)^m = (\sigma\tau)\cdots(\sigma\tau) = \sigma^m\tau^m = \tau^m\sigma^m.$$

(Note the application of associativity, many times, in shifting the brackets around. We usually do this without comment!) Therefore, if $(\sigma\tau)^m = e$ then using the disjointness of σ and τ in full it follows that for all $s \in S$ and $t \in T$ we have:

$$s = (\sigma\tau)^m(s) = \sigma^m(\tau^m(s)) = \sigma^m(s), \qquad \sigma^m(t) = t.$$

Hence $\sigma^m = e$. A similar argument shows that $\tau^m = e$. It follows from Proposition 4.3 (i) that m is a multiple of p *and* a multiple of q; ie. a common multiple of p and q. Since (by definition) the order of $\sigma\tau$ is the *least* such m, we have $o(\sigma\tau) = \mathrm{lcm}(p, q)$. ■

Remark 4.13 (Orders of products of group elements).
Since $\mathrm{lcm}(p, q) = \mathrm{lcm}(q, p)$ the equation for $o(\sigma\tau)$ is symmetric in p and q, which is consistent with the fact that σ and τ commute. In fact, it turns out that:

$$o(\sigma\tau) = o(\tau\sigma),$$

for all $\sigma, \tau \in S_n$, even if σ and τ don't commute! Furthermore, this is a property that generalises to *all* groups $(G, \cdot)$:

$$o(xy) = o(yx),$$

for all $x, y \in G$. To see this, simply note that for any integer $m \geqslant 2$ we have:

$$(xy)^m = (xy)\cdots(xy) = x(yx)\cdots(yx)y = x(yx)^{m-1}y,$$

so multiplying on the right by x:

$$(xy)^m x = x(yx)^m$$

and then by x^{-1}, gives us the equation:

$$(xy)^m = x(yx)^m x^{-1}.$$

It follows from this that $(xy)^m = e$ if and only if $(yx)^m = e$, which by Definition 4.5 implies that xy and yx have the same order (whether finite or infinite).

The general proof is arguably clearer than one specifically tailored to permutations, since it pinpoints the precise abstract algebraic properties that are being utilised. This is a good illustration of the philosophy behind the abstract formulation of groups, mentioned at the beginning of Section 4.2. ◊

The group S_n, although very large is nevertheless finite; so every element has finite order (Proposition 4.3 (iv)). The following result tells us how to compute the orders of disjoint products of cycles. Since we have already hinted that every permutation has such a decomposition (we'll give a formal proof in Proposition 4.6 below) this in effect tells us how to find the order of *every* element of S_n.

Proposition 4.5 (Order of a product of disjoint cycles)**.**

i) *The order of an r-cycle is r.*

ii) *The order of a product of disjoint cycles is the least common multiple (lcm) of the orders of the cycles.*

Proof. The proof of (i) is direct, whereas we prove (ii) by induction on the number k of cycles.

i) This seems more or less obvious. Nevertheless, suppose $\sigma = (c_1 \ \cdots \ c_r)$ for distinct $c_1, \ldots, c_r \in [n]$. First, for all $i = 1, \ldots, r$ we have:

$$\begin{aligned} \sigma^r(c_i) &= \sigma \circ \cdots \circ \sigma(c_i) \quad (r \text{ compositions}) \\ &= c_i. \end{aligned}$$

Since σ fixes the remaining integers it follows that $\sigma^r = e$ (recalling that e is the identity function). Now if $k < r$ then:

$$\sigma^k(c_1) = c_{k+1} \neq c_1,$$

which is all we need to show that $\sigma^k \neq e$. Therefore $o(\sigma) = r$.

ii) This is true if $k = 1$. Suppose $\sigma = \rho_1 \cdots \rho_{k+1}$ where $\rho_1, \ldots, \rho_{k+1}$ are disjoint cycles of orders $r_1, \ldots, r_{k+1}$. Then $\sigma = \rho\rho_{k+1}$ where $\rho = \rho_1 \cdots \rho_k$. Although ρ is not in general a cycle, ρ and ρ_{k+1} are disjoint permutations, so it follows from Proposition 4.4 that:

$$\begin{aligned} o(\sigma) &= \operatorname{lcm}(o(\rho), o(\rho_{k+1})) \\ &= \operatorname{lcm}(\operatorname{lcm}(r_1, \ldots, r_k), r_{k+1}), \quad \text{by the induction hypothesis} \\ &= \operatorname{lcm}(r_1, \ldots, r_{k+1}). \end{aligned}$$

■

In broad terms, there are two things that can happen to an integer when acted on by a permutation: it moves, or it stays put. Here's the formal definition.

Definition 4.10 (Fixed points of permutations)**.**
Suppose $\sigma \in S_n$ and $c \in [n]$. If $\sigma(c) = c$ then we say that σ *fixes* c, or c is a *fixed point* of σ. If $\sigma(c) \neq c$ we say that σ *moves* c. ♦

Proposition 4.6 (Cycle decomposition)**.**
Every permutation is a product of uniquely determined disjoint cycles.

Proof. This is a proof by induction, on the number m of integers moved by the permutation, and uses the strong form of the induction hypothesis.

For all $m = 0, 1, \ldots, n$ let $P(m)$ be the statement:

> *Every permutation $\sigma \in S_n$ that moves m integers is a product of uniquely determined disjoint cycles.*

Then $P(0)$ is true, since in this case $\sigma = e$ which can be written as a product of 1-cycles $(1)(2)\cdots(n)$, and these are of course disjoint. Furthermore, any other product of disjoint cycles moves at least one integer, and therefore cannot be a factorisation of σ. This is the base case.

Now suppose $0 \leqslant m \leqslant n-1$ and $P(0), \ldots, P(m)$ are all true. Suppose σ moves $m+1$ integers. Let c be the *least* integer moved by σ, and consider the subset:

$$C = \{c, \sigma(c), \sigma^2(c), \ldots\} \subseteq [n].$$

Since C is finite there exists a *least* integer $k \geqslant 2$ such that:

$$\sigma^k(c) \in \{c, \sigma(c), \ldots, \sigma^{k-1}(c)\}.$$

Thus $\sigma^k(c) = \sigma^j(c)$ for some $j = 0, \ldots, k-1$. Multiplying both sides by σ^{-j} yields $\sigma^{k-j}(c) = c$, which is only possible if $j = 0$. Hence $\sigma^k(c) = c$. So if we define:

$$c_1 = c,\ c_2 = \sigma(c),\ \ldots,\ c_k = \sigma^{k-1}(c),$$

then we obtain a k-cycle $\rho = (c_1\ \cdots\ c_k)$ with the property that $\sigma = \sigma'\rho$ where $\sigma' \in S_n$ fixes $c_1, \ldots, c_k$. Therefore σ' and ρ are disjoint permutations, and σ' moves at most m integers since ρ moves at least 1. By the induction hypothesis (existence) σ' is a product of disjoint cycles, and therefore so is σ. Furthermore, by its construction ρ must appear in *any* decomposition of σ into disjoint cycles, so by the induction hypothesis (uniqueness) the product is unique. Thus $P(m+1)$ is true.

It remains to note that $P(m)$ is true for all $m > n$. This is a consequence of the following logical loophole, which is in fact a mini proof by contradiction: if $P(m_0)$ were false for some $m_0 > n$ then there would exist a permutation $\sigma \in S_n$ that moves m_0 integers, which is impossible.

We have shown for all $m \geqslant 0$ that if $P(m)$ is true then $P(m+1)$ is true, which completes the induction step. So by the Principle of Induction $P(m)$ is true for all m. ■

Remarks 4.14.

1) The proof of Proposition 4.6 shows how the Principle of Induction may be applied to situations where only a *finite* number of statements $P(n)$ need to be proved.

2) The proof used (twice) the well-ordering principle for $\mathbb{N}$ (cf. Remark 1.24).

3) Combining Propositions 4.5 and 4.6 with Example 4.8 shows that for any permutation σ we have:
$$o(\sigma) = o(\sigma^{-1}).$$
In fact, this is true more generally:
$$o(a) = o(a^{-1}),$$
for *any* element a in *any* group G (cf. Remark 4.13). *(Can you prove it?)* ♢

Example 4.10 (Order of a permutation)**.**
Let's consider the permutation $\sigma \in S_{10}$ from Example 4.7, which we decomposed into the following product of disjoint cycles:
$$\sigma = \begin{pmatrix}3 & 8\end{pmatrix}\begin{pmatrix}1 & 7 & 9\end{pmatrix}\begin{pmatrix}2 & 5 & 4 & 10 & 6\end{pmatrix}.$$
Since $\mathrm{lcm}(2, 3, 5) = 30$ we conclude that $o(\sigma) = 30$. Notice that this is *much less* than the order of the group (which is $10! = 3628800$), and also happens to be a factor of the group order (cf. Example 4.3). □

4.5.2. Transpositions

The simplest kind of permutation (other than the identity) is one that swaps two numbers and leaves everything else fixed. This is in fact a cycle, with just two entries.

Definition 4.11 (Transposition)**.**
A 2-cycle τ is usually called a *transposition.* ♦

Note. Transpositions have the property that $\tau^2 = e$, or equivalently $\tau^{-1} = \tau$; ie. they are involutions. Although we've seen this sort of behaviour before, it's still unusual enough to be noteworthy. In fact there are other permutations of order 2: by Proposition 4.5, any product of disjoint transpositions has this property. ♢

Transpositions are the "building blocks" of all permutations, because of the following result.

Proposition 4.7 (Factorisation into transpositions)**.**
Every permutation σ can be "factorised":

$$\sigma = \tau_k \cdots \tau_1,$$

where $\tau_1, \ldots, \tau_k$ are transpositions.

Note. The transpositions $\tau_1, \ldots, \tau_k$ are not usually disjoint; this will become apparent in the proof. ◊

Proof. This is a direct proof, which begins with a special case.

First, suppose σ is an r-cycle:

$$\sigma = (c_1 \ \cdots \ c_r).$$

Notice that:

$$(c_1 \ c_2)(c_1 \ \cdots \ c_r) = (c_1)(c_2 \ \cdots \ c_r) = (c_2 \ \cdots \ c_r).$$

(Remember to read these products right-to-left.) Applying this a further $r-2$ times gives us:

$$(c_{r-1} \ c_r) \cdots (c_1 \ c_2)\sigma = e.$$

Then multiplying this equation successively on the left by $(c_{r-1} \ c_r), \ldots, (c_1 \ c_2)$ gives us:

$$\sigma = (c_1 \ c_2)(c_2 \ c_3) \cdots (c_{r-1} \ c_r). \tag{4.4}$$

Now, for a general permutation σ, we simply decompose σ into a product of disjoint cycles and apply (4.4) to each cycle. ■

Remark 4.15. Another way of stating Proposition 4.7 is that the symmetric group $G = S_n$ is generated by the subset T of all transpositions; ie. every element of G can be expressed as a product of elements of T. Since $|T| = \binom{n}{2}$ this results in a considerable reduction in the amount of information required to construct G. We can be even more economical by noting that any transposition $(i \ j)$ can be written:

$$(i \ j) = (1 \ i)(1 \ j)(1 \ i).$$

So S_n is generated by the subset:

$$T_1 = \{(1 \ 2), \ldots, (1 \ n)\}$$

containing only $n-1$ transpositions. In fact it is possible to find generating sets for S_n with only 2 elements [24], no matter how large n may be!

The idea of finding small generating subsets is a useful one that can be formulated for groups in general; see for example Definition 4.17. ◊

We refer to a decomposition of σ into transpositions as a *complete factorisation* of σ. Proposition 4.7 guarantees that complete factorisations always exist; however they are far from unique. For example, in S_5 we have:

$$\sigma = (1\ 2\ 3)(4\ 5) = (1\ 2)(2\ 3)(4\ 5),$$

as in the proof of Proposition 4.7, but we can also factorise σ more exotically as:

$$\sigma = (2\ 3)(3\ 5)(1\ 5)(3\ 4)(3\ 5).$$

Not only are the two sets of transpositions completely different, they also differ in number! And here is a complete factorisation of an r-cycle which is fundamentally different to that of (4.4):

$$(c_1\ \cdots\ c_r) = (c_1\ c_r)\cdots(c_1\ c_3)(c_1\ c_2). \tag{4.5}$$

This means that when factorising permutations, transpositions don't behave in quite the same way as, for example, prime numbers in the factorisation of integers (which are unique). Nevertheless, from factorisations such as (4.4) and (4.5) it looks like the minimum number of transpositions required to completely factorise an r-cycle is $r-1$. We won't prove this, or even have need of it, but regard it as motivation for the following definition.

Definition 4.12 (Transposition number)**.**
The *transposition number* $T(\sigma)$ of an arbitrary permutation $\sigma \in S_n$ is defined to be the non-negative integer computed by first decomposing σ into disjoint cycles and then taking the following sum:

$$T(\sigma) = \sum_{r=1}^{n} (r-1)(\#\ r\text{-cycles}).$$

In other words, we take a weighted sum of the number of disjoint cycles, where the weights are what we believe to be the minimum number of transpositions required to completely factorise each cycle.

Since the decomposition into disjoint cycles is unique, $T(\sigma)$ is well-defined. In particular, if σ is an r-cycle then $T(\sigma) = r-1$. Furthermore $T(e) = 0$, since the only way of decomposing e into disjoint cycles is as a string of 1-cycles, each of which has transposition number 0. ♦

Example 4.11 (Transposition number)**.**
The permutation $\sigma \in S_{10}$ in Example 4.10 has transposition number:

$$T(\sigma) = 1.1 + 2.1 + 4.1 = 7. \qquad \square$$

It looks as if $T(\sigma)$ should be the smallest number of transpositions required to completely factorise σ. Again, we won't prove this, or even assume that it's true. However,

$T(\sigma)$ gives us an important piece of information about the possible ways to completely factorise σ. To state this clearly, we say that two integers $k, \ell \in \mathbb{Z}$ have the *same parity* if they're both even, or both odd; in other words, if:

$$(-1)^k = (-1)^\ell.$$

In particular, this classifies 0 as an even integer (as we would expect). Now here's our result. Since it carries a fair bit of clout, we're going to bestow upon it the title "theorem".

Theorem 4.8 (Parity Theorem)**.**
For all $\sigma \in S_n$ the number of transpositions in any complete factorisation of σ has the same parity as $T(\sigma)$.

Proof. We'll subdivide the proof into two parts. The main argument (Part 2) is a proof by induction (on the number of transpositions), preceded by a proof by exhaustion (Part 1), with only two cases (which could conceivably be stated separately as a lemma), that will help us make the induction step.

Part 1. We first observe what happens when σ is multiplied by a transposition $\tau = (a\ \ b)$ to form the new permutation $\sigma' = \tau\sigma$. When σ is decomposed into its disjoint cycles there are two possibilities: either a, b appear in different cycles, or in the same cycle. In the first case σ' contains a product of the form:

$$(a\ \ b)\big(a\ \ c_1\ \cdots\ c_r\big)\big(b\ \ d_1\ \cdots\ d_s\big) = \big(b\ \ d_1\ \cdots\ d_s\ \ a\ \ c_1\ \cdots\ c_r\big),$$

with all other cycles of σ unaffected by τ. Computing transposition numbers, we see that:

$$T(\sigma') = T(\sigma) + 1.$$

In the second case σ' contains a product of the form:

$$(a\ \ b)\big(a\ \ c_1\ \cdots\ c_r\ \ b\ \ d_1\ \cdots\ d_s\big) = \big(b\ \ d_1\ \cdots\ d_s\big)\big(a\ \ c_1\ \cdots\ c_r\big),$$

with all other cycles of σ undisturbed. Hence:

$$T(\sigma') = T(\sigma) - 1.$$

Thus, multiplying any permutation by a transposition changes the parity of the transposition number; ie. switches it from even to odd, and vice versa.

Part 2. We now proceed to the induction. Let $P(k)$ be the statement:

> *If σ is a product of k transpositions then k has the same parity as $T(\sigma)$.*

The base case $P(1)$ is true, because a transposition, being a 2-cycle, has transposition number 1. For the induction step, suppose $P(k)$ is true and σ is a product of $k+1$ transpositions:

$$\sigma = \tau_{k+1}\tau_k \cdots \tau_1.$$

Since transpositions are involutions we can rearrange this equation to:

$$\tau_{k+1}\sigma = \tau_k \cdots \tau_1.$$

Hence by the induction hypothesis $T(\tau_{k+1}\sigma)$ has the same parity as k. Therefore, by the first part of the proof, $T(\sigma)$ has opposite parity to k, hence the same parity as $k+1$. This shows that $P(k+1)$ is true. We therefore conclude that $P(k)$ is true for all k, by the Principle of Induction. ■

4.5.3. Even and odd permutations

In the light of Theorem 4.8 it's natural to introduce the following concept.

Definition 4.13 (Parity of a permutation)**.**
A permutation $\sigma \in S_n$ is said to be *even* or *odd* according as $T(\sigma)$ is even or odd; this property is called the *parity* of σ. ⧫

The key message of Theorem 4.8 is that if σ is an even (resp. odd) permutation then any complete factorisation of σ will always have an even (resp. odd) number of transpositions. So, to determine the parity of σ we can either calculate $T(\sigma)$ by decomposing σ into disjoint cycles and applying Definition 4.12, or completely factorise σ (however we like) and count the number of transpositions. The first route is often more useful in practice, whereas the second route can be very useful for theoretical purposes (see for example Proposition 4.9 below).

Example 4.12. The permutation σ in Example 4.10 has transposition number $T(\sigma) = 7$ and is therefore odd; so any complete factorisation of σ will always contain an odd number of transpositions. □

A simple and useful numerical way to record the parity of a permutation is by its "signature", which is defined as follows.

