Draft 2, Version of 11 October 2012

Oxford University Mathematical Institute

An Introduction to Pure Mathematics

Notes by Peter M. Neumann (Queen's College)

Preface

These notes are intended as a rough guide to the eight-lecture course *Introduction to Pure Mathematics* which is a part of the Oxford 1st-year undergraduate course for the Preliminary Examination in Mathematics. Please do not expect a polished account. They are my personal lecture notes, not a carefully checked textbook. Nevertheless, I hope they may be of some help.

Our 1st-year courses are designed to build on your previous experience and the knowledge of mathematics that you have acquired from school or college. The published synopsis for this one is as follows.

- The natural numbers and their ordering. Induction as a method of proof, including a proof of the binomial theorem with non-negative integral coefficients.
- Sets: examples including the natural numbers, the integers, the rational numbers, the real numbers. Inclusion, union, intersection, power set, ordered pairs and cartesian product of sets.
- Relations. Definition of an equivalence relation. Examples.
- Maps: composition, restriction; injective (one-to-one), surjective (onto) and invertible maps; images and preimages.
- Rules for writing mathematics with examples.
- Hypotheses, conclusions,"if", "only if", "if and only if", "and", "or".
- Formulation of mathematical statements with examples. Quantifiers: "for all", "there exists".
- Problem solving in mathematics: experimentation, conjecture, confirmation, followed by explaining the solution precisely.
- Proofs and refutations: standard techniques for constructing proofs; counter examples. Example of proof by contradiction and more on proof by induction.

As your primary reading to support this course we recommend

(⋆) Charles Batty, *How do undergraduates do Mathematics?* (Mathematical Institute Study Guide, 1994).

Paper copies can be purchased from Reception in the Mathematical Institute. It is also available on-line at `http://www.maths.ox.ac.uk/files/study-guide/guide.pdf`.

Further reading:

(1) Geoff Smith, *Introductory Mathematics: Algebra and Analysis* (Springer-Verlag, London, 1998), Chapters 1 and 2.

(2) Robert G. Bartle and Donald R. Sherbert, *Introduction to Real Analysis* (Wiley, New York, Third Edition, 2000), Chapter 1 and Appendices A and B.

(3) C. Plumpton, E. Shipton and R. L. Perry, *Proof* (Macmillan, London, 1984).

(4) R. B. J. T. Allenby, *Numbers and Proofs* (Arnold, London, 1997).

(5) Richard Earl, *Bridging Course in Mathematics* (Mathematical Institute website: use the search box in `http://www.maths.ox.ac.uk`.)

(6) G. Pólya, *How to solve it* (Second edition 1957, reprinted by Penguin Books, 1990). My personal opinion is that this GREAT Classic should be in every mathematician's personal library.

Oxford synopses provide a good guide to the content of the lectures, but it should never be forgotten that it is the syllabus which defines the course. It is the syllabus which should guide student learning, it is the syllabus which specifies to the examiners what they should test, it is the syllabus which suggests to tutors what they should teach, it is the syllabus which provides the basis of the lecture synopses.

A set of two exercise sheets goes with this lecture course. The questions they contain will be found embedded in these notes along with a number of supplementary exercises.

These notes contain much from the notes of the 2011 version of the course. It is a pleasure to acknowledge my great debt to Dr Robin Knight who wrote them.

I would much welcome feedback. Desy Kristianti of Oriel College has already made useful suggestions for which I am very grateful. Please let me know of any further errors, infelicities and obscurities. Email me at `peter.neumann@queens.ox.ac.uk` or write a note to me at Queen's College.

ΠMN: Queen's: Version of 11 October 2012

CONTENTS

* * * * *

1 Prolegomenon

1.1 Some words of advice

We begin with a word. If you do not know it look it up. The point I want to make with the word 'prolegomenon' is that in mathematics (as in life) we are forever coming across words that are new to us. Do not be afraid to look them up. That is what dictionaries and other reference works, whether on paper or on-line, are for.

Also a word of advice. Mathematics is a complex subject involving theory, problem-solving, discovery, conjecture, proof or refutation, examples or counterexamples, applications, and much else beside. One of its characteristics, especially but far from exclusively on the 'pure' side, is a certainty that can rarely be matched in other subjects. That certainty comes from precision in the use of words and symbols and from precision in the use of logical argument. Many words that we use in mathematics are everyday words such as real, rational, complex, imaginary, set, group, ring, function, relation, to which we give special meanings; others, such as homomorphism, isomorphism, homeomorphism, topology, are rarely met outside of mathematics. Although there may be small local variations, generally, the mathematical community has agreed on the technical meanings of such words. If you come across a word you have not met before, or a familiar word in an unfamiliar context, then make sure to look it up.

1.2 Plan for these lectures

The lecture plan is more or less as listed in the synopsis. We'll begin with some basic mathematics including numbers and mathematical induction. Then we'll move on to some of the language of mathematics, the language of sets, relations and functions, and some particularly important examples of each. Next we'll discuss rules for writing mathematics. The importance of these rules or conventions cannot be over-emphasised. Mathematicians need to be able to express their thoughts very precisely. They need to be able to write what they mean and mean what they write. Their precise meaning needs to be understood by the reader.

Sections 5 and 6 contain discussion of proof as it is commonly understood in mathematics. We will discuss some particularly common logical errors and then examine some of the basic elements of logic used by mathematicians. I will say something about problem-solving, something about experimenting and formulating conjectures, and about the importance of examples and counter-examples. We'll end with a discussion of how to discover and then refine and explain one's own proofs.

Sometimes a concept or technique might be used informally before it has been formally defined or discussed. This is so that when we come to our formal discussion we will already have some context to build on. In particular, although I shall not be discussing proofs until the second week, I will feel free to show some examples right from the first lecture.

1.3 Lectures and lecture notes

Pre-prepared lecture notes are—or should be—primarily for a lecturer's own use. They are, if you like, a script for the live event which is the lecture. Your lecturers should feel free to say things that are not in their notes, and not say things that are in their notes. Therefore do take your own notes in lectures. Doing so has several advantages:

(1) You get a record of what actually happened in the lecture, including explanations that occurred to the lecturer on the spur of the moment; diagrams; questions that members of the audience asked; the responses; what, if anything, was particularly emphasised; what struck you personally as noteworthy;

(2) it can help fix the content of the lectures in your memory;

(3) the ability to take accurate notes is a useful life skill.

On this last point: mathematics undergraduates go on to careers of every kind. Twenty years from now you may remember little or none of the material you studied as an undergraduate, and you may well have no need for it, but if you get the practice now, you should be able to take useful minutes of meetings and take accurate notes of any seminars or presentations you attend later in life.

The taking of accurate notes is not a matter of copying what a lecturer writes on the board. For one thing there is no point—most of our lecturers provide written notes such as these—but for another, you will find that most lecturers go too fast for that. The lecturer leads, the audience follows, and following inevitably involves falling behind. Rather, try to follow the lecturer's meaning, and make notes in the real sense of that word. Jottings. But make sure that your notes contain enough to stimulate your memory so that you can reconstruct the narrative and the arguments for yourself afterwards. For example, make a note of diagrams. To my great embarrassment I have been unable to incorporate diagrams in these notes, mainly for lack of time. But a picture is worth a thousand words, and I am sure to use diagrams in the lectures. Create your own personal version of the lecture notes. That is a powerful technique for learning and understanding.

One final observation on lectures: do not expect to understand them instantly. If you do, that's great; but don't worry if you feel lost or left behind. Try to retain at least some idea of the big picture; work through your lecture notes after the lecture; then discuss points that are still puzzling you with your contemporaries and your college tutors.

1.4 The Greek alphabet

Here is the part of the lower case Greek alphabet that is commonly used in mathematics, not in any particular order, with names and approximate Roman equivalents:

α : alpha, a	β : beta, b	γ : gamma, c
δ : delta, d	ϵ or ε : epsilon, e	
θ or ϑ : theta	ϕ or φ : phi	ψ : psi
ι : iota, i	κ : kappa, k	
λ : lambda, l	μ : mu, m	ν : nu, n
ω : omega, o	π or ϖ : pi, p	ρ : rho, r
σ : sigma, s	τ : tau, t	χ : chi
ξ : xi, x	η : eta	ζ : zeta, z

A few of the upper case (capital letters) are also commonly used in mathematics:

$\gamma \mapsto \Gamma$;	$\delta \mapsto \Delta$;	$\theta \mapsto \Theta$;	$\phi \mapsto \Phi$;
$\lambda \mapsto \Lambda$;	$\omega \mapsto \Omega$;	$\pi \mapsto \Pi$;	$\sigma \mapsto \Sigma$.

You might see Ξ (capital ξ), but it is rare. The letter Σ and the summation symbol $\sum$ look alike, but technically they are typographically different; the same goes for the letter Π and the product symbol $\prod$.

2 Numbers and induction

2.1 Natural numbers

Numbers are, of course, fundamental for most parts of mathematics and its applications. Most of us are comfortable with various different kinds of numbers. Yet even at this very basic level we need to define our terms, or perhaps simply agree usage and conventions.

DEFINITION 2.1. A *natural number* is a member of the sequence 0, 1, 2, 3, ... obtained by starting from 0 and adding 1 successively. We write $\mathbb{N}$ for the collection (technically, the set—see below) $\{0, 1, 2, 3, \ldots\}$ of all natural numbers.

IMPORTANT NOTE. Although nowadays all mathematicians use $\mathbb{N}$ as a standard name for the collection of natural numbers, some authors and lecturers include 0 as a natural number, some do not. Unfortunately, there is no standard convention. In this course, 0 will belong to $\mathbb{N}$. Just make sure when reading texts or discussing mathematics with others that you clarify such notational conventions before you get in deep.

Natural numbers have many familiar and important properties. For example, they can be added and multiplied—that is, if m, n are natural numbers then so are $m+n$ and $m \times n$—and they may be compared: $m < n$ if m occurs earlier in the sequence than n. Furthermore $\mathbb{N}$ is *well-ordered*, that is, any non-empty collection of natural numbers has a least (or first) member. Although there will be further discussion of some of these points later (see § 6.3), all this will, for the purposes of your 1st year course (and for much of the 2nd year too), be taken for granted. Notice, however, that a purist could find several points of attack. In the definition I have used the terms 0, 1, 'member', 'sequence' and 'adding successively', and in the discussion I have used such terms as 'added', 'multiplied', 'occurs earlier', 'least'. Do we all agree what these mean? At this level it would be very surprising if we did not. Nevertheless, there is an exciting area where mathematics and philosophy come together which deals with such questions. It is represented in the Oxford mathematics degree in the Foundations courses (Logic and Set Theory) offered in Parts B and C.

2.2 Mathematical Induction

THEOREM 2.2 [**Mathematical Induction**]. *Let P be a property of natural numbers, that is, a statement $P(x)$ about natural numbers x that may be true for some natural numbers x and false for others. Suppose that $P(0)$ is true, and that for all natural numbers n, if $P(n)$ is true then $P(n+1)$ is also true. Then $P(n)$ is true for all natural numbers n.*

I have called this a theorem, but what proof can we give? Intuitively it seems clear: we are given that $P(0)$ holds, then its truth implies that of $P(1)$, the truth of $P(1)$ implies that of $P(2)$, and so on. Nevertheless, it says something non-trivial. It collects together all the individual true statements, $P(0)$, $P(1)$, $P(2)$, ..., of which there are infinitely many, into just one true statement, namely '$P(n)$ is true for all natural numbers n'. It is so basic that in some versions of the logical foundations of mathematical thinking it is taken as an axiom.

Mathematical Induction is a technique for proving theorems. It goes with a method for defining functions called *recursion*. Two typical definitions by recursion are the following:

DEFINITION 2.3 [**Powers**]. Let a be a number (or a variable that takes numerical values). Define $a^0 := 1$ and then define $a^{n+1} := a^n \times a$ for $n \geqslant 0$

Read the symbol := as 'to be' or as 'is defined to be'. It is quite different from =, 'is equal to', which indicates that two previously defined entities are the same.

DEFINITION 2.4 [**Factorials**]. Define $n!$ for all $n \geqslant 0$ by the rule that $0! := 1$, and thereafter $(n+1)! := n! \times (n+1)$ for all $n \geqslant 0$.

These are typical of recursion in that it is used to define a function of a natural number by specifying what value it takes at 0, and saying also how to get from the value it takes at n to the value it takes at $n+1$. The second function defined above is the familiar factorial function, which we commonly define informally by writing $n! := 1 \times 2 \times 3 \times \cdots \times n$.

Note that the definitions $a^0 := 1$, $0! := 1$ are made for good reason. It makes sense that a product of no factors should be 1. After all, if we have a product of a number of factors, and then add in no more factors, we do not change the product, that is, we have multiplied it by 1. For a very similar reason, we take the sum of no numbers to be 0: if we start with a sum of some numbers and then add no summands we do not change the original sum, so what we have added is 0.

One use of the factorial function is to define the following extremely useful function of two variables:

DEFINITION 2.5 [**Binomial coefficients**]. For natural numbers m, n with $m \leqslant n$ define $\binom{n}{m} := \dfrac{n!}{m!(n-m)!}$.

Famously, the binomial coefficients may be organised into an array commonly called Pascal's Triangle, whose defining property is captured in the following lemma.

LEMMA 2.6. [**Pascal's Triangle**]. *Let m, n be natural numbers such that $1 \leqslant m \leqslant n$. Then*

$$\binom{n}{m-1} + \binom{n}{m} = \binom{n+1}{m}.$$

EXERCISE 2.1. Write out a proof of Lemma 2.6, using just our definitions of the factorial function and binomial coefficients, nothing more.

As a good illustration of how Mathematical Induction may be used we give a proof of a very famous and important theorem:

THEOREM 2.7 [**The Binomial Theorem** (for non-negative integral exponents)]. *Let x, y be numbers (or variables that may take numerical values). Then for every natural number n, $(x+y)^n = \displaystyle\sum_{m=0}^{n} \binom{n}{m} x^{n-m} y^m$.*

Proof. Let $P(n)$ be the assertion that the equation above is true for the natural number n. Certainly $P(0)$ is true since $(x+y)^0 = 1$ while the sum on the right of the equation has just one term, namely $\binom{0}{0} x^0 y^0$, which also is equal to 1.

Now suppose that $P(n)$ is true, so that $(x+y)^n = \displaystyle\sum_{m=0}^{n} \binom{n}{m} x^{n-m} y^m$. Then

$$
\begin{aligned}
(x+y)^{n+1} = (x+y)^n(x+y) &= \Bigg(\sum_{m=0}^{n} \binom{n}{m} x^{n-m}y^m\Bigg)\Big(x+y\Big) \qquad [\text{by } P(n)] \\
&= \sum_{m=0}^{n} \binom{n}{m} x^{n-m+1}y^m + \sum_{m=0}^{n} \binom{n}{m} x^{n-m}y^{m+1} \\
&= x^{n+1} + y^{n+1} + \sum_{m=1}^{n} \left(\binom{n}{m} + \binom{n}{m-1}\right) x^{n+1-m}y^m,
\end{aligned}
$$

that is, by Lemma 2.6 $\left(\text{together with the definitions of } \binom{n+1}{0} \text{ and } \binom{n+1}{n+1}\right)$,

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

which is the statement $P(n+1)$. By induction, therefore, the equation holds for all natural numbers n, as the theorem states.

EXERCISE 2.2. Write careful proofs by induction of formulae for $\sum_{m=0}^{n} m$, $\sum_{m=0}^{n} m^2$, $\sum_{m=0}^{n} m^3$.

EXERCISE 2.3. Show that if $S_k(n) := \sum_{m=0}^{n} m^k$ then $S_k(n)$ may be expressed as a polynomial of degree $k+1$ in n, whose leading term is $\dfrac{n^{k+1}}{k+1}$.

EXERCISE 2.4. The following famous argument purports to show that all natural numbers are small. 'Certainly 0 is small. If n is small then also $n+1$ is small. Therefore by induction all natural numbers are small.' What is missing?

NOTE: I have learned from Desi Kristianti of Oriel College that this is known as Wang's Paradox. It is not what I would call a paradox. When my friend and I were shown it by one of our schoolteachers we all treated it as a joke. I still think of it as a joke—though a joke with a serious moral. The moral is: make sure to define your terms precisely.

EXERCISE 2.5. Show that $\sum_{m=0}^{n} \binom{n}{m} = 2^n$, and that

$$
\sum_{\substack{0 \leqslant m \leqslant n, \\ m \text{ even}}} \binom{n}{m} = \sum_{\substack{0 \leqslant m \leqslant n, \\ m \text{ odd}}} \binom{n}{m} = 2^{n-1}.
$$

EXERCISE 2.6. Show that $\dfrac{4^n}{2n} < \binom{2n}{n} < 4^n$ if $n \geqslant 2$. How far can you refine these inequalities?

3 Sets

Set theory, since its introduction about 120 years ago, has become a sophisticated part of mathematics with deep theorems that can be studied in Parts B and C of the Oxford degree. The basic concept is simple, however, and the beginnings of the subject provide a precise and convenient language that has revolutionised the way mathematicians think about and present their arguments.

3.1 Sets, examples of sets

We begin with a classic description of what the word 'set' should mean:

Note that the membership relation $\in$ should look different from the Greek letter ϵ (epsilon).

DEFINITION 3.1. A *set* is any collection of individuals. We write $x \in X$ to mean that x is a member of a set X. The members of a set are often called its *elements*. Two sets are equal if and only if they have the same elements.

Although this definition is similar to that formulated by Georg Cantor in 1895 there are difficulties with it. For one thing, it defines a technical term 'set' by reference to two undefined terms 'collection' and 'individual'. At this stage that should not create problems. There are, however, more serious problems which are discussed in the second/third year Set Theory course. (Anyone who is interested can look up set theoretic paradoxes, such as the famous one called Russell's Paradox.)

One particularly important set:

Make sure you make $\emptyset$ look different from ϕ, the Greek letter phi.

DEFINITION 3.2. The empty set, written $\emptyset$, is the set with no elements.

DEFINITION 3.3. Curly brackets (braces) are used to show sets. The set whose elements are a_1, a_2, a_3, ..., a_n is written $\{a_1, a_2, a_3, \ldots, a_n\}$. Similarly, the set whose members are those of an infinite sequence a_1, a_2, a_3, ... of objects is denoted $\{a_1, a_2, a_3, \ldots\}$.

EXAMPLES 3.4. The sets $\{0, 1\}$ and $\{1, 0\}$ have the same elements, so they are equal. Similarly, $\{2, 2\}$ and $\{2\}$ have the same elements, and so are equal.

A common error to avoid: never confuse a with $\{a\}$, the set whose only element is a. For example, if $a = \emptyset$, then a has no elements, but $\{a\}$ has one element (namely a), so they cannot be equal. Or if $a = \mathbb{N}$ then a is infinite, but $\{a\}$ is not.

We also have notation for a set whose members are identified by a property.

DEFINITION 3.5. Let P or $P(x)$ be a property, that is, an assertion involving a variable x that may be true (or false) of any given individual x. Then $\{x \mid P(x)\}$, also written $\{x : P(x)\}$, is the set of all objects x having the property $P(x)$. Read it as 'the set of all x such that $P(x)$' or 'the set of all x such that P holds for x'. If A is a set, and $P(x)$ is a property then we write $\{x \in A \mid P(x)\}$ or $\{x \in A : P(x)\}$ for the set consisting of those elements x of A that have the property P.

EXAMPLES 3.6. The set of even natural numbers is $\{n \in \mathbb{N} \mid n \text{ is even}\}$. We could write the set of primes as $\{n \mid n \text{ is a prime number}\}$, or as $\{n \in \mathbb{N} \mid n \text{ is prime}\}$. The set $\{1, 2, 3, 4, 6, 12\}$ is equal to $\{n \in \mathbb{N} \mid n \text{ is a factor of } 12\}$. We could write $\emptyset = \{x \mid x \neq x\}$ or $\emptyset = \{n \in \mathbb{N} \mid n^2 < 0\}$.

Some other important sets:

$\mathbb{N}$ is the set $\{0, 1, 2, 3, \ldots\}$ of all natural numbers (recall the warning on p. 3 above—some authors do not include 0);

$\mathbb{Z}$ is the set of all integers (positive, negative, or zero) [Z is the first letter of the German word *Zahlen* 'numbers'];

$\mathbb{Q}$ is the set of all rational numbers [Q for quotient];

$\mathbb{R}$ is the set of all real numbers;

$\mathbb{C}$ is the set of all complex numbers.

All the above are written in the 'blackboard bold' font which was originally a way of writing bold-face letters on a blackboard, but has since taken on an independent life. A purist (such as myself) would insist that this notation should be reserved for the sets equipped with their most important structure (addition, multiplication, etc.). For the purposes of these lectures, however, we'll use it just for the sets. You'll find that lecturers use variations on this notation to denote closely related sets. Thus for example $\mathbb{R}^+$ or $\mathbb{R}^{>0}$ often denotes the set of positive real numbers; $\mathbb{C}^*$ often denotes the set of non-zero complex numbers. Usually the intention should be clear from the context. Where it is not, ask.

There are many interesting and important examples of sets that consist of real numbers. Perhaps the most commonly occurring are the *intervals* described as follows.