Definition 4.14 (Signature of a permutation)**.**
The *signature* of a permutation $\sigma \in S_n$ is defined:

$$\operatorname{sgn}(\sigma) = \begin{cases} 1, & \text{if } \sigma \text{ is even,} \\ -1, & \text{if } \sigma \text{ is odd.} \end{cases}$$

Following Theorem 4.8, another way of writing this is:

$$\operatorname{sgn}(\sigma) = (-1)^{T(\sigma)} = (-1)^N, \tag{4.6}$$

where N is the number of transpositions in *any* complete factorisation of σ. ⧫

Signatures have the following rather nice property.

Proposition 4.9 (Multiplicative property of signature)**.**
For all $\sigma, \tau \in S_n$ we have:

$$\operatorname{sgn}(\tau\sigma) = \operatorname{sgn}(\tau)\operatorname{sgn}(\sigma).$$

Proof. This is a comparitively simple direct proof; we've already done all the hard work!

Suppose $\sigma = \sigma_r \cdots \sigma_1$ and $\tau = \tau_s \cdots \tau_1$ are complete factorisations. Then by (4.6):

$$\operatorname{sgn}(\sigma) = (-1)^r, \qquad \operatorname{sgn}(\tau) = (-1)^s,$$

and since

$$\tau\sigma = \tau_s \cdots \tau_1 \, \sigma_r \cdots \sigma_1$$

is also a complete factorisation we have:

$$\operatorname{sgn}(\tau\sigma) = (-1)^{s+r} = (-1)^s(-1)^r = \operatorname{sgn}(\tau)\operatorname{sgn}(\sigma). \qquad \blacksquare$$

Remark 4.16. We may regard sgn as a function $\operatorname{sgn}\colon S_n \to \{-1, 1\}$. The set $\{-1, 1\}$ is a (very small) group, under the binary operation of multiplication. Proposition 4.9 then says that sgn is a *group homomorphism.* Although we won't say much more about it in this course, it's a concept that plays a huge rôle in the more advanced study of groups, and indeed throughout the entire area of abstract algebra. ◊

We now return to our development of general group theory.

4.6. Subgroups

It is often useful to analyse whether groups can contain smaller groups, or how a group can fit into a larger one. For example, consider the following subset of the permutation group S_5, presented in cycle notation:

$$H = \{(1\ 2\ 3), (1\ 3\ 2), (1\ 2), (1\ 3), (2\ 3), e\}.$$

This looks suspiciously like the group S_3: it has the same number of elements (namely, 6) and they only permute the integers $1, 2, 3$. Note however that these are permutations of the set $\{1, 2, 3, 4, 5\}$; the cycle notation suppresses the elements that don't move! It turns out that H is a "subgroup" of S_5; a "copy" of S_3 sitting inside S_5. We will verify this after we've given the formal definition.

Definition 4.15 (Subgroup)**.**
Let $(G, \cdot)$ be a group. A *subgroup* of G is a subset $H \subseteq G$ such that $(H, \cdot)$ is a group in its own right. ♦

Note. It's very important that the group multiplication of H is the *same* as that of G; otherwise, there's no connection between the two groups and the concept becomes meaningless. ◊

To verify that a subset $H \subseteq G$ is a subgroup, we need to check the following three things, which we refer to as the *subgroup properties.*

SG1) H is non-empty. This may seem almost too obvious to state!

SG2) For all $x, y \in H$, $x \cdot y \in H$. We say: "H is closed under multiplication".

SG3) For all $x \in H$, $x^{-1} \in H$. We say: "H is closed under inverse".

The subgroup properties make no mention of the group identity element, and the following result explains why.

Proposition 4.10 (Presence of the identity element)**.**
Suppose $(H, \cdot)$ is a subgroup of $(G, \cdot)$. The identity element of G is contained in H, and is the identity element of H.

Proof. Direct and straightforward.

Since $H \neq \emptyset$ it contains an element, say x. So, since H is closed under inverse, it also contains x^{-1}. Then, since H is closed under multiplication, it contains $x \cdot x^{-1} = e$. Finally, since e satisfies:

$$x \cdot e = x = e \cdot x,$$

for all $x \in G$, and hence for all $x \in H$, it follows from Proposition 4.1 (i) that e is the identity element of H. ■

Note. The proof of Proposition 4.10 used all three of the subgroup properties. ◊

Example 4.13 (Subgroup of S_5)**.**
Let's take a closer look at the example from the beginning of this Section, and verify that H is indeed a subgroup of S_5.

SG1) Clearly $H \neq \emptyset$.

SG2) To make a compelling argument for this, we shift attention from those integers moved by elements of H to those that aren't. (Changing our point of view like this can often lead to a clearer way of doing things; proof by contrapositive is a good example.) Thus $\sigma \in H$ if and only if $\sigma(4) = 4$ and $\sigma(5) = 5$. So, suppose $\sigma, \tau \in H$. Then bearing in mind that multiplication in S_5 is composition of functions:

$$\sigma\tau(4) = \sigma(\tau(4)) = \sigma(4) = 4, \qquad \sigma\tau(5) = \cdots = 5.$$

Hence $\sigma\tau \in H$.

SG3) Similarly, if $\sigma \in H$ then:

$$\sigma^{-1}(4) = \sigma^{-1}(\sigma(4)) = 4, \qquad \sigma^{-1}(5) = \cdots = 5.$$

Hence $\sigma^{-1} \in H$.

We have completed our checks: H is indeed a subgroup. □

Remark 4.17. We noted at the beginning of Section 4.6 that, despite their obvious similarity, H is *not* the group S_3. So we can't write $H = S_3$. Instead, we write:

$$H \cong S_3,$$

and say: "H is isomorphic to S_3". We won't give a precise definition (there will certainly be one in the Introduction to Group Theory module in Year 2), and won't be mentioning it much more in this course. Suffice to say that "≅" is an equivalence relation on the set/class of all groups (!), and plays an important rôle in the further development of group theory. ◊

4.6.1. Cyclic subgroups

Let $(G, \cdot)$ be any group, and let $a \in G$ be an element of finite order; say $o(a) = n$ (see Definition 4.5). We would like to construct, in the most economical way possible, a subgroup of G containing a. How should we do this? Since subgroups are closed under multiplication, any such subgroup must contain all powers a^2, a^3 etc. Now, for any $m \in \mathbb{N}$ we have:

$$a^{m+n} = a^m \cdot a^n = a^m \cdot e = a^m.$$

So the list of powers of a repeats itself after the first n entries:

$$a, a^2, \ldots, a^n = e,\ a^{n+1} = a \cdot e = a,\ a^{n+2} = a^2, \ldots$$

We therefore focus our attention on the first n powers of a.

Definition 4.16 (Cyclic subgroup)**.**
Suppose $o(a) = n$ and let $\langle a \rangle \subseteq G$ denote the following subset:

$$\langle a \rangle = \{a, a^2, \ldots, a^n = e\}.$$

This is called the *cyclic subgroup* of G *generated* by a. ♦

To justify the terminology we need to show that $\langle a \rangle$ is a subgroup of G. Let's check the subgroup properties; it's closure under inverse that requires particular scrutiny.

SG1) It's clear that $\langle a \rangle$ is non-empty!

SG2) If $a^j, a^k \in \langle a \rangle$ then by Proposition 4.3 (ii):

$$a^j \cdot a^k = a^{j+k} = a^r,$$

where r is the least residue of $j + k$ modulo n. So $\langle a \rangle$ is closed under multiplication (as expected).

SG3) Suppose $a^k \in \langle a \rangle$, so $k = 1, \ldots, n$. Then $n - k = 0, \ldots, n - 1$, so bearing in mind that $a^0 = e$, we have $a^{n-k} \in \langle a \rangle$ and:

$$a^{n-k} \cdot a^k = a^k \cdot a^{n-k} = a^n = e.$$

Hence $(a^k)^{-1} = a^{n-k}$, so $\langle a \rangle$ is closed under inverse.

The check is complete: $\langle a \rangle$ is indeed a subgroup!

Remark 4.18. We can think of $\langle a \rangle$ as the "smallest" subgroup of G containing a. More precisely, any subgroup containing the element a must also contain every other element of $\langle a \rangle$. For, if $H \subseteq G$ is a subgroup containing a then since H is closed under multiplication it must also contain a^2, a^3 etc. So $\langle a \rangle \subseteq H$. ♢

From its definition it looks as if $\langle a \rangle$ has n elements. However, we don't yet know for sure that $a, a^2, \ldots, a^n$ are distinct. The following result assures us that they are, and in so doing explains the dual use of the word "order" in group theory.

Proposition 4.11 (Order of cyclic subgroups)**.**
For any group $(G, \cdot)$, if $a \in G$ is an element of finite order then $\langle a \rangle$ is a finite subgroup of G whose order is $o(a)$.

Proof. This is a direct proof, that brings into play our definition of the cardinality of a finite set (Definition 1.8).

Define a function $f\colon [n] \to \langle a \rangle$ by the rule $f(r) = a^r$ for all $r = 1, \ldots, n$. Then f is onto, by definition of $\langle a \rangle$ (Definition 4.16). Furthermore f is one-to-one. For, applying the "standard routine" (see Definition 1.3): if $f(r) = f(s)$ then $a^r = a^s$, hence $r \equiv s \pmod{n}$ by Proposition 4.3 (ii). This means that $s = r + kn$ for some $k \in \mathbb{Z}$, which since $0 \leqslant s \leqslant n$ can only happen if $k = 0$. Therefore $r = s$, as required. So f is a bijection, and it follows from Definition 1.8 that:

$$|\langle a \rangle| = n = o(a),$$

which completes the argument. ■

We showed in Proposition 4.3 (iv) that that every element of a finite group has finite order. Proposition 4.11 tells us that this order cannot excede the order of the group.

Corollary 4.12 (Inequality between orders in groups)**.**
If $(G, \cdot)$ is a finite group and $a \in G$ then $o(a) \leqslant |G|$.

Proof. We have $o(a) = |\langle a \rangle|$ from Proposition 4.11. But $\langle a \rangle \subseteq G$, and G is finite, so $|\langle a \rangle| \leqslant |G|$ by Proposition 1.5. ■

It's conceivable that constructing a cyclic subgroup may require *all* the elements of G. If this happens it tells us something rather special about the group G itself.

Definition 4.17 (Cyclic group)**.**
A finite group $(G, \cdot)$ is said to be *cyclic* if $G = \langle a \rangle$ for some $a \in G$; ie. every element of G is a power of a. An element a with this property is said to be a *generator* of G. ♦

Remarks 4.19 (Cyclic groups).

1) Definition 4.17 is not asking that $G = \langle a \rangle$ for *all* $a \in G$. Neither is it suggesting that there is only one element a with this property. Our upcoming examples will show that cyclic groups can more than one generator, but that not every element is necessarily a generator. In fact, determining which elements are generators and which aren't can sometimes be quite a headache!

2) It follows immediately from Definitions 4.17 and 4.16 that cyclic groups are abelian. The converse is false; however, it's almost true: just as every integer can be factored into a product of primes, every finite abelian group can be "factored" into a "product" of cyclic groups. This statement obviously requires some clarification, and then proof, both of which belong to a more advanced study of group theory. ◊

Example 4.14 (Generators of the cyclic group $(\mathbb{Z}_n, +)$)**.**
Let $(G, \cdot)$ be the group $(\mathbb{Z}_n, +)$ (see Section 4.3.2). Because the binary operation is addition, rather than multiplication, powers of elements are iterated sums, rather than products. So, for all non-negative integers r, the r-th "power" of an element $\bar{a} \in \mathbb{Z}_n$ is its r-th multiple $r\bar{a}$, defined:

$$\begin{aligned} r\bar{a} &= \bar{a} + \cdots + \bar{a} \quad (r \text{ times}) \\ &= \overline{a + \cdots + a}, \quad \text{by the laws of modular arithmetic} \\ &= \overline{ra}. \end{aligned}$$

Then for all $a = 0, 1, \ldots, n-1$ we have:

$$\bar{a} = \overline{a1} = a\bar{1},$$

which shows that every element of $\mathbb{Z}_n$ is a "power" of $\bar{1}$. So G is a cyclic group, with generator $\bar{1}$.

If $n > 2$ then there are other generators of G. To see this, note first that if $\bar{c}$ is a generator of $\mathbb{Z}_n$ then in particular $\bar{1} = k\bar{c}$ for some integer k, which by the definition of equality in $\mathbb{Z}_n$ means:

$$1 = kc + \ell n,$$

for some integer ℓ. This implies that 1 is the highest common factor of c and n (because any common factor of c and n must then also divide 1); ie. c and n are coprime. Now suppose c is coprime to n. Then by Bézout's identity (see Appendix A.2) there exist integers $k, \ell \in \mathbb{Z}$ such that:

$$kc + \ell n = 1.$$

Hence:

$$a = akc + a\ell n,$$

therefore:

$$\bar{a} = \overline{ak}\,\bar{c} = \bar{r}\bar{c} = \overline{rc} = r\bar{c},$$

where r is the least residue of ak modulo n. Thus every element of $\mathbb{Z}_n$ is a multiple of $\bar{c}$, so $\bar{c}$ is a generator of G. In summary, $\bar{c}$ is a generator of $\mathbb{Z}_n$ if and only if c is coprime to n.

As an illustration, let's consider the case $n = 6$. To determine all the generators we need only consider the integers $2, 3, 4, 5$, of which only 5 is coprime to 6; so apart from $\bar{1}$ there is only one other generator of $\mathbb{Z}_6$, namely $\bar{5}$. Indeed, we have:

$$2\,\bar{5} = \overline{10} = \bar{4}, \qquad 3\,\bar{5} = \overline{15} = \bar{3}, \qquad 4\,\bar{5} = \overline{20} = \bar{2}, \qquad 5\,\bar{5} = \overline{25} = \bar{1}, \qquad 6\,\bar{5} = \overline{30} = \bar{0},$$

so every element of $\mathbb{Z}_6$ is a multiple of $\bar{5}$, as expected. On the other hand, the possible multiples of $\bar{2}$ and $\bar{4}$ are:

$$\begin{array}{ccc} 2\,\bar{2} = \bar{4}, & 3\,\bar{2} = \bar{6} = \bar{0}, & 4\,\bar{2} = \bar{8} = \bar{2}, \\ 2\,\bar{4} = \bar{8} = \bar{2}, & 3\,\bar{4} = \overline{12} = \bar{0}, & 4\,\bar{4} = \overline{16} = \bar{4}, \end{array}$$

so these are both generators of the cyclic subgroup:

$$\langle\bar{2}\rangle = \langle\bar{4}\rangle = \{\bar{0}, \bar{2}, \bar{4}\}.$$

Finally, it's easy to show that $\langle\bar{3}\rangle = \{\bar{0}, \bar{3}\}$.

This accounts for all possible cyclic subgroups of $\mathbb{Z}_6$; in fact, we have found all possible subgroups, since every subgroup of a cyclic group is cyclic (see Proposition 4.13 below). Notice also that the orders of these subgroups are all factors of 6, which is the order of the group (cf. Example 4.3). This is no coincidence (see Proposition 4.13 again). □

Example 4.15 (Cyclic subgroups of $\mathbb{Z}_{11}^*$).
Let $G = \mathbb{Z}_{11}^*$ with the binary operation of multiplication modulo 11. In Example 4.3 we showed that $\bar{7}$ has order 10; therefore $\mathbb{Z}_{11}^*$ is cyclic, and $\bar{7}$ is a generator. The elements $\bar{2}$, $\bar{6}$ and $\bar{8}$ also have order 10, so these are also generators of G (which shows again that

cyclic groups can have more than one generator). In classical terminology (dating back to Gauss [18], and possibly earlier), generators of $\mathbb{Z}_{11}^*$ are called *primitive roots of unity* modulo 11; the phrase "roots of unity modulo 11" means solutions of the congruence $x^m \equiv 1 \pmod{11}$ for some exponent m (in other words, all integers other than multiples of 11), and "primitive" means "basic", or "fundamental", in the sense that all roots of unity can be expressed in terms of them.

It follows from our calculations for $\bar{3}$, $\bar{4}$, $\bar{5}$ and $\bar{9}$ that:

$$\langle\bar{3}\rangle = \langle\bar{4}\rangle = \langle\bar{5}\rangle = \langle\bar{9}\rangle = \{\bar{3}, \bar{9}, \bar{4}, \bar{5}, \bar{1}\}.$$

So this cyclic subgroup is generated by each of its non-identity elements. Finally:

$$\langle\overline{10}\rangle = \{\overline{10}, \bar{1}\}.$$

We have found all the cyclic subgroups of G, and therefore all possible subgroups (see Proposition 4.13 below). □

Example 4.16 (Cyclic and non-cyclic $\mathbb{Z}_n^\times$)**.**
Suppose first that $G = \mathbb{Z}_8^\times$ with the binary operation of multiplication modulo 8. In Example 4.2 we showed that the three non-identity elements of G have order 2. Therefore G is *not* cyclic. There are three cyclic subgroups of order 2:

$$\langle\bar{3}\rangle = \{\bar{3}, \bar{1}\}, \qquad \langle\bar{5}\rangle = \{\bar{5}, \bar{1}\}, \qquad \langle\bar{7}\rangle = \{\bar{7}, \bar{1}\},$$

and these are the only subgroups of G (see Proposition 4.13 below).