DEFINITION 3.7 [**Real intervals**]. Let a, b be real numbers with $a \leqslant b$. the following are known as *intervals*:

(1) $(a, b) := \{x \in \mathbb{R} \mid a < x < b\}$ [open interval];
(2) $[a, b] := \{x \in \mathbb{R} \mid a \leqslant x \leqslant b\}$ [closed interval];
(3) $(a, b] := \{x \in \mathbb{R} \mid a < x \leqslant b\}$ [half open interval];
(4) $[a, b) := \{x \in \mathbb{R} \mid a \leqslant x < b\}$ [half open interval];
(5) $(a, \infty) := \{x \in \mathbb{R} \mid a < x\}$;
(6) $[a, \infty) := \{x \in \mathbb{R} \mid a \leqslant x\}$;
(7) $(-\infty, b) := \{x \in \mathbb{R} \mid x < b\}$;
(8) $(-\infty, b] := \{x \in \mathbb{R} \mid x \leqslant b\}$;
(9) $(-\infty, \infty) := \mathbb{R}$.

Note that if $a = b$ then $[a, b] = \{a\}$ and $(a, b) = (a, b] = [a, b) = \emptyset$. Check that you understand why this follows from the definitions. Note also that we use the symbol ∞ in this context without giving it an independent meaning. It is NOT a real number. It is easy to see (though perhaps tedious to write out because of the many cases) that an interval S in $\mathbb{R}$ has the property that if $x, y \in S$ and $x \leqslant z \leqslant y$ then also $z \in S$. In fact, the converse holds. Any set S with this property is an interval. But to prove this one needs the *completeness* of $\mathbb{R}$, a matter that will be treated in your Analysis course.

EXERCISE 3.1. Let $S := \{x \in \mathbb{Q} \mid x^2 \leqslant 2\}$. Show that S has the property that if $x, y \in S$ and $x \leqslant z \leqslant y$ then also $z \in S$. Show also that S is not a rational interval—that is, there do not exist $a, b \in \mathbb{Q}$ such that $S = \{x \in \mathbb{Q} \mid a < x < b\}$ (or $S = \{x \in \mathbb{Q} \mid a \leqslant x \leqslant b\}$, or any one of the other possibilities for an interval).

3.2 Some algebra of sets

We begin with set containment or set inclusion.

DEFINITION 3.8 **[Subsets]**. The set A is said to be a *subset* of a set B if every member of A is also a member of B. The notation is $A \subseteq B$ or $B \supseteq A$. If $A \subseteq B$ and $A \neq B$ then we call A a *proper* subset of B.

EXAMPLES 3.9. Note that $\emptyset \subseteq X$ for every set X. Also $\mathbb{N} \subseteq \mathbb{Z} \subseteq \mathbb{Q} \subseteq \mathbb{R} \subseteq \mathbb{C}$, and any real interval S is a subset of $\mathbb{R}$.

The containment $\emptyset \subset X$ is not simply convention. It follows from the definition. After all, it is certainly true that every member of $\emptyset$ is a member of X.

Just as $\neq$ means 'is not equal to' and $\not\leqslant$ means 'is not less than or equal to' so we often draw a line through other relation symbols to negate them. Thus $a \notin A$ means that a is not a member of A and $A \not\subseteq B$ means that A is not a subset of B (that is, there is some object $a \in A$ such that $a \notin B$).

OBSERVATION 3.10. *Let A, B be sets. Then $A = B$ if and only if both $A \subseteq B$ and $B \subseteq A$.*

Proof. Certainly, if $A = B$ then every member of A is a member of B, so $A \subseteq B$, and similarly, $B \subseteq A$. Conversely, if $A \subseteq B$ and $B \subseteq A$ then for every x, $x \in A$ if and only if $x \in B$, so A, B have the same members and therefore, by definition of set equality, $A = B$.

Simple though this observation is, you will often find that when you wish to prove two sets equal, breaking the problem down into the two complementary containments helps greatly.

This is where diagrams, so-called Venn diagrams, are very instructive.

DEFINITION 3.11 **[Set union, intersection, difference]**. Let A, B be sets. We define their *union* (sometimes also called 'join') by

$$A \cup B := \{x \mid x \in A \text{ or } x \in B \text{ (or both)}\}.$$

We define their *intersection* (sometimes also called 'meet') by

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

We define their *set difference* by

$$A \setminus B := \{x \mid x \in A \text{ and } x \notin B\}.$$

EXAMPLES 3.12. If $A := \{n \in \mathbb{N} \mid n \text{ is even}\}$ and $B := \{n \in \mathbb{N} \mid n \text{ is prime}\}$ then $A \cap B = \{2\}$.

$\{0,1,2\} \cup \{2,3\} = \{0,1,2,3\}$; $\{0,1,2\} \cap \{2,3\} = \{2\}$; $\{0,1,2\} \setminus \{2,3\} = \{0,1\}$.

THEOREM 3.13. *Let A, B, C be sets. Then $A \cup (B \cap C) = (A \cup B) \cap (A \cup C)$.*
Also $A \cap (B \cup C) = (A \cap B) \cup (A \cap C)$.

Proof (of first part). We use Observation 3.10. Suppose first that $x \in A \cup (B \cap C)$. Then either $x \in A$ or $x \in B \cap C$. Thus either $x \in A$ or x is in both B and C. If $x \in A$ then $x \in A \cup B$ and $x \in A \cup C$ so $x \in (A \cup B) \cap (A \cup C)$. If x is in both B and C then x is in both $A \cup B$ and $A \cup C$, and so $x \in (A \cup B) \cap (A \cup C)$. Thus every member of $A \cup (B \cap C)$ lies in $(A \cup B) \cap (A \cup C)$. That is $A \cup (B \cap C) \subseteq (A \cup B) \cap (A \cup C)$.

Now suppose that $x \in (A \cup B) \cap (A \cup C)$. Then x is in both $A \cup B$ and $A \cup C$. Thus either $x \in A$ or, if $x \notin A$, then $x \in B$ and also $x \in C$. Thus $x \in A \cup (B \cap C)$. Hence $(A \cup B) \cap (A \cup C) \subseteq A \cup (B \cap C)$. Therefore these two sets are equal.

EXERCISE 3.2. Write a proof of the second part of the theorem. That is, prove that $A \cap (B \cup C) = (A \cap B) \cup (A \cap C)$.

THEOREM 3.14 **[De Morgan's Laws]**. *Let A, B be subsets of a set X. Then*

$$X \setminus (A \cup B) = (X \setminus A) \cap (X \setminus B) \quad \textit{and} \quad X \setminus (A \cap B) = (X \setminus A) \cup (X \setminus B).$$

The proof is left as an exercise:

EXERCISE 3.3. Write a proof of De Morgan's Laws.

Sometimes we have a family of sets $\{A_i\}_{i \in I}$ indexed by a set I. For example, we may have sets $A_1, A_2, \ldots, A_n$, or we may have sets $A_1, A_2, \ldots, A_n, \ldots$, one for each natural number, or we could have sets A_x, one for each $x \in \mathbb{R}$. Then union and intersection are defined by

$$\bigcup_{i \in I} A_i := \{x \mid x \in A_i \text{ for at least one } i \in I\}$$

and, *provided that* $I \neq \emptyset$,

$$\bigcap_{i \in I} A_i := \{x \mid x \in A_i \text{ for every } i \in I\}.$$

Note that if I has two members, say $I = \{1, 2\}$ then the union of the family is simply $A_1 \cup A_2$, and the intersection is just $A_1 \cap A_2$.

EXERCISE 3.4. Let A be a set and B_i a family of sets indexed by a non-empty set I. Show that

$$A \cap \Big(\bigcup_{i \in I} B_i\Big) = \bigcup_{i \in I} (A \cap B_i) \quad \text{and} \quad A \cup \Big(\bigcap_{i \in I} B_i\Big) = \bigcap_{i \in I} (A \cup B_i).$$

EXERCISE 3.5. Let X be a set and A_i a family of subsets of X indexed by a non-empty set I. Show that

$$X \setminus \Big(\bigcup_{i \in I} A_i\Big) = \bigcap_{i \in I} (X \setminus A_i) \quad \text{and} \quad X \setminus \Big(\bigcap_{i \in I} A_i\Big) = \bigcup_{i \in I} (X \setminus A_i).$$

DEFINITION 3.15 **[Cardinality of a set]**. Suppose that $A = \{a_1, a_2, \ldots, a_n\}$ where $a_i \neq a_j$ whenever $i \neq j$. Thus A is a set with n elements (where n is a natural number). Then we write $|A| = n$, and say that n is the *cardinality* (or size) of $|A|$.

Perhaps better would be to use recursion to define finiteness of a set and the cardinality of a finite set: $\emptyset$ is finite and $|\emptyset| = 0$; then if A is finite with $|A| = n$, and $B = A \cup \{b\}$, where $b \notin A$ then also B is finite, and $|B| = n + 1$. It is then a non-trivial fact, but one which we shall take for granted, that if X is finite and $|X| = n + 1$, and Y is obtained from X by removing any member x, no matter which, then Y is finite and $|Y| = n$. You may like to think why I describe this as non-trivial, and how you would set about justifying the assertion.

EXERCISE 3.6. Let A, B be finite sets. Show that

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

Deduce that $|A| + |B| = |A \cup B| + |A \cap B|$.

The sizes, that is cardinalities, of infinite sets will be touched on the Analysis course.

DEFINITION 3.16 **[Power set]**. We define the *power set* of a set A by $\wp A := \{X \mid X \subseteq A\}$. That is, the power set is the set of all subsets of A.

THEOREM 3.17. *Let A be a finite set with $|A| = n$. Then $\wp A$ is finite and $|\wp A| = 2^n$.*

Proof. We use induction. If $|A| = 0$ then A has no members, that is, $A = \emptyset$. Since $\emptyset$ is the only subset of $\emptyset$, $\wp\emptyset = \{\emptyset\}$. Thus $|\wp A| = 1 = 2^0$.

Now suppose that $n \geqslant 0$ and that $|\wp X| = 2^n$ for any set X of size n. Let A be a set with $|A| = n + 1$. Since $n + 1 \geqslant 1$, $A \neq \emptyset$ and so there is a member a of A. Let $B := A \setminus \{a\}$. Then $|B| = n$ and so, by inductive hypothesis, $|\wp B| = 2^n$. Those subsets of A that do not have a as a member are subsets of B, so there are 2^n of them. Those subsets of A that do have a has a member are of the form $\{a\} \cup C$ where C ranges over subsets of B, and so again there are 2^n of them. Since any subset of A does or does not contain a, we see that $|\wp A| = 2^n + 2^n = 2^{n+1}$. Thus, by induction, the theorem is true for all finite sets A.

EXERCISE 3.7. Let k be a natural number, and for a set A let $\wp_k(A)$ be the set of its subsets of size k (that is, $\wp_k(A) := \{B \in \wp A \mid |B| = k\}$). Use induction on n (or perhaps more precisely on $n - k$) together with Lemma 2.6 (Pascal's Triangle) to show that if $k \leqslant n$ then $|\wp_k(A)| = \binom{n}{k}$.