By contrast, suppose now that $G = \mathbb{Z}_9^\times$:

$$G = \{\bar{1}, \bar{2}, \bar{4}, \bar{5}, \bar{7}, \bar{8}\}.$$

Then:

$$\bar{2}^2 = \bar{4}, \qquad \bar{2}^3 = \bar{8}, \qquad \bar{2}^4 = \bar{7}, \qquad \bar{2}^5 = \bar{5}, \qquad \bar{2}^6 = \bar{1};$$

so G is cyclic, with generator $\bar{2}$. Similar calculations reveal that it's also generated by $\bar{5}$, whereas:

$$\langle\bar{4}\rangle = \{\bar{4}, \bar{7}, \bar{1}\} = \langle\bar{7}\rangle, \qquad \langle\bar{8}\rangle = \{\bar{8}, \bar{1}\}.$$

Again, these are the only subgroups of G.

Notice that the order of every subgroup is again a factor of the group order. □

Remark 4.20 (Primitive roots of unity).
Examples 4.15 and 4.16 raise the following interesting question: for which values of n is the group $\mathbb{Z}_n^\times$ cyclic? This was answered by Gauss (although not in the language of group theory) in his ground breaking study of number theory *Disquitiones Arithmeticae* [18]. He showed that the values of n for which a primitive root of unity modulo n exists

are precisely: if n is odd then $n = p^k$ where $p \geqslant 3$ is prime (an "odd prime") and k is a positive integer (we showed that p^k is odd in Remark 3.9 (4)), and if n is even then $n = 2$, $n = 4$ or $n = 2p^k$. Notice that all our examples are consistent with these values; in particular, $n = 8$ is *not* on the list.

This question of which groups $\mathbb{Z}_n^\times$ are cyclic leads immediately to another: given a value of n on Gauss' list, which elements of $\mathbb{Z}_n^\times$ are generators? Gauss wasn't able to answer this; and to date this remains an open question! ♢

Example 4.17 (Permutation groups are not cyclic).
Let $G = S_n$. If $\rho \in S_n$ is an r-cycle (Definition 4.8) then $o(\rho) = r$ by Proposition 4.5 (i), hence $\langle\rho\rangle$ is a subgroup of order r by Proposition 4.11. Then, since S_n has order $n!$ it follows that S_n is generated by a cycle only if $n! = n$, which is the case only when $n = 2$. In general, although the order of a permutation σ can easily be calculated using Proposition 4.5 (ii), determining the *maximum* possible order of σ is a somewhat trickier exercise. However, it follows immediately from Remark 4.19 (2) that no permutation group other than S_2 is cyclic, because S_n is non-abelian when $n \geqslant 3$. □

Some of the features that emerge from this sequence of examples turn out to be true in general. The following result ties them together for us.

Proposition 4.13 (Elementary properties of finite cyclic groups).
Suppose $(G, \cdot)$ is a finite cyclic group.

i) *Every subgroup of G is cyclic.*

ii) *The order of every subgroup of G divides the order of G.*

iii) *The order of every element of G divides the order of G.*

iv) *For all positive integers m, if m divides the order of G then G has an element of order m and a subgroup of order m.*

Proof. Each of these is a direct proof, and part (i) reacquaints us with the Principle of Mutual Containment.

Suppose $|G| = n$, and $a \in G$ is a generator; thus $o(a) = n$ by Proposition 4.11.

(i) Suppose $H \subseteq G$ is a subgroup, and let p be the *least* positive integer such that $a^p \in H$. We claim that $H = \langle b\rangle$ where $b = a^p$. To show this we will demonstrate the inclusions $H \subseteq \langle b\rangle$ and $\langle b\rangle \subseteq H$, and appeal to the Principle of Mutual Containment.

First, since $b \in H$ and H is closed under group multiplication, we have $\langle b\rangle \subseteq H$ (cf. Remark 4.18). On the other hand, if $h \in H$ then $h = a^s$ for some integer $s \geqslant p$. Writing $s = qp + r$ where $q \geqslant 1$ and $0 \leqslant r \leqslant p - 1$, it follows from the elementary laws of exponentiation in groups (4.2) that:

$$h = a^{qp+r} = a^{qp} \cdot a^r = (a^p)^q \cdot a^r = b^q \cdot a^r = k \cdot a^r,$$

where $k \in H$. This equation rearranges to:

$$a^r = k^{-1} \cdot h,$$

whose right hand side is an element of H, since H is closed under inverse. Since a^p is the least positive power that lies in H, this can only happen if $r = 0$, in which case $a^r = e$, and therefore $h = b^q$. Thus $h \in \langle b \rangle$, hence $H \subseteq \langle b \rangle$. This proves our claim, and we conclude that every subgroup H of G is cyclic.

(ii) Suppose $H \subseteq G$ is a subgroup, and $|H| = m$. By (i) we have that $H = \langle b \rangle$ for some element $b \in H$, and $o(b) = m$ by Proposition 4.11. Recalling Definition 4.5, m is the least positive integer satisfying:

$$e = b^m = (a^p)^m = a^{mp}.$$

Hence mp is a multiple of $o(a) = n$, by Proposition 4.3 (i). But $mp \leqslant |G| = n$, by Corollary 4.12. Therefore $mp = n$, which shows that $|H|$ divides $|G|$.

(iii) Every element $b \in G$ generates a cyclic subgroup $\langle b \rangle$ of order $o(b)$ (Proposition 4.11). Therefore $o(b)$ divides $|G|$ by (ii).

(iv) Finally, suppose that m divides $|G|$; thus $n = mp$ for some $p \in \mathbb{N}$. Let $b = a^p$. Then:

$$b^m = (a^p)^m = a^{mp} = a^n = e,$$

and since $o(a) = n$ there is no smaller positive power of b with this property (Definition 4.5). Therefore $o(b) = m$. Hence $\langle b \rangle \subseteq G$ is a subgroup of order m, by Proposition 4.11. ■

Although Proposition 4.13 is only valid for cyclic groups, there are some features that carry across to all finite groups; this is something we will explore further in Set Piece 4.

Remark 4.21 (Infinite cyclic groups).
It is possible to define cyclic groups and subgroups of infinite order. Suppose $a \in G$ is an element of infinite order. Then G is necessarily an infinite group, by Proposition 4.3 (iv) (via the contrapositive!), and we simply define:

$$\langle a \rangle = \{a^k \mid k \in \mathbb{Z}\}.$$

Thus $\langle a \rangle$ is the set of *all* positive powers of a and a^{-1}, together with $a^0 = e$. It follows almost immediately from this definition that $\langle a \rangle$ satisfies the subgroup properties, and is therefore a subgroup of G. It's also easy to check (using Definition 4.5) that all its elements are distinct, so $\langle a \rangle$ is countably infinite; indeed, the function $\mathbb{Z} \to \langle a \rangle$ with rule $k \mapsto a^k$ is a bijection. Again, we call $\langle a \rangle$ the *cyclic subgroup generated by a.* And if $G = \langle a \rangle$ then G is a *cyclic group;* so, every element of G is some power (positive, negative or zero) of a. Like their finite cousins, infinite cyclic groups are abelian.

The most familiar example of an infinite cyclic group is $(\mathbb{Z}, +)$. Since the group operation is addition rather than multiplication (cf. Example 4.15), this means that every element is a multiple (positive, negative or zero) of some integer a, and a moment's thought shows that there are only two possibilities: $a = \pm 1$. ♢

4.6.2. The alternating subgroup

This particular subgroup turns out to have great significance in the theory of finite groups. It's a subgroup of $G = S_n$, the group of permutations of $\{1, \ldots, n\}$ (Definition 4.7), but isn't simply a copy of a symmetric group of lower order. Recalling Definitions 4.13 and 4.14 of the parity and signature of a permutation, it's defined as follows.

Definition 4.18 (Alternating group)**.**
For all $n \geqslant 2$ define the subset $A_n \subset S_n$ by:

$$A_n = \{\sigma \in S_n : \sigma \text{ is even}\} = \{\sigma \in S_n : \operatorname{sgn}(\sigma) = 1\}. \qquad \blacklozenge$$

As implied by the title of Definition 4.18, A_n is a subgroup of S_n. Let's run through the subgroup properties.

SG1) The identity element e of S_n has transposition number 0 and is therefore even; hence $A_n \neq \emptyset$.

SG2) If $\sigma, \tau \in A_n$ then (Proposition 4.9):

$$\operatorname{sgn}(\sigma\tau) = \operatorname{sgn}(\sigma)\operatorname{sgn}(\tau) = 1 \times 1 = 1;$$

hence $\sigma\tau \in A_n$. Thus A_n is closed under multiplication.

SG3) If $\sigma \in A_n$ then taking the signature of both sides of the equation $\sigma\sigma^{-1} = e$ and applying Proposition 4.9 yields:

$$1 = \operatorname{sgn}(e) = \operatorname{sgn}(\sigma)\operatorname{sgn}(\sigma^{-1}) = 1 \times \operatorname{sgn}(\sigma^{-1}) = \operatorname{sgn}(\sigma^{-1});$$

hence $\sigma^{-1} \in A_n$. Thus A_n is closed under inverse.

It follows that A_n is a group in its own right; it's called the *alternating group.*

Remark 4.22. For all $n \geqslant 2$ let's define $B_n \subset S_n$ by:

$$B_n = \{\sigma \in S_n : \sigma \text{ is odd}\} = \{\sigma \in S_n : \operatorname{sgn}(\sigma) = -1\}.$$

However B_n is *not* a subgroup of S_n. *(Can you see which of the subgroup properties fail?)* ♢

The first question that comes to mind concerning A_n is probably: "What is its order?"; in other words, how many of the $n!$ elements of S_n are even permutations? It's certainly the case that the two subsets A_n, B_n partition S_n (see Definition 3.8):

$$S_n = A_n \cup B_n, \qquad A_n \cap B_n = \emptyset,$$

because every permutation is even or odd, and no permutation is even and odd (by Theorem 4.8). We might expect the two subsets to be of equal size, if only because that's a familiar property of even and odd natural numbers (although of course $\mathbb{N}$ is an infinite set, so "equal size" means "same cardinality", which we showed was true in Set Piece 1). In fact, our intuition turns out to be correct!

Proposition 4.14 (Order of the alternating group)**.**
The order of A_n is half that of S_n; ie. $|A_n| = n!/2$.

Note. The order of S_n is even, since $n!$ has a factor of 2 if $n \geqslant 2$. ◊

Proof. This is essentially a counting argument. We show that there are exactly the same number of even permutations as odd ones by constructing a bijection between A_n and B_n (cf. the proof of Proposition 2.8).

Let $\tau \in S_n$ be a fixed transposition and define a function $f\colon A_n \to B_n$ by $f(\sigma) = \tau\sigma$. We note first that $\tau\sigma$ is indeed an odd permutation, since:

$$\operatorname{sgn}(\tau\sigma) = \operatorname{sgn}(\tau)\operatorname{sgn}(\sigma) = -1 \times 1 = -1.$$

Therefore f is well-defined.

Now f is one-to-one, because:

$$f(\sigma) = f(\sigma') \implies \tau\sigma = \tau\sigma' \implies \sigma = \sigma',$$

by left cancellation (Proposition 4.1 (v)) (which is nothing more than multiplication on the left by $\tau^{-1} = \tau$). Furthermore f is onto. For, suppose ρ is an odd permutation. The permutation $\tau\rho$ is then even, because:

$$\operatorname{sgn}(\tau\rho) = \operatorname{sgn}(\tau)\operatorname{sgn}(\rho) = -1 \times -1 = -1,$$

and:

$$f(\tau\rho) = \tau(\tau\rho) = (\tau\tau)\rho = \tau^2\rho = \rho,$$

since $\tau^2 = e$. (We have, perhaps rather pedantically, shown the brackets being shifted to remind ourselves that the associative law is a constant necessity when dealing with groups.) Thus f is a bijection, and we deduce that $|A_n| = |B_n|$. Since the sets A_n, B_n partition S_n we have:

$$|S_n| = |A_n| + |B_n| = 2|A_n|,$$

and the result follows. ■

The even/odd partition of S_n may be summarised by the following simple diagram (Figure 4.1).

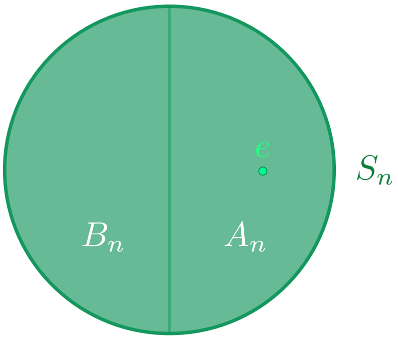


Figure 4.1.: Even/odd partition of the symmetric group

Remark 4.23. The fact that $|A_n|$ divides $|S_n|$ is not a coincidence; it's a special case of a general phenomenon in group theory that we will investigate further in Set Piece 4, where we will also take a closer look at the structure of the alternating group A_4 to see what further subgroups it contains (Example 4.19). ◊

4.6.3. Stabiliser subgroups and infinite permutations

We can find many more subgroups of S_n by generalising the idea that we used to check the subgroup properties in Example 4.13. In fact, with virtually no extra effort this allows us to construct subgroups of *any* symmetric group, not just the finite ones. Recall that the trick was to shift our attention to the elements that are fixed by a permutation, rather than those that are moved by it.

Definition 4.19 (Stabilisers)**.**
Let X be a set and $Y \subseteq X$ any subset. The following subset of the symmetric group S_X:

$$S_{X,Y} = \{\sigma \in S_X : \sigma(y) = y \text{ for all } y \in Y\},$$

is called the *stabiliser* of Y. ♦

Note. Recall that the symmetric group S_X is the set of all bijections $\sigma\colon X \to X$, which we refer to as "symmetries" of X (Definition 4.6). So $S_{X,Y}$ is the set of symmetries of X that are also symmetries of Y. ◊

Here's the nice thing about stabilisers:

Proposition 4.15 (Stabiliser subgroups)**.**
For any $Y \subseteq X$ the stabiliser $S_{X,Y}$ is a subgroup of S_X.

Proof. This is a direct verification of the subgroup properties. We elevate the argument of Example 4.13 to a proof!

SG1) The identity function $1_X\colon X \to X$ satisfies $1_X(y) = y$ for all $y \in Y$, and therefore belongs to $S_{X,Y}$.

SG2) Suppose $\sigma, \tau \in S_{X,Y}$. Then for all $y \in Y$:

$$\sigma\tau(y) = \sigma(\tau(y)) = \sigma(y) = y.$$

Hence $\sigma\tau \in S_{X,Y}$

SG3) Suppose $\sigma \in S_{X,Y}$. Then for all $y \in Y$:

$$\sigma^{-1}(y) = \sigma^{-1}(\sigma(y)) = y.$$

Hence $\sigma^{-1} \in S_{X,Y}$.

This completes the checklist; so $S_{X,Y}$ is a subgroup. ■

Example 4.18 (Infinite permutations)**.**
We illustrate the use of Proposition 4.15 when $X = \mathbb{N}$; thus we look at the group $S_{\mathbb{N}}$ of "infinite permutations".

For $n \in \mathbb{N}$ we define the subset $Y_n \subset \mathbb{N}$ by:

$$Y_n = \{n+1, n+2, \dots\} = \{m \in \mathbb{N} : m > n\} = \mathbb{N} \smallsetminus \{1, \dots, n\},$$

and denote its stabiliser $S_{\mathbb{N},Y_n}$ by $S_{\mathbb{N},n}$. It follows from Proposition 4.15 that $S_{\mathbb{N},n}$ is a subgroup of $S_{\mathbb{N}}$, whose elements permute only the integers $1, \dots, n$; so $S_{\mathbb{N},n}$ is a copy of S_n sitting inside $S_{\mathbb{N}}$. (However, it would not be correct to write $S_{\mathbb{N},n} = S_n$; see Remark 4.17.) In this way, $S_{\mathbb{N}}$ contains copies of *all* the finite symmetric groups S_n. Since S_n gets very large, very quickly (Remark 4.10), we may well ask: "Just how big is $S_{\mathbb{N}}$?". We would certainly expect $|S_{\mathbb{N}}|$ to be infinite (because $\mathbb{N}$ is infinite), but as we saw in Set Piece 2 there are different sizes of infinity. Since $\mathbb{N}$ is countably infinite we might expect the same to be true for $S_{\mathbb{N}}$. On the other hand, since the size of S_n grows so rapidly it is conceivable that this could push $S_{\mathbb{N}}$ towards uncountability. To resolve this teaser we'll use Cantor's Theorem (Set Piece Theorem 2). This requires us to make a link between $S_{\mathbb{N}}$ and the power set $\mathscr{P}(\mathbb{N})$.

Suppose $Y \in \mathscr{P}(\mathbb{N})$; thus $Y \subseteq \mathbb{N}$. Consider the complement Y^c, which is also a subset of $\mathbb{N}$, hence countable (Lemma 1.7). We can therefore list the elements of Y^c:

$$\begin{aligned} &a_1, \dots, a_n, &&\text{if } Y^c \text{ is finite,}\\ &a_1, a_2, \dots, &&\text{if } Y^c \text{ is infinite.}\end{aligned}$$

Let's define $\sigma_Y \in S_{\mathbb{N},Y}$ by:

$$\sigma_Y = (a_1 \cdots a_n), \quad \text{if } Y^c \text{ is finite,}$$

$$\sigma_Y = (a_1\, a_2)(a_3\, a_4)\cdots, \quad \text{if } Y^c \text{ is infinite.}$$

In both cases, σ_Y is a "permutation" that fixes precisely the numbers in Y.

We now define a function $f\colon \mathscr{P}(\mathbb{N}) \to S_{\mathbb{N}}$ by:

$$f(Y) = \sigma_Y.$$

This function is one-to-one. For, if $f(Y) = f(Z)$ then $\sigma_Y = \sigma_Z$, which by the Principle for Equality of Functions implies that σ_Y and σ_Z fix precisely the same numbers, so $Y = Z$. Recalling Definition 2.2, it follows that:

$$|\mathscr{P}(\mathbb{N})| \leqslant |S_{\mathbb{N}}|.$$

Therefore, since $\mathscr{P}(\mathbb{N})$ is uncountable by Cantor's Theorem (Set Piece Theorem 2), $S_{\mathbb{N}}$ is also uncountable. □

The "take away" from Example 4.18 is that $S_{\mathbb{N}}$ is uncountably infinite. From this it follows that if X is *any* infinite set then S_X is uncountable. *(Can you show this? You could try constructing a copy of $S_{\mathbb{N}}$ inside S_X using Proposition 4.15.)* This leaves us with the following interesting/strange fact: symmetric groups are either finite, or uncountably infinite; none are countably infinite!

4.7. Euclidean symmetry

We briefly discuss another important class of finite non-abelian groups, which arise when we study geometric symmetry. This is probably the kind of symmetry we're most familiar with, and it's high time for it to make an appearance!

Suppose we have some shape in the plane (a drawing or design, say). An age-old question is then: "How symmetric is this shape?" Mathematically, we may regard our shape as a subset of the Cartesian plane, say $X \subset \mathbb{R}^2$. One answer would then be to form the symmetric group S_X (Definition 4.6) and analyse its structure. However, we remarked at the end of Section 4.6 that if X contains infinitely many points (which will be the case for most shapes) then S_X is uncountable. For example, the shapes in Figure 4.2 certainly have infinitely many points, and therefore uncountably many symmetries; however the cardioid has only one (non-trivial) geometric symmetry—reflection in the vertical line through its cusp—whereas the windmill has a point of 4-fold rotational symmetry.

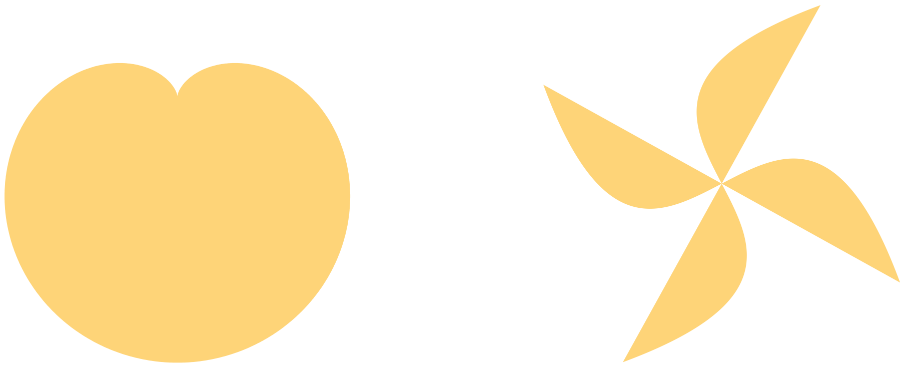


Figure 4.2.: Cardioid (left) and windmill (right)

The problem is that the general notion of a "symmetry" as a bijection (ie. a "rearrangement of points"), has little or nothing to say about "geometric symmetry", such as rotational symmetry, or reflectional symmetry. The "symmetries" that do this job for us are called "Euclidean symmetries", and are of a very special type. To give an accurate definition, we assume that X is a *bounded* subset of $\mathbb{R}^2$:

$$(\exists r > 0)(\forall (x, y) \in X)\ x^2 + y^2 < r^2.$$

This is a precise (but complicated) way of saying that X lies entirely within a disc of some radius r. Geometrically, this means that our shape occupies a finite area. Not all shapes/subsets have this property. *(Can you think of some examples of shapes that are unbounded?)*

Definition 4.20 (Euclidean symmetry)**.**
Suppose $X \subseteq \mathbb{R}^2$ is a bounded subset.

- A *Euclidean symmetry* of X is a bijection $\varphi \colon X \to X$ of the form:

  $$\varphi(x, y) = F(x, y), \quad \text{for all } (x, y) \in X,$$

  where $F \colon \mathbb{R}^2 \to \mathbb{R}^2$ is a rotation or reflection of the entire Cartesian plane (see equations (4.7) and (4.8) below).

- If F is a rotation then φ is called a *rotational symmetry,* and usually denoted (for phonetic reasons) by the Greek letter ρ; the centre of the rotation is called a *point of symmetry* for X.

- If F is a reflection then φ is called a *reflectional symmetry,* and usually denoted by σ; the mirror line is called a *line of symmetry* for X. ♦

Remarks 4.24.

1) Another way of stating Definition 4.20 is that a Euclidean symmetry of X is the restriction to X of a rotation or reflection of $\mathbb{R}^2$; see Remark 1.13.

2) The use of Greek letter σ for a reflection comes from the German "Spiegel", which means "mirror". *(Check out the beautiful piece of music: "Spiegel im Spiegel", by Arvo Pärt.)* It's not to be confused with its use in permutation theory. In fact, this is a situation where we have done a bit of "notation recycling" (in accordance with the "Principle of Conservation of Symbols")! $\Diamond$

Rotations and reflections of $\mathbb{R}^2$ are conceptually familiar. However, for the record, they are functions $\mathbb{R}^2 \to \mathbb{R}^2$ of the following type:

$$R(x,y) = \big(a + c(x-a) - s(y-b), b + s(x-a) + c(y-b)\big), \tag{4.7}$$

for a rotation about a point $(a,b) \in \mathbb{R}^2$, and:

$$S(x,y) = \big(a + c(x-a) + s(y-b), b + s(x-a) - c(y-b)\big), \tag{4.8}$$

for reflection in a line through (a,b). Here $c, s \in \mathbb{R}$ are numbers satisfying $c^2 + s^2 = 1$, which determine the angle of rotation, or the line of reflection. (Yes, they can be thought of as the cosine and sine of some angle!)

Equations (4.7) and (4.8) are clearly very similar, differing only by a couple of changes of sign. (Mathematical moral: changes of sign (in particular, sign errors) can have drastic consequences!) We won't worry about how the equations are derived. *(It's something you can look into if you like.)* From the equations, it's not difficult to show that R, S are bijections, by writing down their inverses (see Proposition 1.2):

$$R^{-1}(x,y) = \big(a + c(x-a) + s(y-b), b - s(x-a) + c(y-b)\big),$$

and noting that $S^{-1} = S$ ("reflecting twice gets us back where we started"; reflections are another example of an involution!). This means that Euclidean symmetries $\varphi\colon X \to X$ are indeed bijections.

4.7.1. Symmetries of a triangle, and dihedral groups

To get a better understanding of all this, let's look at an example. Suppose $X = \Delta \subset \mathbb{R}^2$ is an equilateral triangle, the size, position and orientation of which don't matter; see Figure 4.3. We label the vertices V_1, V_2, V_3, and the edges e_1, e_2, e_3, where e_i is the edge opposite V_i. The line m_i through V_i and the midpoint of e_i is a *median* of Δ. The three medians intersect at a common point O, called the *centre* (or *centroid*) of Δ (this is true for any triangle; it's an old and famous theorem of Euclidean geometry, and can also be proved quite efficiently using vectors).

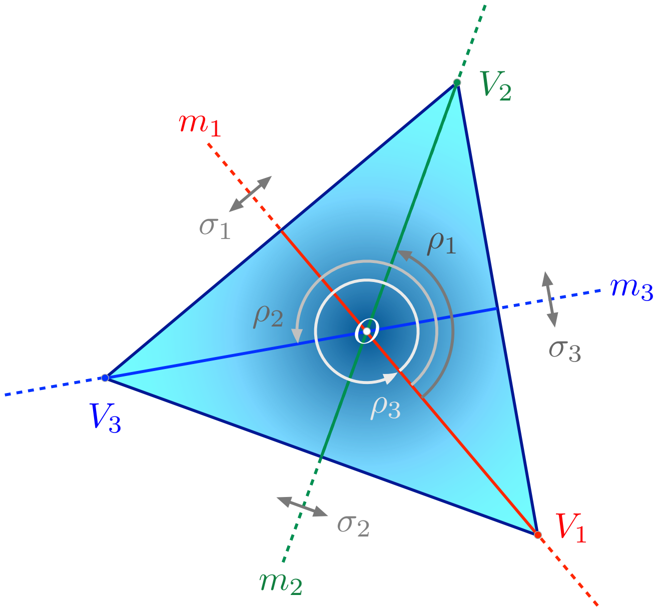


Figure 4.3.: Symmetries of an equilateral triangle

There are some familiar Euclidean symmetries of Δ:

$\sigma_i \colon \Delta \to \Delta, i = 1, 2, 3;$ reflection in m_i,

$\rho_j \colon \Delta \to \Delta, j = 1, 2, 3;$ anticlockwise rotation about O through angle $2j\pi/3$.

Although these are functions $\Delta \to \Delta$, they come from reflections or rotations that act on the entire plane and carry Δ along with them; this is the idea behind Definition 4.20. By the way, although we have measured the angles of rotation anticlockwise, we could equally well have chosen to measure them clockwise; it's a matter of taste, and we have to make a choice. The lines m_1, m_2, m_3 are lines of symmetry for X, and O is a point of (3-fold) symmetry.

Now here's our main result.

Theorem 4.16 (Symmetry group of an equilateral triangle)**.**
Let $\Delta \subset \mathbb{R}^2$ be an equilateral triangle. The set:

$$D_3 = \{\rho_1, \rho_2, \rho_3, \sigma_1, \sigma_2, \sigma_3\}$$

is a group under composition.

Proof. A direct proof, part of which is a "proof by exhaustion" (with three main cases and numerous sub-cases). Although the group operation is composition of functions, to streamline notation we omit the composition symbol "$\circ$" (in accordance with the "Principle of Notational Simplicity").

We check that D_3 is a subgroup of the symmetric group S_Δ. Since D_3 is clearly non-empty, we need only check the closure properties. We'll tackle inverses first, because it's easier, and we'll use it to help us with closure under multiplication.

Closure under inverse. We have $\sigma_i^{-1} = \sigma_i$ and $\rho_3 = e$, the identity function. So it suffices to note that:

$$\rho_1\rho_2 = e = \rho_2\rho_1, \tag{4.9}$$

from which it follows that $\rho_1^{-1} = \rho_2$ and $\rho_2^{-1} = \rho_1$.

Closure under multiplication. There are a number of different cases to consider, depending on whether we're composing two rotations, two reflections, or one of each.

Rotations. It follows from (4.9) and $\rho_3 = e$ that the product of rotations is always a rotation.

For other combinations, it's helpful to note that because rotations and reflections preserve distance they are characterised by their action on the vertices V_1, V_2, V_3, which they permute.

Reflections. To determine all products $\sigma_i\sigma_j$ we track the movement of the vertices. For example:

$$\sigma_1\sigma_2(V_1) = \sigma_1(V_3) = V_2, \quad \sigma_1\sigma_2(V_2) = \sigma_1(V_2) = V_3, \quad \sigma_1\sigma_2(V_3) = \sigma_1(V_1) = V_1,$$

which is precisely the same permutation of V_1, V_2, V_3 as ρ_1; hence:

$$\sigma_1\sigma_2 = \rho_1.$$

Using Proposition 1.3 and (i) then gives us the following "freebie":

$$\sigma_2\sigma_1 = \sigma_2^{-1}\sigma_1^{-1} = (\sigma_1\sigma_2)^{-1} = \rho_1^{-1} = \rho_2.$$

Similarly:

$$\sigma_1\sigma_3 = \rho_2,$$

with freebie:

$$\sigma_3\sigma_1 = \rho_1.$$

A final couple of tricks give us:

$$\sigma_2\sigma_3 = \sigma_2\sigma_1\sigma_1\sigma_3 = \rho_2^2 = \rho_1, \qquad \sigma_3\sigma_2 = \rho_1^{-1} = \rho_2.$$

The remaining relations between reflections are simply:

$$\sigma_i^2 = e = \rho_3.$$

So a product of reflections is always a rotation.

Rotations and reflections. To determine products of the form $\sigma_i\rho_j$ and $\rho_i\sigma_j$ we substitute the rotation by a product of reflections previously calculated. Thus:

$$\sigma_1\rho_1 = \sigma_1\sigma_1\sigma_2 = \sigma_2, \qquad \rho_1\sigma_1 = \sigma_3\sigma_1\sigma_1 = \sigma_3,$$
$$\sigma_2\rho_2 = \sigma_2\sigma_2\sigma_1 = \sigma_1, \qquad \rho_2\sigma_2 = \sigma_3\sigma_2\sigma_2 = \sigma_3,$$

$$\begin{aligned}
\sigma_1\rho_2 &= \sigma_1\sigma_1\sigma_3 = \sigma_3, & \rho_2\sigma_1 &= \sigma_2\sigma_1\sigma_1 = \sigma_2,\\
\sigma_2\rho_1 &= \sigma_2\sigma_2\sigma_3 = \sigma_3, & \rho_1\sigma_2 &= \sigma_1\sigma_2\sigma_2 = \sigma_1,\\
\sigma_3\rho_1 &= \sigma_3\sigma_3\sigma_1 = \sigma_1, & \rho_1\sigma_3 &= \sigma_2\sigma_3\sigma_3 = \sigma_2,\\
\sigma_3\rho_2 &= \sigma_3\sigma_3\sigma_2 = \sigma_2, & \rho_2\sigma_3 &= \sigma_1\sigma_3\sigma_3 = \sigma_1.
\end{aligned}$$

So, the product of a rotation and reflection is always a reflection. ■

Remarks 4.25 (Dihedral groups).

1) The group D_3 is called a *dihedral*[11] *group.* It should (strictly speaking) be denoted $D_3(\Delta)$ because moving, rotating or scaling Δ produces (strictly speaking) a different set of rotations and reflections. However, all these groups are isomorphic.

2) The group D_3 contains *all* the Euclidean symmetries of Δ. This is because, as remarked in the proof, Euclidean symmetries permute the vertices of Δ, and the destinations of all other points are then determined. Since $|D_3| = 6$ and there are precisely 6 permutations of V_1, V_2, V_3, there can be no further Euclidean symmetries.

3) The rotations ρ_1, ρ_2, ρ_3 form a subgroup of their own, which we denote by C_3.

4) The proof makes it clear that D_3 is non-abelian. However C_3 is abelian.

5) As the notation suggests, D_3 belongs to a family D_n of dihedral groups. For $n \geqslant 3$, D_n is the set of all Euclidean symmetries of a regular n-sided polygon; the exceptional cases D_1 and D_2 are the Euclidean symmetries of a (non-equilateral) isosceles triangle and (non-square) rectangle, respectively. It can be shown in all cases that D_n is a group, with $|D_n| = 2n$, and D_n non-abelian for all $n \geqslant 3$. The argument for general n is cleverer than our argument for D_3.

6) The dihedral group D_n contains precisely n rotations, and these form a subgroup C_n, which is abelian.

7) Since every element of D_3 corresponds to a permutation of the three vertices V_1, V_2, V_3, the dihedral group D_3 is isomorphic to the symmetric group S_3. The only other case where the families D_n and S_n overlap is D_1 and S_2, which are also isomorphic. ◊

4.8. Set Piece 4: Lagrange's Theorem

References.
Episode 18 of video lectures.

[11] "Dihedral" comes from the Greek "two faces", or literally "two seats". It refers to the physical interpretation of a reflection in the plane as a rotation in space that flips the shape upside down.

For our final Set Piece of the course, we're going to prove a fundamental result in abstract algebra which dates back to Lagrange, although our formulation and proof in modern mathematical language is rather different to his! This is a lovely proof, using many of the tools we've developed so far. A fitting *dénouement!*

To set the scene, it is worth noting that in all the examples of subgroups of finite groups that we've seen so far (cyclic subgroups the alternating subgroup, rotation subgroups, stabiliser subgroups) the order of the subgroup invariably divides the order of the group. In Proposition 4.13 we proved that this always happens in cyclic groups. Lagrange's Theorem tells us that the group doesn't have to be cyclic! Here's the statement.

Set Piece Theorem 4 (Lagrange's Theorem)**.**
Suppose $(G, \cdot)$ is a finite group and H is a subgroup of G. Then $|H|$ divides $|G|$.

Like many "good" theorems, Lagrange's theorem is easy to state. However, the proof involves a "big idea", which we introduce in the next section. After this, the proof will follow the same lines as that of Proposition 4.14.

4.8.1. Cosets

In order to prove Proposition 4.14 we partitioned the group in question (the symmetric group S_n) into the subgroup of interest (the alternating group A_n) and its complementary subset (the odd permutations B_n). We then constructed a bijection between these two subsets, thereby showing that they have the same number of elements, and allowing us to conclude that the order of A_n is half that of S_n. The situation was summarised diagrammatically in Figure 4.1. In the more general context of Lagrange's theorem we will adopt the same strategy: partition the group G into H and a number of other subsets, called "cosets" (short for "complementary subsets"), and then show that every set in the partition has the same size. The question is, therefore, how to construct such a partition. It turns out that there is more than one!

In Proposition 3.3 we showed that partitions are "the same as" equivalence relations. So we begin by turning the even/odd partition of S_n into an equivalence relation. Following the recipe of Proposition 3.3 this relation is:

$$\sigma \sim \tau \quad \text{if and only if} \quad \sigma, \tau \in A_n \ \text{ or } \ \sigma, \tau \in B_n, \tag{4.10}$$

for all $\sigma, \tau \in S_n$; in other words, two permutations are related if and only if they have the same parity (Definition 4.13). Notice that we can simplify this to:

$$\sigma \sim \tau \quad \text{if and only if} \quad \sigma\tau \in A_n, \tag{4.11}$$

because, recalling the definition (Definition 4.14) and fundamental property (Proposition 4.9) of the signature of permutations, σ and τ have the same parity if and only if:

$$\operatorname{sgn}(\sigma\tau) = \operatorname{sgn}(\sigma)\operatorname{sgn}(\tau) = 1.$$

Or, simply notice that if we factorise σ and τ into transpositions then by placing these side-by-side we get a factorisation of $\sigma\tau$ into an *even* number of transpositions, since σ and τ are both even or both odd.