3.3 Ordered pairs; cartesian products of sets

DEFINITION 3.18. The ordered pair whose first element is a and whose second element is b, is written (a, b). The defining property is that $(a, b) = (c, d)$ if and only if $a = c$ and $b = d$.

The point is that, in an ordered pair one member is first, the other second. Contrast this with the unordered pair $\{a, b\}$, where we cannot distinguish first and second elements; $\{a, b\}$ and $\{b, a\}$ have the same elements, and so they are equal.

Warning: there is a problem with notation here. If $a, b \in \mathbb{R}$ are real numbers and $a < b$, then the ordered pair whose first element is a and whose second element is b, and the open interval between a and b, are both written (a, b). Usually the context will indicate what is intended, but if, in your work, there is the possibility of confusion, then remove the ambiguity using words to clarify. Write something like 'the open interval (a, b)', or 'the ordered pair (a, b)'. We could of course change our notation: for example, we could write the ordered pair as $\langle a, b\rangle$ or the open interval as $]a, b[$ (though it is bad policy to use brackets the wrong way round since the eye expects to pair brackets as $[...]$). No such scheme is accepted by everybody. One reason is that there are more mathematical concepts than notations to represent them, so ambiguities like this will unfortunately but inevitably occur.

We can also define ordered triples (a, b, c), ordered quadruples (a, b, c, d), etc. in the same manner. A sequence n long is called an n-tuple (though NOT if n is small).

DEFINITION 3.19. The Cartesian product of sets A, B (which may be the same) is defined by

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

If $A = B$, we also write $A \times A$ as A^2. More generally, we define $A_1 \times A_2 \times \cdots \times A_n$ to be the set of all ordered n-tuples $(a_1, a_2, \ldots, a_n)$ such that $a_i \in A_i$ for $1 \leqslant i \leqslant n$.

The product of n copies of A may be written as A^n. Note that $A^1 = A$ and $A^0 = \{\emptyset\}$. Note also that the elements of the Cartesian product $A \times B$ are ordered pairs (a, b). They are never written $a \times b$.

The sets $A \times B \times C$, $(A \times B) \times C$ and $A \times (B \times C)$ are not equal. Members of the first are ordered triples with first member from the set A, second from the set B, third from the set C, whereas those of the second are ordered pairs whose first member is an ordered pair, and those of the third are ordered pairs whose second member is an ordered pair. Nevertheless, they may be identified in a natural way, and few of us would be so pedantic as to distinguish them except in very special circumstances. Thus to write $A \times B \times C = (A \times B) \times C = A \times (B \times C)$ is wrong, but usually harmless.

EXAMPLES 3.20. The most familiar example of a Cartesian product is the Euclidean plane which we regard as being equal to the set of ordered pairs of real numbers $\mathbb{R}^2$ or $\mathbb{R} \times \mathbb{R}$. Indeed, it is by rather imaginative and far-fetched analogy with this important example that mathematicians have come to associate Descartes' name with the general set-theoretic construction.
If $A = \{1, 2\}$ and $B = \{3, 4, 5\}$, then $A \times B = \{(1, 3), (1, 4), (1, 5), (2, 3), (2, 4), (2, 5)\}$, while $B \times A = \{(3, 1), (3, 2), (4, 1), (4, 2), (5, 1), (5, 2)\}$.

THEOREM 3.21. *Let A, B, C, D be sets. Then $(A \times B) \cap (C \times D) = (A \cap C) \times (B \cap D)$.*

Proof. Let $(x, y) \in (A \times B) \cap (C \times D)$. Then $(x, y) \in A \times B$ and $(x, y) \in C \times D$. Since $(x, y) \in A \times B$, $x \in A$ and $y \in B$. Similarly $x \in C$ and $y \in D$. Therefore $x \in A \cap C$ and $y \in B \cap D$. So $(x, y) \in (A \times B) \cap (C \times D)$. Thus $(A \times B) \cap (C \times D) \subseteq (A \cap C) \times (B \cap D)$.

Now let $(x, y) \in (A \cap C) \times (B \cap D)$. Then $x \in A \cap C$ and $y \in B \cap D$. Therefore $x \in A$ and $y \in B$, so $(x, y) \in A \times B$, but also $x \in C$ and $y \in D$, so $(x, y) \in C \times D$. Thus $(x, y) \in (A \times B) \cap (C \times D)$. This shows that $(A \cap C) \times (B \cap D) \subseteq (A \times B) \cap (C \times D)$. By Observation 3.10 therefore, $(A \times B) \cap (C \times D) = (A \cap C) \times (B \cap D)$, as required.

EXERCISE 3.8. Is it the case that for all sets A, B, C, D,

$$(A \times B) \cup (C \times D) = (A \cup B) \times (C \cup D)?$$

EXERCISE 3.9. Prove that if A, B are finite sets then $A \times B$ is finite and $|A \times B| = |A| \times |B|$.

4 Relations and functions

In ordinary conversation a relationship can hold between one or more objects: 'tar is black' uses being black as a unary (that is, 1-place) relationship; 'Elizabeth is the mother of Charles' uses being mother of as a binary (2-place) relationship; 'Milton Keynes lies between Oxford and Cambridge' uses being between as a ternary (3-place) relationship. In these lectures we will focus on binary relationships as they are needed for mathematics.

4.1 Binary relations

In mathematics a binary relation is something like $=$, $\leqslant$ or $\subseteq$ which asserts a certain relationship between two objects. We formalise this idea by identifying a relationship $a\,R\,b$ with the set of ordered pairs (a, b) which are connected by the relation.

DEFINITION 4.1. A *binary relation* between sets A and B is a subset of $A \times B$. A binary relation on a set A is a subset of $A \times A$. If R is a binary relation, we write $(a, b) \in R$ and $a\,R\,b$ interchangeably.

Although unary, ternary, ..., k-ary relations do occur in mathematics, they are far less common than binary relations. It is usual therefore to use 'relation' to mean 'binary relation', and I shall do so from now on.

EXAMPLES 4.2. *The successor relation on* $\mathbb{N}$ *is the set* $\{(a, b) \in \mathbb{N}^2 \mid b = a + 1\}$.
The order relation on the set of real numbers is the set $\{(a, b) \in \mathbb{R}^2 \mid a \leqslant b\}$.
For a set X *the subset relation on* $\wp X$ *is the relation* $\{(A, B) \in (\wp X)^2 \mid A \subseteq B\}$.

There are very many different kinds of relations. One of the most important kinds is the equivalence relation, which asserts that two objects are, in some sense, to be treated as being the same. To prepare for the notion we need some further terminology.

DEFINITION 4.3. Let R be a relation on a set A. To say that

R is *reflexive* means that $a\,R\,a$ for all $a \in A$;

R is *symmetric* means that whenever $a, b \in A$ and $a\,R\,b$ also $b\,R\,a$;

R is *transitive* means that whenever $a, b, c \in A$ and both $a\,R\,b$ and $b\,R\,c$ also $a\,R\,c$.

EXAMPLES 4.4. The relations $=$, $\leqslant$, $\subseteq$ are reflexive, the relations $\neq$, $<$ are not. Relations $=$, $\neq$, 'have the same size' (for sets) are symmetric; relations $<$, $\subseteq$ are not. Relations $=$, $\leqslant$, $\subseteq$ are transitive; relations $\neq$, 'is the immediate successor of ' (on $\mathbb{N}$) are not transitive.

EXERCISE 4.1. Which of the following relations on $\mathbb{N}$ are reflexive, which are symmetric, which are transitive?

(i) the relation $a \mid b$ (read as 'a divides b');
(ii) the relation $a \nmid b$ (does not divide);
(iii) a, b are related if a, b leave the same remainder after division by 2012;
(iv) a, b are related if $\mathrm{hcf}\,(a, b) > 2012$.

DEFINITION 4.5. A relation R on a set A is said to be an *equivalence relation* if and only if R is reflexive, symmetric and transitive. Symbols $\sim$, $\approx$, $\simeq$, $\equiv$, and others like them, are often used to denote a relation that is known to be an equivalence relation.

EXAMPLES 4.6. The relation of equality on any set is always an equivalence relation. For any set A, if $R := A \times A$ then R is an equivalence relation (the 'universal' relation). The relation R on $\mathbb{N}$ such that $m\,R\,n$ if and only if m and n have the same number of prime factors, is an equivalence relation.
The relation of being congruent is an equivalence relation on the set of triangles in $\mathbb{R}^2$.
The relation of being similar is an equivalence relation on the set of triangles in $\mathbb{R}^2$.
The relation $\leqslant$ on $\mathbb{R}$ is not symmetric, so is not an equivalence relation.
The relation R on $\mathbb{R}$ such that $x\,R\,y$ if and only if $|x-y|<1$ is not transitive, so is not an equivalence relation.

It is very often the case in mathematics that a situation can be fruitfully viewed in more than one way. That is certainly the case with equivalence relations. An equivalence relation on a set A is a way of saying that two elements of A are 'essentially' the same. It divides A up into subsets of elements that are in some way the same as each other. That is, it gives rise to a partition of A.

DEFINITION 4.7. A *partition* of a set A is a set Π of subsets of A with the following properties:

(1) $\emptyset \notin \Pi$ (that is, all the sets in Π are non-empty);
(2) $\bigcup_{P\in\Pi} P = A$ (that is, every member of A lies in one of the members of Π);
(3) if $P, Q \in \Pi$ and $P \neq Q$ then $P \cap Q = \emptyset$.

EXAMPLES 4.8. If $\Pi := \{\{2n : n \in \mathbb{N}\},\ \{2n+1 : n \in \mathbb{N}\}\}$ then Π is a partition of $\mathbb{N}$ (into two parts);
If $\Pi := \{\{0\},\ \{1,4,5\},\ \{2,3\}\}$ then Π is a partition of $\{0,\ 1,\ 2,\ 3,\ 4,\ 5\}$ (into three parts).

EXERCISE 4.2. How many partitions are there of a set of size 1? of size 2? of size 3? of size 4? of size 5?

OBSERVATION 4.9. *Each partition of a set A is naturally associated with an equivalence relation on A. Indeed, given the partition Π we define $a \sim b$ to mean that the elements a, b of A lie in the same part of Π.*

Formally this says that $a \sim b$ if there is some member P of Π such that $a, b \in P$. Since the union of the members of Π is the whole of A, for any $a \in A$ there must exist some $P \in \Pi$ with $a \in P$. Then, trivially, a and a both lie in P, so $a \sim a$, that is, the relation is reflexive. Also, if $a \sim b$ then $a, b \in P$ where $P \in \Pi$, and then of course also $b, a \in P$, so $b \sim a$. Thus the relation is symmetric. Lastly, if $a \sim b$ and $b \sim c$ then there exist $P, Q \in \Pi$ such that $a, b \in P$ and $b, c \in Q$. But now $b \in P \cap Q$, so $P \cap Q \neq \emptyset$, and therefore by condition (3) for a partition, $P = Q$. Therefore $a, c \in P$ so $a \sim c$, and so the relation is transitive. Being reflexive, symmetric and transitive $\sim$ is an equivalence relation.