Now, the relation (4.11) can easily be generalised to all groups G and subgroups H by defining:

$$a \sim b \quad \text{if and only if} \quad a \cdot b \in H, \tag{4.12}$$

for all $a, b \in G$. Unfortunately, however, this is not an equivalence relation! For example, it's not reflexive, since there is no guarantee in general that $a^2 \in H$; that this happens when $G = S_n$ and $H = A_n$ is just a quirk! So, although (4.11) is arguably the neatest way to condense (4.10) into a single condition, it's not going to give us what we want.

Luckily, there are other ways to express the relation (4.10) more concisely; for example, here are two:

$$\begin{aligned} \sigma \sim \tau \quad &\text{if and only if} \quad \sigma\tau^{-1} \in A_n, \\ \sigma \sim \tau \quad &\text{if and only if} \quad \sigma^{-1}\tau \in A_n. \end{aligned}$$

These are valid because a permutation always has the same parity as its inverse; we can see this by either applying the sgn function to both sides of the equation $\sigma\sigma^{-1} = e$, or observing that the inverse of a product of transpositions is the product of the same transpositions but in reverse order (see Proposition 4.1 (iv)). Both relations generalise easily to any group G (finite or infinite) and subgroup H. However, there is no reason to expect their generalisations to necessarily coincide, so we will exercise caution, denoting one by $\sim_R$ and the other by $\sim_L$ as follows:

$$a \sim_R b \quad \text{if and only if} \quad a \cdot b^{-1} \in H, \tag{4.13}$$

$$a \sim_L b \quad \text{if and only if} \quad a^{-1} \cdot b \in H, \tag{4.14}$$

for all $a, b \in G$. In each case, the introduction of a group inverse element neatly fixes the problem we were having with relation (4.12), and in fact produces an equivalence relation. Let's check this for $\sim_R$; the verification for $\sim_L$ is almost identical.

Reflexive. Since H is a subgroup it contains the identity element e (Proposition 4.10). Hence for all $a \in G$ we have $a \cdot a^{-1} = e \in H$. Thus $a \sim_R a$ for all $a \in G$.

Symmetric. Suppose $a \sim_R b$; thus $a \cdot b^{-1} \in H$. Then, using Proposition 4.1 (iv) we have:

$$b \cdot a^{-1} = (a \cdot b^{-1})^{-1} \in H,$$

since H is closed under inverse. Hence $b \sim_R a$.

Transitive. Suppose $a \sim_R b$ and $b \sim_R c$; thus $a \cdot b^{-1} \in H$ and $b \cdot c^{-1} \in H$. Then:

$$a \cdot c^{-1} = (a \cdot b^{-1}) \cdot (b \cdot c^{-1}) \in H,$$

since H is closed under multiplication. Hence $a \sim_R c$.

Remarks 4.26.

1) The subscripts "R" and "L" decorating the two relations are *(as you may have already guessed)* short for "right" and "left", because (for example) in relation (4.13) the inverse element appears on the right hand side of the product. However, some mathematicians adopt the opposite convention, labelling (4.13) as "left" and (4.14) as "right"!

2) If G is an abelian (ie. commutative) group, and H is any subgroup, then the relations $\sim_R$ and $\sim_L$ are the same. For:

$$\begin{aligned} a \sim_R b &\iff a \cdot b^{-1} \in H \\ &\iff a \cdot b^{-1} = h, \quad \text{for some } h \in H \\ &\iff b \cdot a^{-1} = h^{-1}, \quad \text{by Proposition 4.1 (iv)} \\ &\iff a^{-1} \cdot b = h^{-1}, \quad \text{since } G \text{ is abelian} \\ &\iff a^{-1} \cdot b \in H, \quad \text{since } H \text{ is closed under inverse} \\ &\iff a \sim_L b. \end{aligned}$$

3) If G is non-abelian then the argument of (2) breaks down, right at its midpoint, and it's not hard to find counter-examples which show that in general the two relations are different (see Remark 4.27 (4) below). However, this doesn't rule out the existence of particular combinations of G and H for which $\sim_R$ and $\sim_L$ coincide; for example, as we have already seen, when $G = S_n$ and $H = A_n$. When this happens we say that H is a *normal subgroup* of G; thus A_n is a normal subgroup of S_n, and all subgroups of an abelian group are normal. Despite the terminology, subgroups of non-abelian groups are not normally normal!

4) Another motivation for the relations $\sim_R$ and $\sim_L$, which provides further insight into the appearance of inverse elements, is to regard them as generalisations of congruence modulo n on $\mathbb{Z}$ (Section 3.3):

$$a \sim_n b \quad \text{if and only if} \quad a - b = kn \text{ for some } k \in \mathbb{Z}.$$

It's not hard to show that the following subset of $\mathbb{Z}$:

$$n\mathbb{Z} = \{nk : k \in \mathbb{Z}\}$$

is a subgroup. *(See whether you can do this.)* The relation of congruence can then be rephrased as:

$$a \sim_n b \quad \text{if and only if} \quad a - b \in n\mathbb{Z}.$$

This is precisely the relation (4.13) with $H = n\mathbb{Z}$, bearing in mind that group "multiplication" in $\mathbb{Z}$ is addition, so the inverse element of b is $-b$. The relation could equally well be written as:

$$a \sim_n b \quad \text{if and only if} \quad -a + b \in n\mathbb{Z},$$

which is precisely (4.14) in this context. Since $\mathbb{Z}$ is abelian, the difference between $\sim_R$ and $\sim_L$ doesn't show up in this example. ◊

Having defined two new equivalence relations, a natural question to ask is: "What are their equivalence classes?" Let's look at the relation $\sim_L$. We claim that the equivalence class $[a]_L$ of $a \in G$ is the following subset of G:

$$aH = \{a \cdot h \mid h \in H\}.$$

(The notation "aH" is "good", because it's not only easy on the eye but also reminds us of the definition; this checks all the boxes of the "Principle of Notational Simplicity".) To prove this equality of sets, we run through a sequence of bi-implications which shows that they have precisely the same elements, thereby directly verifying the Principle for Equality of Sets.

$$\begin{aligned} x \in [a]_L &\iff x \sim_L a, \quad \text{by Proposition 3.1} \\ &\iff a \sim_L x, \quad \text{by symmetry} \\ &\iff a^{-1} \cdot x \in H \\ &\iff a^{-1} \cdot x = h, \quad \text{for some } h \in H \\ &\iff x = a \cdot h, \quad \text{multiplying both sides on the right by } x \\ &\iff x \in aH. \end{aligned}$$

A similar argument shows that the equivalence class $[a]_R$ of a under the relation $\sim_R$ is the subset $Ha \subseteq G$ defined as the notation suggests:

$$Ha = \{h \cdot a \mid h \in H\}.$$

This brings us to the definition we have been building up to.

Definition 4.21 (Cosets)**.**
Let G be any group, and H any subgroup of G. For any element $a \in G$, the subset aH (resp. Ha) of G is called the *left* (resp. *right*) coset of H through a. ♦

Remarks 4.27 (Cosets).

1) Definition 4.21 does not assume that either G or H is finite; in fact, one or both could be infinite, the cosets themselves can also be infinite sets, and there could be infinitely many of them! Of course, if G is finite then so are H and all its cosets (Proposition 1.5), and there can only be a finite number.

2) Since $He = eH = H$, the subgroup H qualifies as both a left and a right coset. Although this is somewhat at odds with the "coset" terminology, it's more than outweighed by the convenience of having everything under the same umbrella.

3) Since $e \in H$ we have $a = e \cdot a \in Ha$ and $a = a \cdot e \in aH$ for all $a \in G$. Thus, the right and left cosets of H through a both contain a, as the terminology suggests.

4) In general, however, $aH \neq Ha$. For an easy counter-example, suppose $G = S_3$ and H is the following subgroup of order 2:

$$H = \{e, (1\ 2)\}.$$

The right cosets of H are:

$$H(1\ 3) = \{(1\ 3), (1\ 3\ 2)\}, \qquad H(2, 3) = \{(2\ 3), (1\ 2\ 3)\},$$

whereas its left cosets are:

$$(1\ 3)H = \{(1\ 3), (1\ 2\ 3)\}, \qquad (2\ 3)H = \{(2\ 3), (1\ 3\ 2)\}.$$

Since the right and left cosets are the equivalence classes of $\sim_R$ and $\sim_L$ respectively, this also shows that the two relations are different. On the other hand, notice that the various cosets of H all contain the same number (viz. 2) of elements.

5) Another way of saying that H is a normal subgroup (see Remark 4.26 (5)) is:

$$aH = Ha, \quad \text{for all } a \in G.$$

6) If $G = \mathbb{Z}$ and $H = n\mathbb{Z}$ (see Remark 4.26 (4)) then the right and left cosets of H coincide (since $\mathbb{Z}$ is abelian), and are nothing other than the congruence classes modulo n. ♢

The partition of G into the right and left cosets of H may be represented diagrammatically as curvaceous "slices of a pie" (Figure 4.4). To get the complete picture we need to know the comparitive size of each "slice". This is the essence of Lagrange's theorem.

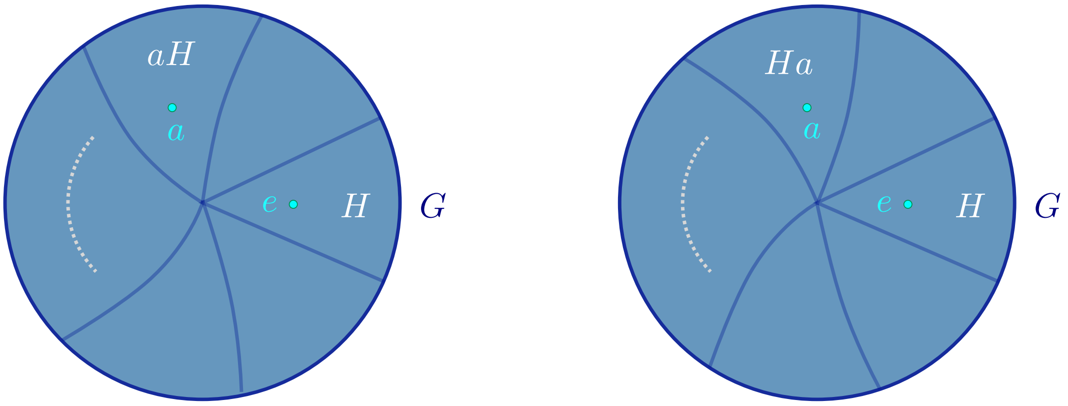


Figure 4.4.: Left (on the left) and right (on the right) cosets of a subgroup

4.8.2. The proof

We now have all the machinery we need to prove Lagrange's theorem.

Proof. This is essentially a counting argument.

We first show that each left coset of H has the same number of elements. To do this, we first fix an element $a \in G$ and then define a function: $f \colon H \to aH$ by:

$$f(h) = a \cdot h, \quad \text{for all } h \in H.$$

Then f is surjective by definition, and by left cancellation (Proposition 4.1 (v)):

$$f(h) = f(h') \;\Rightarrow\; a \cdot h = a \cdot h' \;\Rightarrow\; h = h',$$

so f is one-to-one. Thus f is a bijection, and it follows that $|aH| = |H|$.

We now count the elements of G, noting that since G is finite so is H, and there can be only finitely many cosets; say m. Then, since the cosets partition G and each coset has $|H|$ elements we have $|G| = m|H|$, which gives the result. ■

Remarks *4.28.*

1) In the proof we could equally well have used the right cosets of H, to conclude that $|Ha| = |H|$ for all $a \in G$. This not only proves Lagrange's Theorem (again!), but also shows that the right and left cosets have the same size, and there are exactly the same number of them.

2) The first part of the proof is valid for infinite groups and subgroups, and shows that all cosets have the same cardinality.

3) The number of right cosets of H is called the *index* of H, and denoted $[G : H]$. A more detailed statement of Lagrange's Theorem is therefore:

$$|G| = [G : H]\,|H|,$$

which is how it usually appears in textbooks.

4) The quotient set (see Definition 3.9) for the relation $\sim_L$ is usually denoted G/H, rather than $G/\sim_L$. This not only looks nicer, but also reminds us that its elements are the left cosets of H:

$$G/H = \{aH : a \in G\}.$$

Likewise, the quotient set for $\sim_R$ is denoted $H\backslash G$, reminding us that its elements are right cosets:

$$H\backslash G = \{aH : a \in G\}.$$

Notice that:

$$|G/H| = [G : H] = |H\backslash G|.$$

5) If H is a normal subgroup then since its right and left cosets are the same, the quotient sets G/H and $H\backslash G$ coincide. We usually choose to denote this set by G/H, rather than $H\backslash G$, simply because it looks somewhat prettier. The big surprise is that G/H is itself a group, called the *quotient group.* This is a fundamental construction of group theory, which features prominently in the further development of the subject; however it's beyond the scope of this course. ♢

Lagrange's Theorem is a striking result with many applications. One immediate consequence is that the order of a group element *always* divides the group order, something we already proved for cyclic groups (Proposition 4.13) and permutation groups (Proposition 4.5).

Corollary 4.17 (Divisibility property of orders of group elements)**.**
The order of any element of a finite group G divides the order of G. In particular, for all $a \in G$ we have $a^n = e$ where $n = |G|$.

Proof. This is a direct argument, combining what we know about orders of elements with Lagrange's theorem.

The order of a is the order of the cyclic subgroup $\langle a \rangle$ (Proposition 4.13), and the order of $\langle a \rangle$ divides the order of G by Lagrange's Theorem. Therefore, since n is a multiple of $o(a)$ it follows from Proposition 4.3 (i) that $a^n = e$. ■

Whenever we prove a theorem in mathematics the question invariably arises: "Is the converse true?" This is a very good question, because the statements $P \Rightarrow Q$ and $Q \Rightarrow P$ are logically independent. Now, the converse of Lagrange's Theorem is the following statement:

> *If G is a finite group and m is a positive integer that divides $|G|$ then G has a subgroup of order m.*

We have seen that this statement is true if G is a cyclic group (Proposition 4.13), which might raise our hopes of it being true in general. So it comes as a bit of a surprise to discover that it is in fact false! To show this we need a counter-example, and our next example (Example 4.19) is exactly that. (Remember: to disprove a statement it's enough to find a single counter-example.) Although we like our counter-examples to be as simple as possible, this one, despite being the simplest for the case at hand, is nevertheless quite complicated!

Example 4.19 (Counter-example to the converse of Lagrange's Theorem)**.**
Let $G = A_4$, the subgroup of S_4 containing all even permutations of the set $\{1, 2, 3, 4\}$ (see Section 4.6.2).

We know that (Proposition 4.14):

$$|A_4| = \frac{4!}{2} = 12.$$

After some thought, we see that A_4 contains eight 3-cycles (recall that an r-cycle is even if and only if r is odd, so every 3-cycle of S_4 belongs to A_4), together with three pairs of disjoint transpositions (which by definition are even permutations), plus of course the identity. This "head count" shows that we have found all the elements of A_4. *(You can* *write them all down if you like; however, our argument doesn't require this amount of detail.)* Lagrange's Theorem implies that, other than the trivial subgroup $\{e\}$ and H itself, the only possible subgroups of A_4 have order 2, 3 or 6. We claim that the last of these is impossible, and prove it by contradiction.

So, suppose $H \subset A_4$ is a subgroup of order 6. Then H must contain at least one 3-cycle (because there are fewer than six even permutations that aren't). We'll denote this 3-cycle by $(a\ \ b\ \ c)$; in other words, a, b, c are labels for three of the four integers $1, 2, 3, 4$, but we're not specifying which is which. Then by closure under inverse, H must also contain the inverse of this 3-cycle, which is $(b\ \ a\ \ c)$.

It's conceivable that H contains just this pair of 3-cycles, plus all (three) pairs of disjoint transpositions, plus the identity (for a total of six). However, notice that if $\{a, b, c, d\} = \{1, 2, 3, 4\}$ then:

$$(a\ \ b)(c\ \ d)(a\ \ b\ \ c) = (b\ \ d\ \ c),$$

which is a *different* 3-cycle, contradicting the closure of H under multiplication.

It follows that H contains another 3-cycle (and its inverse). Since every pair of 3-cycles must have precisely two elements in common, by relabelling if necessary we may assume that this second pair of 3-cycles are $(a\ \ b\ \ d)$ and $(b\ \ a\ \ d)$. The four 3-cycles plus the identity bring the number of elements in H up to five. Now notice that:

$$(a\ \ b\ \ c)(a\ \ b\ \ d) = (a\ \ c)(b\ \ d),$$
$$(b\ \ a\ \ c)(b\ \ a\ \ d) = (b\ \ c)(a\ \ d),$$

which are *different* pairs of disjoint transpositions, again contradicting the closure of H under multiplication.

Having shown that all possibilities for H lead to contradictions, we conclude that our original assumption must be incorrect; so H cannot have order 6. □

4.8.3. Fermat's Little Theorem and Carmichael numbers

Lagrange's Theorem and its group theoretical consequences may seem somewhat abstract, perhaps even a little rarefied. However, as we mentioned at the beginning of Section 4.2, one of the motivations for studying group theory is to assemble a tool kit that can be applied whenever we see a group. A good example of such an application is the following result from classical number theory, known to Fermat as early as 1640 and named after him[12], but which with the benefit of mathematical hindsight can be derived rather easily as a corollary of Corollary 4.17 (hence a corollary of a corollary of Lagrange's Theorem!). The result ("Fermat's Little Theorem") may be regarded as a slightly more sophisticated version of Proposition 3.5.