Conversely, any equivalence relation on a set A naturally defines a partition of A.

DEFINITION 4.10. Let $\sim$ be an equivalence relation on a set A. For $a \in A$ define $[a] := \{b \in A \mid a \sim b\}$, the *equivalence class* of a.

That the equivalence classes form a partition of the set A is a theorem that is to be proved in a later Prelim lecture course. (It is not particularly hard, so you may like to anticipate and find a proof for yourself.) Thus equivalence relations and partitions correspond to each other in a natural way.

4.2 Functions

The concept of a function from a set A to a set B is simple enough: it is a rule assigning exactly one element of B to each element of A. We formalise the concept by formulating it in set-theoretic language as a special kind of binary relation:

DEFINITION 4.11. A *function* from a set A to a set B is a binary relation f between A and B such that for each $a \in A$ there is exactly one $b \in B$ such that $(a, b) \in f$. We write $f(a) = b$ or sometimes $f : a \mapsto b$. We write $f : A \to B$ to mean that f is a function from A to B.

If $f : A \to B$, we refer to A as its *domain* and B as its *codomain*.

This definition makes clear that if $f : A \to B$ and $g : A \to B$ then $f = g$ if and only if $f(a) = g(a)$ for every $a \in A$.

Warning: always make sure that your recipe for defining a function makes sense. For example, if we are seeking to define a function $f : \mathbb{R} \to \mathbb{R}$, then the recipe $f(x) := 1/x$ fails since $f(0)$ is undefined; similarly, the recipe $f(x) := y$ where $y^2 = x$ fails since it does not return a unique value—is $f(4)$ equal to 2 or -2? In such cases, where either $f(x)$ cannot always be defined, or where $f(x)$ appears to take more than one value, there is something wrong with the definition: we say that f is *ill-defined*. Our interest is in *well-defined* functions.

DEFINITION 4.12. For $f : A \to B$ the set of values $\{f(a) \in B \mid a \in A\}$ is known as the *range* or the *image* of f. We emphasize that it is a subset of B.

More generally, for $f : A \to B$ and $X \subseteq A$ we define the *image* of X (under f) by $f(X) := \{f(x) \in B \mid x \in X\}$. Thus the range of f is $f(A)$.

Warning: the terminology of codomains and ranges is not always used consistently. Make sure that you work out from the context how a writer intends these words.

DEFINITION 4.13. For $f : A \to B$ and $Y \subseteq B$ we define the *preimage* of Y (under f) by $f^{-1}(Y) := \{x \in A \mid f(x) \in Y\}$

Warning: there are serious possibilities of notational confusion here. If $X \subseteq A$ and $x \in A$ then $f(X)$ and $f(x)$ look similar, even though they are different kinds of object: the former is a set (a subset of B), the latter a single value (a member of B). It is even worse with the preimage $f^{-1}(Y)$: it is an important piece of notation for an important concept even when f^{-1} has no meaning on its own—as often happens.

DEFINITION 4.14. If $f : A \to B$ and $X \subseteq A$ the *restriction* $f \restriction X$ of f to X is the function $X \to B$ such that $(f \restriction X)(x) = f(x)$ for all $x \in X$.

Thus the restriction of f to X is little different from f; only its domain is a subset of that of f. Restrictions are written in various ways. You might see $f \restriction X$ written as $f_{\restriction X}$, $f_{|X}$, $f|X$, for example.

EXERCISE 4.3. Let $A := \{0, 1, 2\}$ and $B := \{0, 1, 2, 3, 4\}$. Let $f : A \to B$ be defined by $f : x \mapsto x + 1$ and let $g : B \to A$ be defined by $g(x) := 0$ if x is even, $g(x) := 1$ if x is odd.

(i) How many different sets $f(X)$ are there for $X \subseteq A$?
(ii) How many different sets $g(Y)$ are there for $Y \subseteq B$?
(iii) How many different sets $f^{-1}(Y)$ are there for $Y \subseteq B$?
(iv) How many different sets $g^{-1}(X)$ are there for $X \subseteq A$?

EXERCISE 4.4. Define $f : \mathbb{R} \to \mathbb{R}$ by $f(x) := \sin x$ for all real numbers x.

(i) What are $f([0, \pi])$, $f([0, 2\pi])$, $f([0, 3\pi])$?
(ii) What are $f^{-1}(\{0\})$, $f^{-1}(\{1\})$, $f^{-1}(\{2\})$?
(iii) Let $A := [0, \pi]$ and $B := [2\pi, 3\pi]$. Show that $f(A \cap B \neq f(A) \cap f(B)$.
(iv) Let $A := [0, \pi]$. Find $f(A)$, $f^{-1}(f(A))$, and $f(f^{-1}(A))$.

EXERCISE 4.5. Let X, Y be sets and $f : X \to Y$, let $A, B \subseteq X$ and $C,\ D \subseteq Y$. Prove that:

(i) $f(A) \cup f(B) = f(A \cup B)$;
(ii) $f^{-1}(C) \cup f^{-1}(D) = f^{-1}(C \cup D)$;
(iii) $f^{-1}(C) \cap f^{-1}(D) = f^{-1}(C \cap D)$.

EXERCISE 4.6. Let $f : A \to B$, where A, B are sets.

Show that if $X \subseteq A$ then $X \subseteq f^{-1}(f(X))$. Must equality hold?

Is it always true that if $Y \subseteq B$ then $Y \subseteq f(f^{-1}(Y))$?

DEFINITION 4.15. A function $f : A \to B$ is said to be

(1) one-to-one, or *injective*, or an injection if whenever $a_0, a_1 \in A$ and $a_0 \neq a_1$ also $f(a_0) \neq f(a_1)$; equivalently, $|f^{-1}(\{b\})| = 1$ for every $b \in f(A)$;
(2) onto, or *surjective*, or a surjection if for every $b \in B$ there exists $a \in A$ such that $f(a) = b$; equivalently, $f(A) = B$; or equivalently, $f^{-1}(\{b\}) \neq \emptyset$ for every $b \in B$.
(3) one-to-one and onto, or *bijective*, or a bijection, if and only if it is both injective and surjective.

EXAMPLES 4.16. Our examples will be functions from $\mathbb{R}$ to $\mathbb{R}$.

(1) The function $f : x \mapsto x^2$ is not injective because $f(1) = f(-1)$ while $1 \neq -1$. It is not surjective either because there is no real number x for which $f(x) = -1$ (so -1 is not in the range of f).
(2) The function $g : x \mapsto \mathrm{e}^x$ is one-to-one. It is not surjective, however, because again -1 is not in the range of g.
(3) The function $h : x \mapsto x^3 - x$ is onto (can you see why?). However it is not one-to-one, because, for example, $h(0) = h(1)$, whereas of course $0 \neq 1$.
(4) The function $k : x \mapsto x^3$ is both one-to-one and onto, so it is a bijection.

EXERCISE 4.7. Let $A := \{0, 1, 2\}$ and $B := \{0, 1, 2, 3, 4\}$.

(i) How many $f : A \to B$ are there?
(ii) How many $f : B \to A$ are there?
(iii) How many injective $f : A \to B$ are there?
(iv) How many injective $f : B \to A$ are there?
(v) How many surjective $f : A \to B$ are there?
(vi) How many surjective $f : B \to A$ are there?

4.3 Algebra of functions

There is a very important way in which functions can be combined.

DEFINITION 4.17. If $f : A \to B$ and $g : B \to C$ then the *composition* of f and g, written $g \circ f$, is the function $A \to C$ defined by the equation $(g \circ f)(a) := g(f(a))$ for every $a \in A$.

Composition is familiar in calculus as 'function of a function'.

EXAMPLES 4.18. Consider functions $f, g : \mathbb{R} \to \mathbb{R}$: if $f(x) := x^2$ and $g(x) := \cos(x)$ then $(g \circ f)(x) = \cos(x^2)$, while $(f \circ g) = (\cos x)^2$ (more usually written $\cos^2 x$); if $f(x) := x^6$ and $g(x) := \mathrm{e}^x$ then $(g \circ f)(x) = \mathrm{e}^{x^6}$, while $(f \circ g)(x) = (\mathrm{e}^x)^6 = \mathrm{e}^{6x}$.

Notice that if $f : A \to B$ and $g : B \to C$ then both $g \circ f$ and $f \circ g$ are defined only if $C = A$. These examples show that it can very well happen that then $g \circ f \neq f \circ g$. Indeed, generally speaking, it is very rare that equality holds. That is to say, composition of functions is not, in general, *commutative*.

THEOREM 4.19. *Let A, B, C be sets, $f : A \to B$, $g : B \to C$.*

(1) *If f, g are injective then $g \circ f$ is injective.*

(2) *If f, g are surjective then $g \circ f$ is surjective.*

(3) *If f, g are bijective then $g \circ f$ is bijective.*

Proof. Clearly, (3) is the conjunction of (1) and (2), so it is only these that need to be demonstrated. We show (1) and leave (2) as an exercise.

Suppose that both f and g are injective. Let a_0, $a_1 \in A$ and suppose that $(g \circ f)(a_0) = (g \circ f)(a_1)$. This means that $g(f(a_0)) = g(f(a_1))$. Since g is injective it must be the case that $f(a_0) = f(a_1)$. And now, since f is injective, also $a_0 = a_1$. Therefore $g \circ f$ is injective.

EXERCISE 4.8. Write out a proof that if A, B, C are sets and both $f : A \to B$, $g : B \to C$ are surjective then $g \circ f$ is surjective.

EXERCISE 4.9. Let $f : A \to B$ and $g : B \to C$, where A, B, C are sets. Prove that if $g \circ f$ is injective then f is injective. Prove also that if $g \circ f$ is surjective then g is surjective.

DEFINITION 4.20. The identity function on a set A is the function $A \to A$ defined by $a \mapsto a$ for all $a \in A$. It is denoted 1_A (or just 1 when no ambiguity threatens) or id_A.

OBSERVATION 4.21. If A, B are sets and $f : A \to B$ then $1_B \circ f = f$ and $f \circ 1_A = f$. In particular, for any set A and any function $f : A \to A$, $1_A \circ f = f \circ 1_A = f$.

Although the operation of composition of functions is not usually commutative it is what is called *associative*. Indeed, this is one of the reasons why the associative law (which you will come across many times very soon) is so very important in mathematics.