[12] Fermat never actually supplied a proof, stating: "I would send you a demonstration of it, if I did not fear going on for too long." The first published proof appeared in 1736, in a paper of Euler.

Theorem 4.18 (Fermat's Little Theorem)**.**
Suppose p is a prime number. Then $a^{p-1}-1$ is divisible by p, for all integers a other than multiples of p.

Note. If a is a multiple of p then so is any power of a, so the statement is clearly false in this case. However $a^p - a$ is then divisible by p, and sometimes Fermat's Little Theorem is stated this way in order to achieve a more elegant form. ♢

Proof. This is a direct proof, using the group exponentiation laws.

The theorem is equivalent to the congruence:

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

so it makes sense to work with congruence classes modulo p. Since a isn't a multiple of p we have $\bar{a} \neq \bar{0}$, hence $\bar{a}$ is an element of the multiplicative group $G = \mathbb{Z}_p^*$. By Corollary 4.17 the order n of $\bar{a}$ divides the order of G, which is $p-1$; thus $p-1 = nq$ for some integer q. Hence:

$$(\bar{a})^{p-1} = (\bar{a})^{nq} = ((\bar{a})^n)^q = (\bar{1})^q = \bar{1},$$

which is precisely the required congruence when written as an equation in G. ■

Remark 4.29*.* It's easy to see that Fermat's Little Theorem can go wrong if p isn't prime. For example, if we try $p = 4$ (clearly not prime) then the result fails for "base" integers $a = 2$, $a = 3$ and $a = 4$, and hence for all integers congruent to 0, 2 or 3 modulo 4; in other words, the result holds only for base integers congruent to $1 \pmod{4}$. However, perhaps surprisingly, this is not invariably the case; see Example 4.21 below. ♢

Raising an integer to a power changes only the multiplicities of its prime factors; the factors themselves are unaltered. Fermat's Little Theorem can help us to work out remainders when a (large) power is divided by other primes.

Example 4.20 (Least residues of a large power)**.**
Consider the power 6^{251}. This is a big number! *(Try evaluating it on your calculator, and see what happens!)* Let's call in N. It's divisible by its prime factors 2 and 3; and 4 is also a factor, as are all powers of 2 and 3 up to 251, and their products. By Proposition 3.5 it has remainder 1 when divided by 5. Of course 6 is a factor. When dividing by 7, note that 6^2 leaves remainder 1, so that N has remainder 6. We have seen that 8 and 9 are factors. When dividing by 10 we note that $6^2 \equiv 6 \pmod{10}$, so by the laws of modular arithmetic:

$$6^{251} \equiv 6^{250} \equiv \cdots \equiv 6 \pmod{10};$$

so N has remainder 6 when divided by 10.

Now we reach $p = 11$. By Fermat's Little Theorem, working with congruence classes modulo 11 we have:

$$(\bar{6})^{10} = \bar{1}.$$

Hence writing $251 = (25 \times 10) + 1$ we obtain:

$$(\bar{6})^{251} = \left((\bar{6})^{10}\right)^{25} \times \bar{6} = \bar{1} \times \bar{6} = \bar{6};$$

so dividing N by 11 leaves remainder 6. We can continue this procedure for as long as we please. □

An immediate and interesting question that arises from Fermat's Little Theorem is whether there exist non-prime moduli for which the theorem holds; in other words, what is the logical status of the converse? If p is replaced by a composite natural number n then the theorem will clearly hold for any "base" integer a satisfying $a \equiv 1 \pmod{n}$, or $a \equiv -1 \pmod{n}$ if n is odd, but in general will break down for other values of a, as we observed in Remark **??**. For example, if n is a prime power, say $n = p^m$ for $m \geqslant 2$, then since:

$$n - 1 = p^m - 1 \geqslant m$$

(can you prove this?) we can write $n - 1 = qm + r$ for integers $q \geqslant 1$ and $r \geqslant 0$; so if a is a multiple of p, say $a = kp$, then:

$$a^{n-1} = k^{n-1}(p^m)^q p^r = k^{n-1} p^r n^q \equiv 0 \pmod{n}.$$

This is consistent with both Fermat's Little theorem and its converse. To disprove the converse we need to exhibit a counter-example, which means finding a *composite* integer n with the property that $a^{n-1} \equiv 1 \pmod{n}$ for *all* (sensible) base integers a. Here's one.

Example 4.21 (Counter-example to converse of Fermat's Little Theorem)**.**
Consider $n = 561$. This integer isn't prime (its digit sum is a multiple of 3), and can be factorised as follows:

$$561 = 3 \times 187 = 3 \times 11 \times 17.$$

Now suppose a is any integer other than a multiple of 3, 11 or 17; in other words, a is coprime to n. Then by Fermat's Little Theorem:

$$a^2 \equiv 1 \pmod{3}, \qquad a^{10} \equiv 1 \pmod{11}, \qquad a^{16} \equiv 1 \pmod{17}.$$

Hence:

$$a^{560} = (a^2)^{280} \equiv 1 \pmod 3, \qquad a^{560} = (a^{10})^{56} \equiv 1 \pmod{11}, \qquad a^{560} = (a^{16})^{48} \equiv 1 \pmod{17}.$$

We now apply the Chinese Remainder Theorem (Theorem 3.2), twice, to conclude that:

$$a^{560} \equiv 1 \pmod{561}.$$ □

Example 4.21 was first given by Carmichael[13] [9], and gives rise to the following general definition.

[13] Robert Carmichael (1879–1967): American mathematician, interested in both number theory and group theory.

Definition 4.22 (Carmichael number)**.**
A composite integer n is called a *Carmichael number* if $a^{n-1} - 1$ is divisible by n for all integers a coprime to n. ♦

Every Carmichael number is by definition a counter-example to the converse of Fermat's Little Theorem, and in addition to the integer 561 (which happens to be the smallest Carmichael number) a few more were constructed in [9], such as 1105, 1729 and 2565. More recently, it has been shown that there are in fact infinitely many Carmichael numbers [2]. Interestingly, with $295,486,761,787$ decimal digits, the largest known Carmichael number [1] is (considerably) greater than the largest known prime, which has a mere $24,862,048$ digits (see Remark 2.5).

The practical significance of Carmichael numbers is that they provide a check list of "false positives" for the use of Fermat's Little Theorem in primality testing, which is an important aspect of the black art of encryption in modern cryptography. Because they are rather sparse, the statistical likelihood of mistaking Carmichael numbers for primes is rather low; so in practice Fermat's Little Theorem is a reasonably effective method for deciding whether a natural number is "probably prime".

A. The Fine Detail

A.1. Finite sets

We first revisit Proposition 1.4 from Section 1.2.7 of Chapter 1, which at the time we left unproven. An inductive proof now allows us to make amends. It's interesting to note that the argument requires us to take care of quite a few small details, which when written out in full make the proof seem quite complicated, although the underlying structure is, as always with induction, the same.

Proposition A.1 (Functions between finite sets)**.**
Suppose $f\colon [m] \to [n]$ for natural numbers m, n.

i) *If f is one-to-one then $m \leqslant n$.*

ii) *If f is onto then $m \geqslant n$.*

Proof. For (i) we use induction on m, whereas for (ii) we use induction on n.

(i) Let $P(m)$ be the statement:

> *For all n, if $f\colon [m] \to [n]$ is one-to-one then $m \leqslant n$.*

Then $P(1)$ is true, since $1 \leqslant n$ for all $n \in \mathbb{N}$.

Now assume $P(m)$ is true, and suppose $f\colon [m+1] \to [n]$ is one-to-one. If $c = f(m+1)$ then restricting f to $[m]$ gives us a "new" function $g\colon [m] \to [n] \smallsetminus \{c\}$, which is also one-to-one. Now define a function $h\colon [n] \smallsetminus \{c\} \to [n-1]$ by:

$$h(r) = \begin{cases} r, & \text{if } r < c, \\ r-1, & \text{if } r > c. \end{cases}$$

Then h is one-to-one; for $h(r) = h(s)$ is only possible if $r, s < c$ or $r, s > c$, and in both cases it follows immediately that $r = s$. The composition $h \circ g\colon [m] \to [n-1]$ is one-to-one by Proposition 1.3, and therefore by the induction hypothesis $m \leqslant n-1$. Hence $m+1 \leqslant n$, so $P(m+1)$ is true. The induction step is now complete, so by the Principle of Induction $P(m)$ is true for all m.

(ii) Let $P(n)$ be the statement:

> *For all m, if $f\colon [m] \to [n]$ is onto then $m \geqslant n$.*

Then $P(1)$ is true, since $m \geqslant 1$ for all $m \in \mathbb{N}$.

Now assume $P(n)$ is true, and suppose $f\colon [m] \to [n+1]$ is onto. Let $f(m) = c$. Since $n+1 \geqslant 2$ the set $[n+1] \smallsetminus \{c\}$ is non-empty, and therefore contains an element p, say. We use p to "divert" f away from c, defining a new function $g\colon [m-1] \to [n+1] \smallsetminus \{c\}$ by:

$$g(r) = \begin{cases} f(r), & \text{if } f(r) \neq c, \\ p, & \text{if } f(r) = c. \end{cases}$$

Since f is onto so is g. Having "stepped down" the domain from $[m]$ to $[m-1]$ we now "step down" the codomain as in (i) by defining a function $h\colon [n+1] \smallsetminus \{c\} \to [n]$ as follows:

$$h(r) = \begin{cases} r, & \text{if } r < c, \\ r-1, & \text{if } r > c. \end{cases}$$

Then h is onto; for, if $s < c$ then $s = h(s)$, whereas if $s \geqslant c$ then $s = h(s+1)$. The composition $h \circ g\colon [m-1] \to [n]$ is onto by Proposition 1.3, and therefore by the induction hypothesis $m-1 \geqslant n$. Hence $m \geqslant n+1$, so $P(n+1)$ is true. This completes the induction step, and the result follows from the Priniciple of Induction. ■

We now revisit Proposition 1.5 from Section 1.2.7 of Chapter 1, which was also left unproven. This result seems entirely obvious! However, as remarked at the time, we are using a specific and precise definition of finite sets (Definition 1.8), for which the result serves as a validation. The proof again involves induction, and will probably once again seem unduly complicated. It is important to put all our preconceptions of finiteness on one side, and work only with the formal definition (Definition 1.8); notice how many times it appears!

Proposition A.2 (Subsets of finite sets)**.**
If A is a finite set and $B \subseteq A$ then B is finite and $|B| \leqslant |A|$.

Proof. This comes in two parts: Part 1 is a proof by induction, and Part 2 is a direct proof that makes use of Part 1 and the properties of bijections (Propositions 1.1 and 1.3).

Part 1. We first prove by induction that the following statement $P(n)$ is true for all $n \geqslant 1$:

If $C \subseteq [n]$ then C is finite with $|C| \leqslant n$.

The base case $P(1)$ is true, since the only possible subsets of $[1]$ are $[1]$ and $\emptyset$.

Now assume $P(n)$ is true, and suppose $C \subseteq [n+1]$. Then $C \smallsetminus \{n+1\} \subseteq [n]$, so by the induction hypothesis there exists a bijection $f\colon [m] \to C \smallsetminus \{n+1\}$ for some $m \leqslant n$ (Definition 1.8). We now simply define $f_C\colon [m+1] \to C$ by:

$$f_C(r) = \begin{cases} f(r), & \text{if } r \leqslant m, \\ n+1, & \text{if } r = m+1. \end{cases}$$

Then f_C is a bijection. It's onto because:

$$\begin{aligned}\operatorname{im}(f_C) &= \operatorname{im}(f) \cup \{n+1\} = (C \smallsetminus \{n+1\}) \cup \{n+1\}, \quad \text{since } f \text{ is onto}\\ &= C,\end{aligned}$$

and it's one-to-one because if $r \neq s$ then:

- if $r, s \leqslant m$ then $f_C(r) \neq f_C(s)$, since f is one-to-one;
- if $r = m+1$ then $s \leqslant m$, so $f_C(r) = n+1$ and $f_C(s) = f(s) \in C \smallsetminus \{n+1\}$, hence $f_C(r) \neq f_C(s)$.

(Note that we've used the direct definition of one-to-one, rather than the "standard routine"; see Definition 1.3.) Therefore, applying Definition 1.8, C is finite and:

$$|C| = m+1 \leqslant n+1,$$

since $m \leqslant n$. Hence $P(n+1)$ is true. This completes the induction step, and the result follows from the Principle of Induction.

Part 2. Now suppose $|A| = n$, so there exists a bijection $f_A\colon [n] \to A$ (Definition 1.8). Then the inverse function $f_A^{-1}\colon A \to [n]$ is also a bijection (Proposition 1.1). Let $C = f_A^{-1}(B)$, the image of B under f_A^{-1}, or equivalently the preimage of B under f_A (see Sections 1.2.2 and/or 1.2.6). Then $C \subseteq [n]$, so it follows from Part 1 that C is finite with $|C| \leqslant n$. Hence there exists a bijection $f_C\colon [m] \to C$ for some $m \leqslant n$ (Definition 1.8).

We now note that:

$$f_A(C) = f_A(f_A^{-1}(B)) = B.$$

Therefore by restricting f_A to act only on elements of C we obtain a bijection $g\colon C \to B$. It follows that the composition $f_B = g \circ f_C\colon [m] \to B$ is a bijection (Proposition 1.3), and referring to Definition 1.8 once more shows that the proof is finally complete. ■

A.2. The Euclidean algorithm and Bézout's identity

Also known as the *division algorithm,* the *Euclidean algorithm* allows us to methodically compute the greatest common divisor (aka. highest common factor) d of any pair of integers a, b. Having done this, we can "extend" the algorithm by backtracking through it to obtain an expression for d as a "linear combination" of a and b, known as *Bézout's identity:*

$$d = \alpha a + \beta b,$$

for integers α and β. This identity has a variety of uses; for example, it was the principal ingredient in our proof of Proposition 4.2.

We begin with an example, which shows how the algorithm and its extension work in practice, and indicates how to construct a general proof.

Example A.1 (Greatest common divisor)**.**
Suppose $a = 434$ and $b = 133$. We first divide a by b, expressing the result in terms of "quotient" and "remainder":

$$434 = 3.133 + 35. \tag{$\spadesuit$}$$

By rearranging this equation we see that d also divides the remainder 35, and is therefore a common divisor of 133 and 35. Moreover, from the equation as written it follows that any common divisor of 133 and 35 also divides 434, and therefore, being a common divisor of 133 and 434, can be no greater than d. Hence d is the greatest common divisor of 133 and 35. This is the essence of the algorithm: simplify the problem by successively expressing d as the greatest common divisor of two smaller numbers.

To continue the algorithm we now divide b by 35:

$$133 = 3.35 + 28, \tag{$\heartsuit$}$$

from which it follows that d is the greatest common divisor of 35 and 28. Once again, dividing 35 by 28 we obtain:

$$35 = 28 + 7, \tag{$\diamondsuit$}$$

and deduce that d is the greatest common divisor of 28 and 7. One final division yields:

$$28 = 4.7, \tag{$\clubsuit$}$$

which tells us (if we weren't already aware of the fact) that 7 is a factor of 28, so their greatest common divisor is 7. We conclude that $d = 7$. Since the remainder is now 0, the algorithm also terminates here.

Now, to obtain Bézout's identity we rearrange the penultimate equation $(\diamondsuit)$ and successively substitute the preceding equations into it, to eventually arrive at an expression for 7 in terms of a and b. So, we first write:

$$7 = 35 - 28,$$

and then use $(\heartsuit)$ to eliminate 28:

$$\begin{aligned} 7 &= 35 - 133 + 3.35 \\ &= 4.35 - 133. \end{aligned}$$

Finally we use $(\spadesuit)$ to eliminate 35:

$$\begin{aligned} 7 &= 4.434 - 12.133 - 133 \\ &= 4.434 - 13.133, \end{aligned}$$

which is Bézout's identity for the case at hand, with $\alpha = 4$ and $\beta = -13$. Amongst other things, it confirms the fact that any common divisor of a and b must also divide 7, hence $d \leqslant 7$; however it doesn't contain enough information for us to deduce that 7 is a common factor of a and b, and therefore that $d = 7$. □

It is quite straightforward to turn Example A.1 into a general argument. Suppose we have an arbitrary pair of integers $a, b \in \mathbb{N}$ with $a > b$. To determine the greatest common divisor $d = \gcd(a, b)$ we first divide a by b to obtain the equation:

$$a = q_1 b + r_1, \quad \text{where } 0 \leqslant r_1 < b.$$

If $r_1 = 0$ then a is an exact multiple of b, so $d = b$. If $r_1 \neq 0$ then we deduce, as in the example, that $d = \gcd(b, r_1)$, and continue the algorithm by dividing b by r_1 to get:

$$b = q_2 r_1 + r_2, \quad \text{where } 0 \leqslant r_2 < r_1.$$

If $r_2 = 0$ then $d = r_1$. Otherwise $d = \gcd(r_1, r_2)$ and the algorithm continues. After n iterations we reach the equation:

$$r_{n-2} = q_n r_{n-1} + r_n, \quad \text{where } 0 \leqslant r_n < r_{n-1}. \tag{$*$}$$

We also know from the previous iteration that $d = \gcd(r_{n-2}, r_{n-1})$. Therefore if $r_n = 0$ then $d = r_{n-1}$ and the algorithm terminates, whereas if $r_n \neq 0$ we deduce that $d = \gcd(r_{n-1}, r_n)$ and the algorithm continues.

The Euclidean algorithm cannot continue indefinitely, because the sequence:

$$a > b > r_1 > r_2 > \cdots > r_n$$

can accomodate at most b non-negative integers r_i. There must therefore exist some natural number $N \leqslant b$ such that $r_N = 0$, at which point the algorithm terminates. The final equation is then:

$$r_{N-2} = q_N r_{N-1}. \tag{$**$}$$

From this, together with $d = \gcd(r_{N-2}, r_{N-1})$ from the penultimate equation, we conclude that $d = r_{N-1}$.

We have proved the first part of the following theorem.

Theorem A.3 (Euclidean algorithm)**.**
The greatest common divisor d of $a, b \in \mathbb{N}$ is the least positive remainder when the Euclidean algorithm is applied to a, b. Furthermore, there exist integers $\alpha, \beta \in \mathbb{Z}$ such that $d = \alpha a + \beta b$.

Note. The equation $d = \alpha a + \beta b$ is *Bézout's identity.* ◊

Proof. It remains to derive Bézout's identity, for which we will use proof by induction. At first sight this might seem strange, because the identity does not depend on a natural number n. However, each stage of the Euclidean algorithm produces a new (and simpler) expression for d, for which there should be a corresponding Bézout identity, and it is in proving this that induction comes into play.

Since our induction will begin at the end of the Euclidean algorithm we first relabel the remainders in ascending order using $d_1, \dots, d_n$, thus:

$$d_1 = d = r_{N-1}, \quad d_2 = r_{N-2}, \ \dots \ , d_{N-1} = r_1, \quad d_N = b, \quad d_{N+1} = a,$$
$$d_n = d, \quad \text{for all } n > N.$$

With this notation equations $(**)$ and $(*)$ of the Euclidean algorithm become:

$$d_2 = q_N d_1, \tag{‡}$$
$$d_{n+1} = q_{N-n+1} d_n + d_{n-1}, \tag{†}$$

for all $n = 2, \dots, N$. *(Try substituting various values of n, such as 2 or N, to convince yourself that the indexing q_{N-n+1} is correct.)*

Now, for all $n \in \mathbb{N}$ let $Q(n)$ be the statement:

> *There exist $\alpha_n, \beta_n \in \mathbb{Z}$ such that:* $d = \alpha_n d_{n+2} + \beta_n d_{n+1}$.

Rearranging (†) with $n = 2$ (the relabelled penultimate equation of the Euclidean algorithm) yields:

$$d = d_3 - q_{N-1} d_2.$$

So $Q(1)$ is true, with $\alpha_1 = 1$ and $\beta_1 = -q_{N-1}$. This proves the base case.

For the induction step, let $1 \leqslant n \leqslant N - 2$ and suppose $Q(n)$ is true. We rearrange equation (†) for d_{n+3} as follows:

$$d_{n+1} = d_{n+3} - q_{N-n+3} d_{n+2}.$$

Then by the induction hypothesis:

$$\begin{aligned} d &= \alpha_n d_{n+2} + \beta_n d_{n+1} \\ &= \beta_n d_{n+3} + (\alpha_n - q_{N-n+3}\beta_n) d_{n+2}. \end{aligned}$$

Therefore $Q(n + 1)$ is true, with $\alpha_{n+1} = \beta_n$ and $\beta_{n+1} = \alpha_n - q_{N-n+3}\beta_n$. Furthermore $Q(n)$ is true for all $n \geqslant N$, by simply choosing $\alpha_n = 1$ and $\beta_n = 0$. This completes the induction step.

The Principle of Induction now ensures that $Q(n)$ is true for all $n \in \mathbb{N}$. In particular, $Q(N - 1)$ is true, and this is Bézout's identity for d in terms of a and b, as stated. ■

Remark A.1. The integers α, β in Bézout's identity aren't unique; in fact, far from it. For example, if $a = 3$ and $b = 2$ then $\gcd(a, b) = 1$ and we can write:

$$1 = 3 - 2 = 3.3 - 4.2 = 5.3 - 7.2 = \cdots \qquad \lozenge$$

A.3. The Fundamental Theorem of Arithmetic

The Fundamental Theorem of Arithmetic, or FTA, is a very well-known theorem, which as its name implies lies at the heart of classical number theory.

Theorem A.4 (Fundamental Theorem of Arithmetic)**.**
Every natural number $2, 3, 4, \ldots$ is either prime, or a unique product of primes.

In Proposition 2.9 we proved (by induction) the existence of a prime factorisation; so it remains to prove that the factorisation is unique. For this we use an elementary property of prime numbers, dating back to Euclid (Elements [15], Book VII, Proposition 30).

Lemma A.5 (Euclid's Lemma)**.**
Suppose p is a prime number, and m, n are natural numbers. If p divides mn then p divides m or p divides n.

Notes.

1) It's essential that p is prime. For example, if $m = 2$ and $n = 9$ then 6 divides $mn = 18$, but 6 doesn't divide m or n.
2) When viewed with the benefit of hindsight, with the FTA in hand, Euclid's lemma is obvious; it says that any prime factor of a product of integers must be a prime factor of one integer or the other. However, we can't use this as a proof, since we're proposing to use Euclid's lemma to assist in proving the FTA.

Proof. This is a direct proof, which uses the Euclidean algorithm.

We have $mn = pc$ for some natural number c. Suppose p doesn't divide m. Since p is prime this means that the greatest common divisor of p and m is 1. Hence by Bézout's identity:

$$1 = \alpha p + \beta m$$

for integers α, β. Multiplying this equation by n:

$$n = \alpha pn + \beta mn = \alpha pn + \beta pc = (\alpha n + \beta c)p,$$

which shows that p divides n, as required. ■

Remark A.2. This is the "standard proof" of Euclid's lemma. However, the use of Bézout's identity, and the fact that Bézout lived 2000 years after Euclid, is a strong hint that it's not the original one! ♢

Corollary A.6 (Generalised Euclid's Lemma)**.**
Suppose p is prime and $n_1, \dots, n_r$ are natural numbers. If p divides $n_1 \cdots n_r$ then p divides n_i for some $i = 1, \dots, r$.

Proof. This is an easy proof by induction, using Euclid's lemma for the induction step. We leave it as an exercise. ■

Proof. [FTA]
We'll prove uniqueness by induction. Let $P(n)$ be the statement:

> *Any product of n primes has a unique prime factorisation.*

Then $P(1)$ is true, by the definition of prime numbers (Definition 2.1).

Assume $P(n)$ is true, for $n \geqslant 1$. Let $p_1, \dots, p_{n+1}$ be prime, and suppose:

$$p_1 \cdots p_{n+1} = q_1 \cdots q_s,$$

for primes $q_1, \dots, q_s$. Since $n+1 \geqslant 2$ the left hand side of this equation is not a prime number, so $s \geqslant 2$. It follows from Corollary A.6 that q_s divides one of the p_i, and therefore must equal one of them, since the p_i are prime. By relabelling the p_i if necessary we may assume that $q_s = p_{n+1}$. Then by the multiplicative cancellation law (Proposition 2.1):

$$p_1 \dots p_n = q_1 \dots q_{s-1}.$$

It therefore follows from the induction hypothesis that the primes $p_1, \dots, p_n$ may be relabelled, if necessary, so that $q_1 = p_1, \dots, q_{s-1} = p_n$. Thus $P(n+1)$ is true. This completes the induction step, and the result follows from the Principle of Induction. ■

The Fundamental Theorem of Arithmetic has numerous applications. We give just one, that we've already mentioned and made use of in our proof of the Chinese Remainder Theorem back in Section 3.3.

Proposition A.7 (The GCD-LCM formula)**.**
For all $m, n \in \mathbb{N}$ we have:

$$\gcd(m,n)\operatorname{lcm}(m,n) = mn.$$

Proof. This is a direct proof, which comes in two parts.

We abbreviate $\gcd(m,n) = d$ and $\operatorname{lcm}(m,n) = M$. Then $M \leqslant mn$. Suppose the prime factorisations of m and n are:

$$m = p_1 \cdots p_r, \qquad n = q_1 \cdots q_s.$$

Case 1. If the prime factors of m are entirely different to those of n then m and n have no common factors, so $d = 1$. Furthermore, since every prime factor of m and n must appear in the prime factorisation of M, we have:

$$M \geqslant p_1 \cdots p_r q_1 \cdots q_s = mn.$$

Therefore $dM = M = mn$, so the formula holds in this case.

Case 2. If m and n share some prime factors then by relabelling if necessary there is an index k such that $p_1 = q_1, \ldots, p_k = q_k$ and all the remaining primes are different. Then $d = p_1 \cdots p_k$. Hence:

$$m = da, \qquad n = db,$$

where a (resp. b) is the product of the remaining prime factors of m (resp. n). Again, since every prime factor of m and n must appear in the prime factorisation of M, we have:

$$M \geqslant dab = mb = an.$$

Therefore, since M is the *least* common multiple of m and n, we have $M = abd$. It follows that:

$$dM = d^2ab = mn,$$

as required. ■

Example A.2 (Greatest common divisor and least common multiple)**.**
Suppose $m = 156$ and $n = 225$. We first determine their greatest common divisor d using the Euclidean algorithm (Section A.2):

$$\begin{aligned} 225 &= 156 + 69, \\ 156 &= 2.69 + 18, \\ 69 &= 3.18 + 15, \\ 18 &= 15 + 3 \\ 15 &= 5.3. \end{aligned}$$

Therefore $d = 3$, hence by Proposition A.7:

$$\begin{aligned} \operatorname{lcm}(m,n) = \frac{mn}{d} = \frac{m}{d}n &= 52.225 = 50.225 + 2.225 \\ &= 11250 + 450 = 11700. \end{aligned}$$

Notice that we divided through by d before taking the product, which allowed us to (just about) do the arithmetic without using a calculator!

We could have attempted to calculate $\operatorname{lcm}(m,n)$ by factorising m, n into primes and "sieving out" duplicates of those that appear in both m and n, as in the proof of Proposition A.7. That would be a reasonable strategy in this example, since m, n are quite small. However, in general prime factorisation is far more difficult than applying the Euclidean algorithm. We're in the interesting situation where the method of a proof turns out to be not so great in practice! □

A.4. Associativity of matrix multiplication

To show that matrix multiplication is associative we need first and foremost an efficient notation for the entries in a product of matrices. This involves the use of “double subscripts”. If A is a $m \times n$ matrix then we denote by a_{ij} its ij-th entry; in other words, the element (usually, but not always, a real number) at the intersection of its i-th row and j-th column. We then write:

$$A = (a_{ij}),$$

where $1 \leqslant i \leqslant m$ and $1 \leqslant j \leqslant n$. This saves us from having to write out something monstrous like:

$$A = \begin{pmatrix} a_{11} & a_{12} & \dots & a_{1n} \\ a_{21} & a_{22} & \dots & a_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ a_{m1} & a_{m2} & \dots & a_{mn} \end{pmatrix},$$

although it is perhaps helpful to bear in mind that this is what we are really dealing with.

If B is a $n \times p$ matrix:

$$B = (b_{jk}),$$

where $1 \leqslant j \leqslant n$ and $1 \leqslant k \leqslant p$, the ik-th element of AB is the following sum:

$$a_{i1}b_{1k} + a_{i2}b_{2k} + \cdots a_{in}b_{nk},$$

which is obtained by multiplying corresponding elements of the i-th row of A and the k-th column of B, then adding them together. It is more convenient to write this using a summation symbol:

$$\sum_{j=1}^{n} a_{ij}b_{jk}.$$

It therefore follows that $AB = P = (p_{ik})$ where:

$$p_{ik} = \sum_{j=1}^{n} a_{ij}b_{jk}. \tag{A.1}$$

Now suppose that C is a $p \times q$ matrix:

$$C = (c_{kl}),$$

where $1 \leqslant k \leqslant p$ and $1 \leqslant l \leqslant q$. We want to compare the matrix products $(AB)C$ and $A(BC)$. If $AB = P = (p_{ik})$ and $BC = Q = (q_{jl})$ then by (A.1):

$$p_{ik} = \sum_{j=1}^{n} a_{ij}b_{jk}, \qquad q_{jl} = \sum_{k=1}^{p} b_{jk}c_{kl}. \tag{A.2}$$

Therefore, if $PC = X = (x_{il})$ and $AQ = Y = (y_{il})$ then:

$$x_{il} = \sum_{k=1}^{p} p_{ik}c_{kl}, \qquad y_{il} = \sum_{j=1}^{n} a_{ij}q_{jl}, \tag{A.3}$$

and substituting (A.2) into (A.3) yields:

$$x_{il} = \sum_{k=1}^{p}\Big(\sum_{j=1}^{n} a_{ij}b_{jk}\Big)c_{kl}, \qquad y_{il} = \sum_{j=1}^{n} a_{ij}\Big(\sum_{k=1}^{p} b_{jk}c_{kl}\Big).$$

We then apply the distributive law and associative law of multiplication in $\mathbb{R}$ (or possibly $\mathbb{C}$, if we're working with complex matrices) to expand the brackets:

$$x_{il} = \sum_{k=1}^{p}\sum_{j=1}^{n} a_{ij}b_{jk}c_{kl}, \qquad y_{il} = \sum_{j=1}^{n}\sum_{k=1}^{p} a_{ij}b_{jk}c_{kl},$$

which are sums of n^2 identical terms. So, the matrices X, Y have exactly the same entries, which means that $X = Y$, as required.

A.5. Massive multiplication of brackets

The "sigma summation" $\sum$ and "pi product" $\prod$ notations were mentioned right back in Preliminaries and Notation, since when we have made frequent use of the former, but none whatsoever of the latter. We will redress that, by developing a general formula for the familiar routine of "multiplying out brackets", where to make things interesting an arbitrarily large number of brackets are permitted, and each bracket can be an arbitrarily long summation! This formula will also address the basic but important question of when and how the two symbols $\sum$ and $\prod$ can be interchanged, and, perhaps surprisingly, gives a hint how the abstract algebraic ideas of group theory from Chapter 4 can be generalised even further.

We begin with a very simple and familiar example, mutliplying out the following pair of brackets:

$$(c_{11} + c_{12})(c_{21} + c_{22}),$$

where $c_{11}, \ldots, c_{22}$ are arbitrary real numbers. We're using "double subscript" notation, because that allows us to now introduce $\sum$ and $\prod$ as follows:

$$(c_{11} + c_{12})(c_{21} + c_{22}) = \prod_{i=1}^{2}\sum_{j=1}^{2} c_{ij}.$$

Now, expanding the brackets in the usual way gives:

$$\prod_{i=1}^{2}\sum_{j=1}^{2} c_{ij} = c_{11}c_{21} + c_{11}c_{22} + c_{12}c_{21} + c_{12}c_{22}, \tag{A.4}$$

On the other hand, if we take products first and then sum we obtain the far simpler expression:

$$\sum_{j=1}^{2}\prod_{i=1}^{2} c_{ij} = c_{11}c_{21} + c_{12}c_{22}. \tag{A.5}$$

We see that although all the terms in (A.5) are present in (A.4), there are some absentees. It's clear that interchanging the order of $\sum$ and $\prod$ is not permissible!

To fix this, notice that all the terms in (A.4) are products of the form:

$$c_{1f(1)}c_{2f(2)},$$

where $f\colon \{1,2\} \to \{1,2\}$; in other words, a function from the 2-element set $\{1,2\}$ to itself. (Notation-wise, although this may look sligthly strange at first, we've simply replaced the second subsrcipt by $f(1)$ or $f(2)$.) In total there are $2^2 = 4$ such functions (because each integer $1,2$ can be mapped to either 1 or 2 without restriction), accounting for the four terms in (A.4). So if we define the following set:

$$F_2 = \big\{\text{functions } f\colon \{1,2\} \to \{1,2\}\big\},$$

then equation (A.4) can be expressed as follows:

$$\prod_{i=1}^{2}\sum_{j=1}^{2} c_{ij} = \sum_{f\in F_2}\prod_{i=1}^{2} c_{if(i)}.$$

This takes care of the simplest case, and suggests how to resolve the general case of a product of n brackets each of which contains n summands. We define:

$$F_n = \big\{\text{functions } f\colon \{1,\dots,n\} \to \{1,\dots,n\}\big\}.$$

Note that we are not asking for f to be a bijection, as we did in Definitions 4.6 and 4.7. The bijections in F_n are of course the permutations of $\{1,\dots,n\}$; so F_n contains the symmetric group S_n. However, F_n itself is *not* a group, because not all its elements have inverses. Nevertheless, we still have the associative law (which holds for compositions of functions in general; see Remark 1.17 (2)), and an identity element (the identity function); this means that F_n is an example of a more general type of algebraic structure called a *monoid,* and more generally still a *semigroup.* (We won't say any more about these areas of abstract algebra, except that they can be developed axiomatically in much the same way as group theory, and semigroup theory in particular has become an important area of modern algebra, with applications to theoretical computer science amongst others.) In total there are n^n elements of F_n, of which $n!$ belong to S_n.

Proposition A.8 (Massive multiplication of brackets)**.**
Suppose c_{ij} is a collection of real numbers, with $1 \leqslant i,j \leqslant n$. Then:

$$\prod_{i=1}^{n}\sum_{j=1}^{n} c_{ij} = \sum_{f\in F_n}\prod_{i=1}^{n} c_{if(i)}.$$

Note. Multiplying out the n brackets on the left hand side yields a sum with n^n terms, the same number as the n^n functions on the right hand side. ◊

Proof. A typical term in the product of sums on the left hand side of the equation is obtained by selecting an element of c_{1j_1} of the first sum $c_{11}+\cdots+c_{1n}$, where $j_1 \in \{1,\ldots,n\}$, multiplying it by an element c_{2j_2} of the second sum $c_{21}+\cdots+c_{2n}$, and so on, completing the product by picking an element c_{nj_n} from the final sum $c_{n1}+\cdots+c_{nn}$. Define a function $f\colon \{1,\ldots,n\} \to \{1,\ldots,n\}$ by:

$$f(1) = j_1,\ f(2) = j_2, \ldots,\ f(n) = j_n.$$

Then our term is the following product:

$$c_{1\,f(1)} \cdots c_{n\,f(n)}.$$

Moreover, every function $f \in F_n$ contributes such a term. Now notice that the right hand side of the equation is precisely the sum of all such products. ■

Our main example of the use of Proposition A.8 will be its application to the computation of matrix determinants, in Appendix A.6.