THEOREM 4.22 **[Composition of functions is associative]**. *Let $f : A \to B$, $g : B \to C$, and $h : C \to D$ where A, B, C, D are any sets. Then*

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

Proof. For any $a \in A$, let $b := f(a) \in B$, $c := g(b) \in C$, and $d := h(c) \in D$. Then $(g \circ f)(a) = g(f(a)) = g(b) = c$, and so $(h \circ (g \circ f))(a) = h((g \circ f)(a)) = h(c) = d$. Also, $(h \circ g)(b) = h(g(b)) = h(c) = d$, whence $((h \circ g) \circ f)(a) = (h \circ g)(f(a)) = (h \circ g)(b) = d$. Thus $(h \circ (g \circ f))(a) = ((h \circ g) \circ f)(a)$ for every $a \in A$, that is, $h \circ (g \circ f) = (h \circ g) \circ f$, as required.

OBSERVATION 4.23 *Let A, B be sets, $f : A \to B$ a function. If $g, h : B \to A$ are such that $g \circ f = h \circ f = 1_A$ and $f \circ g = f \circ h = 1_B$ then $g = h$.*

Proof. For, then $g = g \circ 1_B = g \circ (f \circ h) = (g \circ f) \circ h = 1_A \circ h = h$.

DEFINITION 4.24. A function $f : A \to B$ is said to be *invertible* if there exists a function $g : B \to A$ such that $g \circ f = 1_A$ and $f \circ g = 1_B$.

By Observation 4.23, g is then unique. It is called the *inverse* of f and we write $g = f^{-1}$. Note that, directly from this definition, g is also invertible and $g^{-1} = f$, that is, $(f^{-1})^{-1} = f$.

THEOREM 4.25. *Let A, B, C be sets, $f : A \to B$, $g : B \to C$. If f, g are invertible then $g \circ f$ is invertible and $(g \circ f)^{-1} = f^{-1} \circ g^{-1}$.*

Proof. For, using associativity several times, together with the definition of inverses, we see that

$$\begin{aligned}(f^{-1} \circ g^{-1}) \circ (g \circ f) &= ((f^{-1} \circ g^{-1}) \circ g) \circ f \\ &= (f^{-1} \circ (g^{-1} \circ g)) \circ f \\ &= (f^{-1} \circ 1_B) \circ f = f^{-1} \circ f = 1_A,\end{aligned}$$

and similarly $(g \circ f) \circ (f^{-1} \circ g^{-1}) = 1_B$. Therefore $g \circ f$ is invertible and its inverse is $f^{-1} \circ g^{-1}$, as claimed.

The following is an important and useful criterion for invertibility.

THEOREM 4.26. *A function $f : A \to B$ is invertible if and only if it is bijective.*

Proof. Suppose first that $f : A \to B$ is invertible. If $f(a_0) = f(a_1)$, then $f^{-1}(f(a_0)) = f^{-1}(f(a_1))$; that is, $(f^{-1} \circ f)(a_0) = (f^{-1} \circ f)(a_1)$; so $1_A(a_0) = 1_A(a_1)$, which means that $a_0 = a_1$. Therefore f is injective (one-to-one). Also f is surjective (onto), because if $b \in B$, then $f(f^{-1}(b)) = (f \circ f^{-1})(b) = 1_B(b) = b$, so $f^{-1}(b)$ is a member of A whose image under f is b.

Now suppose that $f : A \to B$ is bijective. Define $g : B \to A$ by the rule that $g(b) := a$ if $f(a) = b$. We must show that g is well-defined. If $b \in B$ then, because f is surjective (onto), there exists $a \in A$ such that $f(a) = b$, so there do exist candidates for $g(b)$. Now if $f(a) = b$ and also $f(a') = b$ (where of course $a, a' \in A$) then $f(a) = f(a')$ and so, since f is injective (one-to-one), $a = a'$, which means that there is a unique possibility for $g(b)$. So g is well-defined. For $a \in A$, if $b := f(a)$, then by definition

$g(b) = a$, that is $g(f(a)) = a$ or $(g \circ f)(a) = a$: thus $g \circ f = 1_A$. Similarly, if $b \in B$ and $a := g(b)$ then by definition of g, it must be the case that $f(a) = b$,whence $f(g(b)) = b$: thus $(f \circ g)(b) = b$ and since this is true for every $b \in B$, $f \circ g = 1_B$. Therefore f is invertible (and $g = f^{-1}$), as required.

There are important 'one-sided' analogues of Theorem 4.26.

DEFINITION 4.27. Let A, B be sets. A function $f : A \to B$ is said to be *left invertible* if there exists $g : B \to A$ such that $g \circ f = 1_A$. Then g is called a *left inverse* of f.

Similarly, $f : A \to B$ is said to be *right invertible* if there exists $h : B \to A$ such that $f \circ h = 1_B$, and then h is called a *right inverse* of f.

THEOREM 4.28. *Let A, B be sets, and suppose that $A \neq \emptyset$.*

(1) *A function $f : A \to B$ is left invertible if and only if it is injective.*

(2) *A function $f : A \to B$ is right invertible if and only if it is surjective.*

Proof. Although it is very similar to the proof of Theorem 4.26, we show why (1) is true. We leave you to write a proof of (2).

Suppose that $f : A \to B$ is left invertible, and let $g : B \to A$ be a left inverse. If $a_0, a_1 \in A$ and $a_0 \neq a_1$ then $1_A(a_0) \neq 1_A(a_1)$, so $(g \circ f)(a_0) \neq (g \circ f)(a_1)$, that is $g(f(a_0)) \neq g(f(a_1))$, and so, since g is a function, $f(a_0) \neq f(a_1)$. Therefore f is injective.

Now suppose that f is injective. Since $A \neq \emptyset$ we may choose $z \in A$. Define $g : B \to A$ as follows:

$$g(b) := \begin{cases} a & \text{if } f(a) = b, \\ z & \text{if } b \notin f(A). \end{cases}$$

Given $b \in B$ either $b \in f(A)$ or $b \notin f(A)$. In the former case there exists $a \in A$ with $f(a) = b$ and this element a is unique because of the injectivity of f, and so it is legitimate to define $g(b) := a$. Thus our prescription yields a well-defined function $g : B \to A$. And now for any $a \in A$, $g(f(a)) = a$ by definition of g, that is $g \circ f = 1_A$. Hence f is left invertible.

EXERCISE 4.10. Prove part (2) of the theorem. That is, prove that a function $f : A \to B$ is right invertible if and only if it is surjective. [Note that for this we do not, in fact, need that $A \neq \emptyset$.]

5 Writing mathematics

Mathematics is notorious for having a language of its own. Why? Well, there are many answers. We deal with concepts such as numbers, sets, relations, functions, that are subtly different from their counterparts in ordinary discourse. They are different in that their definitions have been carefully formulated, and the words have acquired precise technical meanings within mathematics. To be acceptable, the reasoning we employ about these objects also has to be very precise.

5.1 Errors to avoid

In everyday life we use methods of reasoning that might be wrong; in ordinary life, uncertain knowledge can be better than no knowledge at all. Here are some examples of how mathematical language and reasoning can differ from what people are used to.

"THEOREM". *All odd numbers greater than* 1 *are prime.*

"*Proof*." 3 is prime, 5 is prime, 7 is prime, etc.

In mathematics, we never make a claim about all members of some set, unless either we examine every single one (which is practical only if the set is finite and smallish), or we have some method that works equally well for all members of the set. Compare the above with the following argument that all primes greater than 2 are odd.

> If $n > 2$ and n is even then 2 is a proper divisor of n, so n is not prime.
> Therefore if $n > 2$ and n **is** prime then n cannot be even, hence n is odd.

The form of generalisation in the absurd "proof" above often works for us in real life (to make inferences about all wolves, all electrons, and the like from a limited sample), but it is illegitimate in mathematics.

"THEOREM". $0 = 1$.

"*Proof*". If $a = 2$, then $a^2 = 4$. Now let $x := 0$. Then $(4x - 2)^2 = 4$. Therefore $4x - 2 = 2$, so $x = 1$. That is, $0 = 1$.

This argument contains a slightly hidden step looking like this: if P then Q; Q; therefore P (with P being the statement '$a = 2$', and Q the statement '$a^2 = 4$'). This rule is of course very doubtful in any situation, but in ordinary life it might allow us to guess that someone is in a room because we've heard their voice even although it might be a recording (if Elizabeth is in that room then that is the voice I would hear; that is the voice I have heard; therefore Elizabeth is in that room), or that allows Sherlock Holmes to deduce that horses have been past from their hoofprints (they could be a fake). In mathematics such reasoning is not allowed—we want certainty.

Some ways of reasoning are wrong under any circumstances. 'Begging the question' means assuming what you set out to prove, as in the following example:

"Theorem". $2 + 2 = 5$.

"*Proof*". If $2 + 2 = 5$, then, cancelling 2 from both sides, $2 = 3$. Subtract 5 to get that $-3 = -2$ and then subtract 2: $-5 = -2 - 2$. Now multiply by -1 to see that $5 = 2 + 2$, that is, $2 + 2 = 5$, as required. Hm!

Another rule we use in everyday life is what we might call deference to experts; if someone is an expert in a particular area, then we (often) believe what they say just for that reason. Use this rule only with great care! Do not automatically believe what you read in books. And if your lecturers say something strange then they may have made a mistake; think critically and if you are right then please correct them (politely).

Exercise 5.1. Explain what is wrong with the following (distressingly common) attempted proof of the AM-GM Inequality for two non-negative real numbers, and show how to turn it into a correct argument.

> We are given that $a \geqslant 0$, $b \geqslant 0$. Suppose that $\sqrt{ab} \leqslant \frac{1}{2}(a + b)$. Then $2\sqrt{ab} \leqslant a + b$, so $4ab \leqslant (a + b)^2$. Expanding and tidying we see that $0 \leqslant a^2 - 2ab + b^2$, that is, $0 \leqslant (a - b)^2$. Since squares are always positive this is true. Therefore $\sqrt{ab} \leqslant \frac{1}{2}(a + b)$, which is the AM-GM Inequality.

5.2 The language of mathematical reasoning

Among the most important words in a mathematical argument are the logical words. They need to be used carefully. Most of them hold no surprises, but some have meanings that are a little different in mathematics from their common meanings in everyday life.

If, only if. In ordinary discourse the word 'if' usually carries an implication that may have something to do with causation or necessity. For example, when we say 'if you throw a stone at a window, then the glass will break', then we are not merely making a prediction, we are implying that the stone will cause the glass to break. Such implications are absent in mathematics. In mathematics, 'if P then Q' (where P and Q are assertions) simply means that whenever P holds, Q does too; equivalently, that either P is false, or Q is true. So the statement 'if P then Q' is not held to assert any other connection, causal or otherwise, between P and Q. Thus, for example, both 'If Paris is the capital of France then the Thames flows through London' and 'If Oxford is on Mars then I am 100 metres tall' are true statements.

The following all mean the same:

(1) if P then Q;
(2) P implies Q;
(3) P only if Q;
(4) P is a sufficient condition for Q;
(5) Q is a necessary condition for P;
(6) if Q does not hold then P does not hold.
(7) whenever P holds, Q also holds.

In order to prove a statement of the form 'if P then Q', one typically starts by assuming that P holds and one tries to derive Q, or one starts by assuming that Q does not hold and tries to derive that then also P must not be true. We'll return to this point later.

Notice that 'if P then Q' and 'if not Q then not P' are different ways of saying the same thing. After all, if P is false whenever Q is false, then when P is true Q must necessarily be true too. The assertion 'if not Q then not P' is known as the *contrapositive* of 'if P then Q'. We'll return to this point later too.

Note that the contrapositive is very different from the converse. The *converse* of 'if P then Q' is 'if Q then P'. The former can very well be true without the latter being true. For example 'if Cambridge is on the moon then Oxford is in England' is true because Cambridge is not on the moon, but its converse 'if Oxford is in England then Cambridge is on the moon' is false since here our assertion P is true whereas our Q is false. Again, it is true that if $1 = 0$ then $1 < 2$, but it is not true that if $1 < 2$ then $1 = 0$.

The symbol $\Rightarrow$ is used to mean 'implies'. We use it **only** in formulae and only when an implication is the mathematical statement we wish to make. Thus for example to say that a relation R on a set A is symmetric (Definition 4.3) is to say that $a\,R\,b \Rightarrow b\,R\,a$ whenever $a, b \in A$; the definition of transitivity could be written ($a\,R\,b$ and $b\,R\,c$) $\Rightarrow a\,R\,c$ for all $a, b, c \in A$.

Warning: **never** misuse $\Rightarrow$ to mean 'then' (as in 'if $x = -1 \Rightarrow x^2 = 1$'). **Never** misuse $\Rightarrow$ to mean 'therefore' (as in '$(\sqrt{a} - \sqrt{b})^2 \geqslant 0 \Rightarrow a + b - 2\sqrt{ab} \geqslant 0 \Rightarrow \sqrt{ab} \leqslant \frac{1}{2}(a + b)$'). We *always* read $P \Rightarrow Q$ as 'if P then Q' or as 'P implies Q'.

If and only if. The statement 'P if and only if Q' means 'if P then Q AND if Q then P'. This can be rephrased 'P and Q are equivalent'. Usually one proves such a statement by proving 'if P then Q' and 'if Q then P' separately. The phrase 'P is a necessary and sufficient condition for Q' means exactly the same thing.

You'll find that some (many) people use 'iff' as an abbreviation for 'if and only if'. Personally, I dislike it. To my eye it looks like a misprint—and indeed, it is often misprinted or misread.

Not, and, or. There is little to be said about the so-called *connectives* 'not' and 'and'. An assertion 'not P' will be true when P is false and false when P is true. An assertion 'P and Q' will be true when P and Q are both true, false otherwise. In ordinary discourse the word 'or' in 'one or the other' sometimes carries overtones of 'but not both'. That is never the case in mathematical usage. We always interpret 'P or Q' to mean that P holds or Q holds or both do.

Quantifiers. Quantifiers are expressions like *for all* or *for every*, which are known as *universal* quantifiers; *for some* or *there exist* (or *there exists*), known as *existential* quantifiers. Examples of statements with quantifiers:

- every prime number greater than 2 is odd;
- for every natural number n, either n is a perfect square, or $\sqrt{n}$ is irrational;
- there exists a real number x such that $x^3 - 103x^2 + 2 = 0$;
- some prime numbers have two decimal digits.

Note that a quantifier includes specification of a range: all prime numbers, some real number(s), or whatever. We have symbols $\forall$, $\exists$ for use in formulae (and never be used as lazy abbreviations for words in text). Thus, for example, if we use $\mathbb{P}$ to denote the subset of $\mathbb{N}$ consisting of prime numbers then these statements could be formulated as:

- $\forall p \in \mathbb{P}$: if $p > 2$ then p is odd;

- $\forall n \in \mathbb{N} : (\exists m \in \mathbb{N} : m = n^2)$ or $(\sqrt{n} \notin \mathbb{Q})$;
- $\exists x \in \mathbb{R} : x^3 - 103x^2 + 2 = 0$;
- $\exists p \in \mathbb{P} : 10 \leqslant p < 100$.

EXERCISE 5.2. Which of the following statements about natural numbers are true?

(a) 2 is prime or 2 is odd.
(b) 2 is prime or 2 is even.
(c) If 2 is odd then 2 is prime.
(d) If 2 is even then 2 is prime.
(e) For all $n \in \mathbb{N}$, if n is a square number then n is not prime.
(f) For all $n \in \mathbb{N}$, n is not prime if and only if n is a square number.
(g) For all even primes $p > 2$, $p^2 = 2012$.

The mathematical meanings of quantifiers can be a little different from what they are in ordinary English. When we say 'for all positive real numbers x, there exists a real number y such that $x = y^2$', the meaning—that every positive real number has a real square root—is completely clear. Perhaps slightly less clear is that the assertion 'all even primes p greater than 3 have exactly nine digits' is true. There are no even primes p greater than 3, so all of them do have exactly nine digits. The statement is, as we say, *vacuously* true. So is 'all members of $\emptyset$ are infinite', which may look paradoxical as an English sentence, but happens to be true. In ordinary language, when I say that there exist people who live in France, I assert that there is at least one person who lives in France, but I also suggest that there are some people who do not. Such suggestions are absent in mathematics. Thus, a statement $\exists x \in \mathbb{R} : P(x)$ means precisely that there is at least one real number that has the property P. It means neither more nor less.

Warning: **never get quantifiers in the wrong order**. Consider the statement: for every house H, there exists A such that A is the address of H. That is a ponderous way of saying that every house has an address. Now consider the statement: there exists A such that for every house H, A is the address of H. What does this statement mean? Well, if it is true, then there is an address A—it might be 10 Downing Street for example—which has the remarkable property that for every house H, the address of H is 10 Downing Street. The statement means, in fact, that every house has the same address. Thus if $\mathcal{H}$ is the set of all houses and $\mathcal{A}$ the set of all addresses then

$$\forall H \in \mathcal{H} : \exists A \in \mathcal{A} : A = \text{address}(H) \quad \text{and} \quad \exists A \in \mathcal{A} : \forall H \in \mathcal{H} : A = \text{address}(H)$$

say very different things. The former is true the latter false. The order of quantifiers really matters. Great care is needed because English can be ambiguous. For example, what does the following statement mean?

For all natural numbers x, $x < y$ for some natural number y.

Does it mean

for all $x \in \mathbb{N}$, there exists $y \in \mathbb{N}$ such that $x < y$? $\qquad (\star)$

or does it mean

there exists $y \in \mathbb{N}$ such that for all $x \in \mathbb{N}$, $x < y$? $\qquad (\star\star)$

Of these $(\star)$ is true since y could be $x + 1$ for example, while $(\star\star)$ is false because no matter how big y is there is some x which is bigger. In ordinary language, it is often unclear what logical order quantifiers are supposed to come in; we rely on context and common sense to guess intelligently (and our guesswork is so intelligent that we usually do not notice that there could be a problem). Mathematics, however, is unforgiving. Sloppiness with quantifiers is inexcusable.

EXERCISE 5.3. Formulate Theorem 2.2, Mathematical Induction, using the symbols $\forall$ and $\Rightarrow$.

EXERCISE 5.4. Let $f : \mathbb{R} \to \mathbb{R}$. Translate the following formula into English:

$$\forall a \in \mathbb{R} : \forall \varepsilon \in \mathbb{R}^{>0} : \exists \delta \in \mathbb{R}^{>0} : \forall x \in \mathbb{R} : |x - a| < \delta \Rightarrow |f(x) - f(a)| < \varepsilon\,.$$

EXERCISE 5.5. Let $f : \mathbb{R} \to \mathbb{R}$. Express the following using the symbols $\forall$ and $\Rightarrow$:

> for all real numbers a and x and for every positive real number ε there exists a positive real number δ such that $|f(x) - f(a)| < \varepsilon$ whenever $|x - a| < \delta$.

Does it say much the same as the formula in the preceding exercise?

5.3 Handling negation

Let us return briefly to the word 'not'. Given an assertion P the assertion *not* P is true if P is false and it is false if P is true. That is, *not* P holds if and only if P does not. The following rules, in which $\Leftrightarrow$ is used as a symbol for *if and only if* or *is equivalent to*, are basic.

THEOREM 5.1 **[Some basic rules for negation]**. *Let P, Q be propositions, that is, assertions, perhaps about a member x of a set X. Then*

(1) $\mathit{not}\,(\mathit{not}\,P) \Leftrightarrow P$;

(2) $\mathit{not}\,(P \ \mathit{and}\ Q) \Leftrightarrow (\mathit{not}\,P)\ \mathit{or}\ (\mathit{not}\,Q)$;

(3) $\mathit{not}\,(P \ \mathit{or}\ Q) \Leftrightarrow (\mathit{not}\,P)\ \mathit{and}\ (\mathit{not}\,Q)$;

(4) $\mathit{not}\,\forall x \in X : P(x) \Leftrightarrow \exists x \in X : \mathit{not}\,P(x)$;

(5) $\mathit{not}\,\exists x \in X : P(x) \Leftrightarrow \forall x \in X : \mathit{not}\,P(x)$.

Where do these come from? Well, (1) should be clear. As for (2), the conjunction P *and* Q is false if and only if it is not the case that both P and Q hold, that is to say, one of $\mathit{not}\,P$ and $\mathit{not}\,Q$ must be true, so $\mathit{not}\,P$ *or* $\mathit{not}\,Q$ holds. The justification for (3) is similar. What (4) is saying is that if it is not the case that $P(x)$ holds for every $x \in X$ then at least one $x \in X$ fails to satisfy P, and of course vice versa. And what (5) says is that if there are no members of the set X for which $P(x)$ holds then every member of X fails to satisfy the condition P, and vice versa.

EXAMPLES 5.2. Let $f : \mathbb{R} \to \mathbb{R}$ and let $a \in \mathbb{R}$. The two assertions

$$\forall \varepsilon \in \mathbb{R}^{>0}\ \exists \delta \in \mathbb{R}^{>0}\ \forall x \in \mathbb{R} : \text{if } |a - x| < \delta \text{ then } |f(a) - f(x)| < \epsilon,$$

$$\exists \varepsilon \in \mathbb{R}^{>0}\ \forall \delta \in \mathbb{R}^{>0}\ \exists x \in \mathbb{R} : |a - x| < \delta \text{ but } |f(a) - f(x)| \geqslant \epsilon$$

are negations of each other.