A.6. Determinants

In Section 4.3.5 we showed that the set of all $n \times n$ matrices with non-zero determinant is a group under matrix multiplication, the "general linear group". For this we needed the "multiplicative property" of determinants:

$$\det(AB) = \det(A)\det(B).$$

There are many other properties of determinants, but this one is arguably the most important and tricky to establish, so it is the one that we are going to concentrate on here! First, we need a usable definition of the determinant of a matrix.

The determinant of a 2×2 matrix A is the following familiar expression:

$$\det(A) = \det\begin{pmatrix} a & b \\ c & d \end{pmatrix} = \begin{vmatrix} a & b \\ c & d \end{vmatrix} = ad - bc.$$

When the entries of A are written in "double subscript" notation this becomes:

$$\det(A) = \det\begin{pmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{pmatrix} = \begin{vmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{vmatrix} = a_{11}a_{22} - a_{12}a_{21},$$

which looks more complicated, but is in fact much more useful.

Now suppose A is a 3×3 matrix. Let's evaluate its determinant by expansion along the top row:

$$\det(A) = \det\begin{pmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{pmatrix}$$

$$\begin{aligned}
&= a_{11}\begin{vmatrix} a_{22} & a_{23} \\ a_{32} & a_{33} \end{vmatrix} - a_{12}\begin{vmatrix} a_{21} & a_{23} \\ a_{31} & a_{33} \end{vmatrix} + a_{13}\begin{vmatrix} a_{21} & a_{22} \\ a_{31} & a_{32} \end{vmatrix} \\
&= a_{11}(a_{22}a_{33} - a_{23}a_{32}) - a_{12}(a_{21}a_{33} - a_{23}a_{31}) + a_{13}(a_{21}a_{32} - a_{22}a_{31}) \\
&= a_{11}a_{22}a_{33} - a_{11}a_{23}a_{32} - a_{12}a_{21}a_{33} + a_{12}a_{23}a_{31} + a_{13}a_{21}a_{32} - a_{13}a_{22}a_{31}.
\end{aligned}$$

Looking back at the 2×2 determinant, notice that each of its terms is of the form:

$$a_{1\sigma(1)}a_{2\sigma(2)}$$

where $\sigma \in S_2$; ie. a permutation of the set $\{1, 2\}$. (Notation-wise, we've simply replaced the second subsrcipt by $\sigma(1)$ or $\sigma(2)$.) Now S_2 is a group with $2! = 2$ elements, which tallies with the number of terms in the determinant. Notice further that the sign of each term is "$+$" or "$-$" according as σ is even or odd. Thus, taking their sign into account the terms of $\det(A)$ have the form:

$$\operatorname{sgn}(\sigma)a_{1\sigma(1)}a_{2\sigma(2)},$$

where $\operatorname{sgn}(\sigma)$ is the "signature" of σ (Definition 4.14).

Moving up to the 3×3 determinant, we note that, again taking into account their sign, the terms in the expansion of $\det(A)$ have the form:

$$\operatorname{sgn}(\sigma)a_{1\sigma(1)}a_{2\sigma(2)}a_{3\sigma(3)},$$

where σ is a permutation of $\{1, 2, 3\}$; ie. $\sigma \in S_3$. Notice that there are $n! = 6$ elements of S_3, and 6 terms in the expansion of the determinant.

This suggests how we should define the determinant of an arbitrary (square) matrix.

Definition A.1 (Determinant of a matrix)**.**
Suppose $A = (a_{ij})$ is an $n \times n$ matrix. It's *determinant* is defined:

$$\det(A) = \sum_{\sigma \in S_n} \operatorname{sgn}(\sigma)a_{1\sigma(1)} \cdots a_{n\sigma(n)}.$$

There are $n!$ terms in this sum, which is a finite (albeit very large) number. So the order of summation is immaterial. ♢

The essence of Definition A.1 is that the determinant is the alternating sum of products of entries from each row and each column of the matrix. ("Alternating" simply means that the terms in the sum alternately change sign.) We can use the "pi product" notation to write this even more succinctly as follows:

$$\det(A) = \sum_{\sigma \in S_n} \operatorname{sgn}(\sigma) \prod_{i=1}^{n} a_{i\sigma(i)}. \tag{A.6}$$

With this definition tucked away we will now prove the multiplicative property of determinants.

Proposition A.9 (Multiplicative property of determinants)**.**
For all $n \times n$ matrices A, B we have:

$$\det(AB) = \det(A)\det(B).$$

Proof. This is a direct proof, essentially a calculation, that might be desribed as "traditional algebra on steroids". Because the calculation is quite long and rather delicate we have broken it up into three pieces.

Stage 1. If $AB = P = (p_{ij})$ then by (A.6):

$$\begin{aligned}\det(AB) &= \sum_{\rho \in S_n} \operatorname{sgn}(\rho) \prod_{i=1}^{n} p_{i\rho(i)} \\ &= \sum_{\rho \in S_n} \operatorname{sgn}(\rho) \prod_{i=i}^{n} \sum_{k=1}^{n} a_{ik} b_{k\rho(i)},\end{aligned}$$

where we have used equation (A.1) for the entries of a matrix product

$$\begin{aligned}&= \sum_{\rho \in S_n} \operatorname{sgn}(\rho) \sum_{f \in F_n} \prod_{i=1}^{n} a_{if(i)} b_{f(i)\rho(i)}, \quad \text{by Proposition A.8} \\ &= \sum_{f \in F_n} \sum_{\rho \in S_n} \operatorname{sgn}(\rho) \prod_{i=1}^{n} a_{if(i)} b_{f(i)\rho(i)},\end{aligned}$$

since in a finite sum the order of summation is immaterial.

Stage 2. This is the clever bit. We claim that if f *isn't* a permutation then:

$$\sum_{\rho \in S_n} \operatorname{sgn}(\rho) \prod_{i=1}^{n} a_{if(i)} b_{f(i)\rho(i)} = 0.$$

To see this, note that if f isn't a bijection then, since f maps a finite set to itself, f isn't one-to-one; thus $f(r) = f(s)$ for some $r \neq s$. We now pair a permutation ρ with the permutation $\rho' = \rho \cdot (r\ s)$, where $(r\ s)$ is the transposition swapping r and s. These have opposite parity, so $\operatorname{sgn}(\rho') = -\operatorname{sgn}(\rho)$. Hence, assuming for the sake of argument that $r < s$:

$$\begin{aligned}\operatorname{sgn}(\rho)\Big(\operatorname{sgn}(\rho) \prod_{i=1}^{n} a_{if(i)} b_{f(i)\rho(i)} &+ \operatorname{sgn}(\rho') \prod_{i=1}^{n} a_{if(i)} b_{f(i)\rho'(i)}\Big) \\ &= \prod_{i=1}^{n} a_{if(i)} b_{f(i)\rho(i)} - \prod_{i=1}^{n} a_{if(i)} b_{f(i)\rho'(i)} \\ &= \prod_{i=1}^{n} a_{if(i)} b_{f(i)\rho(i)} - \prod_{i=1}^{n} a_{if(i)} b_{f(i)\rho((r\ s)i)} \\ &= \Big(\prod_{i=1}^{n} a_{if(i)}\Big)\Big(\prod_{i=1}^{n} b_{f(i)\rho(i)} - \prod_{i=1}^{n} b_{f(i)\rho((r\ s)i)}\Big)\end{aligned}$$

$$\begin{aligned}&= \Big(\prod_{i=1}^{n} a_{if(i)}\Big)\Big(b_{f(1)\rho(1)}\cdots b_{f(r)\rho(r)}\cdots b_{f(s)\rho(s)}\cdots b_{f(n)\rho(n)}\\ &\qquad\qquad - b_{f(1)\rho(1)}\cdots b_{f(r)\rho(s)}\cdots b_{f(s)\rho(r)}\cdots b_{f(n)\rho(n)}\Big)\\ &= 0,\end{aligned}$$

since $f(r) = f(s)$. So, when we take the sum over all permutations ρ the terms cancel in pairs, and the entire sum therefore collapses to 0.

Stage 3. It follows from Stage 2 that we only need to consider permutations $\sigma \in S_n$ rather than arbitrary functions $f \in F_n$, so our expression simplifies to:

$$\det(AB) = \sum_{\sigma\in S_n}\sum_{\rho\in S_n} \operatorname{sgn}(\rho)\prod_{i=1}^{n} a_{i\sigma(i)}b_{\sigma(i)\rho(i)}.$$

Since S_n is a group we can factorise $\rho = \tau\sigma$, where τ is the permutation $\rho\sigma^{-1}$. Hence:

$$\begin{aligned}\det(AB) &= \sum_{\sigma\in S_n}\sum_{\tau\in S_n} \operatorname{sgn}(\tau\sigma)\prod_{i=1}^{n} a_{i\sigma(i)}b_{\sigma(i)\tau(\sigma(i))}\\ &= \sum_{\sigma\in S_n}\sum_{\tau\in S_n} \operatorname{sgn}(\tau\sigma)\Big(\prod_{i=1}^{n} a_{i\sigma(i)}\Big)\Big(\prod_{i=1}^{n} b_{\sigma(i)\tau(\sigma(i))}\Big),\end{aligned}$$

and then relabelling by $\sigma(i) = j$ in the second product:

$$\begin{aligned}&= \sum_{\sigma\in S_n}\sum_{\tau\in S_n} \operatorname{sgn}(\tau\sigma)\Big(\prod_{i=1}^{n} a_{i\sigma(i)}\Big)\Big(\prod_{j=1}^{n} b_{j\tau(j)}\Big)\\ &= \sum_{\sigma\in S_n}\sum_{\tau\in S_n} \operatorname{sgn}(\tau)\operatorname{sgn}(\sigma)\Big(\prod_{i=1}^{n} a_{i\sigma(i)}\Big)\Big(\prod_{j=1}^{n} b_{j\tau(j)}\Big), \qquad \text{by Proposition 4.9}\\ &= \Big(\sum_{\sigma\in S_n}\operatorname{sgn}(\sigma)\prod_{i=1}^{n} a_{i\sigma(i)}\Big)\Big(\sum_{\tau\in S_n}\operatorname{sgn}(\tau)\prod_{j=1}^{n} b_{j\tau(j)}\Big)\\ &= \det(A)\det(B),\end{aligned}$$

by equation (A.6). ■

Example A.3 (Non-singular matrices have non-zero determinant)**.**
Suppose A is an invertible (aka. *non-singular*) $n \times n$ matrix; in other words, there exists an $n \times n$ matrix B such that:

$$AB = \mathbb{I}_n = BA,$$

where $\mathbb{I}_n$ is the $n \times n$ identity matrix, defined in (4.1). It is an immediate consequence of Definition A.1 that $\det(\mathbb{I}_n) = 1$. Therefore by Proposition A.9:

$$\det(A)\det(B) = 1,$$

which implies that $\det(A) \neq 0$. The converse (if $\det(A) \neq 0$ then A is non-singular) is also true, but requires additional matrix technology. □

Index of Notation

Index of Definitions

Index of Results

Index of Examples

Index of Selected Remarks

Bibliography

[1] W. R. Alford, Jon Grantham, Steven Hayman and Andrew Shallue, *Constructing Carmichael numbers through improved subset-product algorithms,* Mathematics of Computation **83** (2014), 899–915.

[2] W. R. Alford, Andrew Granville and Carl Pomerance, *There are infinitely many Carmichael numbers,* Annals of Mathematics **139** (1994), 703–722.

[3] R. B. J. T. Allenby, *Numbers and Proofs,* Arnold, 1997.

[4] Tom Apostol, *Calculus, Volume 1,* Wiley International Edition, 1967 (2nd ed).

[5] Kenneth Appel and Wolfgang Haken, *Every Planar Map is Four Colorable,* Contemporary Mathematics **98** (1989), American Mathematical Society.

[6] Bill Bruce, *A really trivial proof of the Lucas-Lehmer test,* American Math. Monthly **100** (1993), 370–371.

[7] Georg Cantor, *Über eine Eigenschaft des Inbegriffes aller reellen algebraischen Zahlen,* Journal für die Reine und Angewandte Mathematik **77** (1874), 258–262.

[8] Norman Biggs, *Discrete Mathematics,* Oxford University Press, 2003 (2nd ed).

[9] Robert Carmichael, *On composite numbers P which satisfy the Fermat congruence $a^{P-1} \equiv 1 \pmod{P}$,* American Mathematical Monthly **19** (1912), 22–27.

[10] Paul Cohen, *The independence of the continuum hypothesis, I and II,* Proceedings of the National Academy of Sciences of the United States of Amereica **50** (1963), 1143–48, and **51** (1964), 105–110.

[11] Daniel Cunningham, *Set Theory: A First Course,* Cambridge University Press, 2016.

[12] Keith Devlin, *The Joy of Sets,* Undergraduate Texts in Mathematics, Springer Verlag, 1993 (2nd ed).

[13] Keith Devlin, *Sets, Functions, and Logic: an Introduction to Abstract Mathematics,* Chapman and Hall/CRC Press, 2003 (3rd ed).

[14] Whitfield Diffie and Martin Hellman, *New directions in cryptography,* IEEE Transactions on Information Theory **22** (1976), 644–654.

[15] Euclid (trans. Thomas Heath), *The Thirteen Books of Euclid's Elements,* Dover Publications, 1956.

Online version available at:

`https://mathcs.clarku.edu/~djoyce/java/elements/elements.html`

[16] Leonhard Euler, *Theoremata arithmetica nova methodo demonstrata,* Novi commentarii academiae scientiarum imperialis Petropolitanae **8** (1763), 74–104.

[17] James Franklin & Albert Daoud, *Introduction to Proofs in Mathematics,* Prentice Hall, 1988.

Also available to download at:

`https://web.maths.unsw.edu.au/~jim/proofs.html`

[18] Carl Friedrich Gauss (trans. Arthur Clarke), *Disquitiones Arithmeticae,* Springer, 1986 (2nd ed), from the original (Latin) edition (1801).

[19] Kurt Gödel, *The consistency of the Axiom of Choice and of the Generalized Continuum-Hypothesis,* Proceedings of the National Academy of Sciences of the United States of America **24** (1938), 556–57.

[20] S. H. Gould, *The origin of Euclid's axioms,* Mathematical Gazette **46** (1962), 269–290.

[21] Paul R. Halmos, *Naive Set Theory,* Undergraduate Topics in Mathematics, Springer Verlag, 1974.

[22] David Hilbert *Über das Unendliche,* Mathematische Annalen **95** (1926), 161–190.

[23] Kevin Houston, *How to Think Like a Mathematician,* Cambridge University Press, 2009.

[24] I. M. Isaacs and Thilo Zieschang, *Generating symmetric groups,* American Math. Monthly **102** (1995), 734–739.

[25] Felix Klein, *Vergleichende Betrachtungen über neuere geometrische Forschungen,* Mathematische Annalen **43** (1893), 63–100.

[26] Wilbur Knorr, *The Evolution of the Euclidean Elements*, Synthese Historical Library **15** (1975), D. Reidel Publishing Co.

[27] Steven G. Krantz, *A Primer of Mathematical Writing,* American Mathematical Society, 1997.

[28] Dick Lehmer, *Mathematical methods in large-scale computing units,* Proceedings of a Second Symposium on Large-Scale Digital Calculating Machinery, 1949; Annals of the Computation Laboratory of Harvard University **26** (1951), 141–146.

[29] Martin Liebeck, *A Concise Introduction to Pure Mathematics,* Chapman and Hall/CRC Press, 2011 (3rd ed).

[30] Zoran Lučić, *Irrationality of the Square Root of 2: The Early Pythagorean Proof, Theodorus's and Theaetetus's Generalizations,* Math. Intelligencer **37** (2015), issue 3, 26–32.

[31] Joseph Rotman, *The Theory of Groups: An Introduction,* Allyn and Bacon, 1973 (2nd ed).

[32] Bertrand Russell, *Principles of Mathematics,* Norton, 1938 (first pub. 1903).

[33] Simon Singh, *Fermat's Last Theorem,* Fourth Estate, 2002 (first pub. 1997).

[34] Ian Stewart & David Tall, *The Foundations of Mathematics,* Oxford University Press, 2015 (2nd ed).

[35] Robert R. Stoll, *Set Theory and Logic,* Dover Publications, 2003 (first pub. 1963).

[36] M. G. Teigen & D. W. Hadwin, *On generating Pythagorean triples,* American Math. Monthly **78** (1971), 378–379.

[37] Daniel Velleman, *How to Prove It,* Cambridge University Press, 2019 (3rd ed.).

[38] Franco Vivaldi, *Mathematical Writing,* Springer Undergraduate Mathematics Series, Springer Verlag, 2014.

[39] Walther von Dyck, *Gruppentheoretische Studien,* Mathematische Annalen **20** (1882), 1–44.

[40] Alfred North Whitehead & Bertrand Russell, *Principia Mathematica, Volumes 1–3,* Cambridge University Press, 1925–27 (2nd ed).

[41] Andrew Wiles, *Modular elliptic curves and Fermat's Last Theorem,* Annals of Mathematics **141** (1995), 443–551.

[42] Don Zagier, *Newman's short proof of the Prime Number Theorem,* American Math. Monthly **104** (1997), 705–708.

Department of Mathematics
University of York

Autumn 2022