For, by (4) and (5) of Theorem 5.1, we can move *not* past a quantifier provided that we change $\forall$ to $\exists$ and vice versa. Thus an assertion of the form $\mathit{not}\,\forall\exists\forall : P$ is the same as $\forall\exists\forall : \mathit{not}\,P$. In the example P is of the form $Q \Rightarrow R$ and the negation of this (which must be true if and only if $Q \Rightarrow R$ is false) is Q *but not* R because $Q \Rightarrow R$ is equivalent to *not* Q *or* R. (Notice that 'but' is another form of 'and', though in English it carries overtones of negative expectations.)

NOTE: It is common to write 'for every $\varepsilon > 0$' as an abbreviation for 'for every real number $\varepsilon > 0$'. You will often see $\forall\varepsilon > 0$ standing for $\forall\varepsilon \in \mathbb{R}^{>0}$. Thus the assertions in Example 5.2 would often be written

$$\forall\varepsilon > 0\ \exists\delta > 0\ \forall x \in \mathbb{R} : |a - x| < \delta \Rightarrow |f(a) - f(x)| < \epsilon,$$
$$\exists\varepsilon > 0\ \forall\delta > 0\ \exists x \in \mathbb{R} : |a - x| < \delta \text{ but } |f(a) - f(x)| \geqslant \epsilon.$$

EXERCISE 5.6. Let $f : \mathbb{R} \to \mathbb{R}$. Organise the following assertions about the function f into pairs such that one member of the pair is true if and only if the other is false:

(1) for all real numbers x, y, there exists a real number $z > x$ such that $f(z) > y$;
(2) for every real x there exists a real number y such that for all real z either $z \leqslant y$ or $f(z) \neq x$;
(3) there exist real numbers x, y such that for every real number z if $z \leqslant y$ then $f(z) = x$;
(4) there is a real number x such that for every real y there is a real number $z > y$ for which $f(z) = x$;
(5) for all real x and y there is a real number z such that $z \leqslant y$ and $f(z) \neq x$.
(6) There exist real numbers x and y such that for every real z either $z \leqslant x$ or $f(z) \leqslant y$.

5.4 Formulation of mathematical statements

It is important to understand correctly the logical form of a theorem or a problem. The most common form is 'If P then Q' though there are a number of variations on the actual words we use. In this context, P is the *hypothesis* and Q the *conclusion*. In the following examples, the hypothesis is introduced with the symbol $\triangleleft$, the conclusion with $\triangleright$.

EXAMPLES 5.3.

THEOREM A. $\triangleleft$ *Suppose that the polynomial $p(x)$ with real coefficients has odd degree.* $\triangleright$ *Then $p(x)$ has a real root.*

THEOREM B. $\triangleleft$ *If n is a natural number* $\triangleright$ *then n has a unique prime factorisation.*

THEOREM C. $\triangleleft$ *Whenever f is a continuous function on $\mathbb{R}$, and a, b are real numbers such that $a < b$, $f(a) < 0$ and $f(b) > 0$,* $\triangleright$ *there exists a real number $c \in (a, b)$ such that $f(c) = 0$.*

Note that hypothesis and conclusion are not always quite so clearly visible. For example, Theorem A could have been put in the form

Every polynomial with real coefficients and odd degree has a real root.

It is, of course, important to interpret statements of theorems correctly. It is also important to write down your own theorems clearly, so that someone else can easily work out what the hypothesis is, and what is the conclusion. For example, the formulation of Theorem C above is not particularly reader-friendly. Hypothesis and conclusion could be exhibited more clearly, for example by breaking the one long sentence into two or more:

> *Let f be a continuous function on $\mathbb{R}$, and let a, b be real numbers such that $a < b$, $f(a) < 0$ and $f(b) > 0$. Then there exists a real number $c \in (a, b)$ such that $f(c) = 0$.*

Here is a much worse example:

> THEOREM D. *Whenever $f : [0, 1] \to \mathbb{R}$ is a continuous function, f is differentiable and $f(0) = f(1)$, f attains a greatest and a least value, and there exists $c \in (0, 1)$ such that $f'(c) = 0$.*

It seems to start by saying that every continuous function from $f : [0, 1] \to \mathbb{R}$ is differentiable and satisfies $f(0) = f(1)$, but that is nonsense. Don't write like this: instead be clear and orderly.

> EXERCISE 5.7. Reformulate Theorem D to make completely clear what is meant.

6 Proofs and refutations

Proofs in mathematics have to stand up to rigorous examination. They need to be completely logical; they need to be capable of being thoroughly checked. Ideally, though, they should be more than that. They should help the intuition to understand what lies behind a theorem, what its context is and what it 'means'. Thus a proof should have a clear structure. Let's examine some of the possibilities.

6.1 Direct proof

The concept of direct proof is very simple: to prove a statement of the form *if P then Q* we start from P and seek to reach Q by legitimate reasoning. Here's an example—one possible response to Exercise 5.1 above.

THEOREM 6.1. *Let a, b be non-negative real numbers. Then $\sqrt{ab} \leqslant \frac{1}{2}(a+b)$.*

Proof. Being non-negative, a, b have non-negative real square roots. All real squares are non-negative, so $(\sqrt{a} - \sqrt{b})^2 \geqslant 0$ (moreover, equality holds if and only if $a = b$). Expanding, we see that $a - 2\sqrt{a}\sqrt{b} + b \geqslant 0$. Rearranging this we get that $\sqrt{ab} \leqslant \frac{1}{2}(a+b)$, as required. Moreover, we see that equality holds if and only if $a = b$.

Notice that this is a good (if particularly simple) example of a direct proof. It starts from the assertion P that a, b are non-negative real numbers and moves forward to the conclusion Q which is the AM-GM Inequality. Notice also that it gives a little more information than is in the statement of the theorem.

EXERCISE 6.1. Use the Binomial Theorem to prove that if n is any natural number then $n < 2^n$.

[Although this lends itself to proof by induction, the exercise is to give a simple direct proof.]

EXERCISE 6.2. Use the Binomial Theorem to prove that if a is any positive real number and k is any natural number then there exists $N \in \mathbb{N}$ such that $n^k < (1+a)^n$ for all natural numbers n such that $n > N$. In symbols the theorem is $\forall a > 0\ \forall k \in \mathbb{N}\ \exists N \in \mathbb{N}\ \forall n \in \mathbb{N}$: if $n > N$ then $n^k < (1+a)^n$.

6.2 Proof by contradiction

Since the contrapositive *if not Q then not P* is equivalent to *if P then Q* one can prove the latter by giving a direct proof of the former. This is known as *proof by contradiction* or, in older books, as *reductio ad absurdum*, reduction to an absurdity. Here is a simple example—another possible response to Exercise 5.1 above.

THEOREM 6.2. *Let a, b be non-negative real numbers. Then $\sqrt{ab} \leqslant \frac{1}{2}(a+b)$.*

Proof. Suppose, seeking a contradiction, that $\sqrt{ab} > \frac{1}{2}(a+b)$. Then $2\sqrt{ab} > a+b$ and so $4ab > (a+b)^2$, that is, $4ab > a^2 + 2ab + b^2$. Subtract $4ab$ from both sides: $0 > a^2 - 2ab + b^2 = (a-b)^2$. This is a contradiction since squares of real numbers are non-negative. Therefore the assumption must be wrong, that is, $\sqrt{ab} \leqslant \frac{1}{2}(a+b)$, as required.

There are two criticisms to be made of this argument. One is that it hides the need for the assumption that both a and b are non-negative. Where does that enter? The other that it does not quite so easily tell us the condition for equality. This example illustrates two matters of style: first, always indicate early and explicitly that you are using the technique; second, that if you find a proof by contradiction, it is always worth thinking whether you might turn the ideas round and derive a simple direct proof. After all, a long proof by contradiction can be a little unconvincing: the final contradiction should have arisen because the assumption *not Q* is untenable, but it might have arisen from an error in the argument. Take a look at the famous and very important paper 'Solvability of groups of odd order' by WALTER FEIT and JOHN G. THOMPSON in *Pacific Journal of Mathematics*, 13 (1963), pp. 775–1029. It is a huge proof by contradiction which concludes on its 253rd page of exceptionally technical and intricate and clever argument with the words 'This contradicts Lemma 38.11, and completes the proof of the main theorem of this paper.' How sure can we be that the contradiction is not the consequence of an error made somewhere around the 200th page? As it happens the proof is correct. And we know of no way other than use of contradiction to prove the Feit–Thompson Theorem. Even so, for less complex theorems it is a technique to be used with great care. It is, however, a powerful technique. Here is another example.

THEOREM 6.3. *Let X be any set. No functions $f : X \to \wp X$ are surjective.*

Proof. Suppose, seeking a contradiction, that there does exist a surjective function $f : X \to \wp X$. For each $x \in X$ the image $f(x)$ is a subset of X and we can ask whether or not x lies in it. We focus on those x that do not lie in their image, and define $A := \{x \in X \mid x \notin f(x)\}$. Then $A \subseteq X$, that is, $A \in \wp X$. Since f is surjective there exists $a \in X$ such that $A = f(a)$. Now either $a \in A$ or $a \notin A$. If $a \in A$ then $a \in f(a)$, but, by definition of A, $a \notin f(a)$. And if $a \notin A$, so $a \notin f(a)$, then, by definition of A, $a \in A$. This is a contradiction. Therefore (unless there is a slip in the reasoning) the assumption must be wrong, that is, there is no surjective function $X \to \wp X$.

The above theorem and proof are a version of Cantor's argument that for any set, whether finite or infinite, the cardinal number (that is, the size) of X is strictly smaller than the cardinal number of its power set $\wp X$. This is of great importance in mathematics. It is also of great importance historically—it led directly to Cantor's Paradox, that on the one hand there cannot be a largest set, but on the other hand the set of all sets should obviously be a largest set, and it led nearly as directly to Russell's Paradox focussing on the set of all sets that are not members of themselves. These paradoxes led in their time to an understanding that Set Theory needed to be formulated considerably more carefully than I have done in my naive description of it in § 3. That is why Set Theory proper is an advanced topic offered in the third and fourth years of the Mathematics course (though in the second year of Mathematics & Philosophy). The naive version, however, serves adequately as a good language for mathematics—just so long as we do not try to push it too far.

Here are two classical examples where proof by contradiction is used.

EXERCISE 6.3. Let $n \in \mathbb{N}$. Show that either n is a perfect square (that is, there exists $m \in \mathbb{N}$ such that $n = m^2$), or $\sqrt{n} \notin \mathbb{Q}$.

EXERCISE 6.4. Let $\mathbb{P} := \{n \in \mathbb{N} \mid n \text{ is prime}\}$, the set of prime numbers. Show that $\mathbb{P}$ is infinite.

6.3 More on Mathematical Induction

Mathematical Induction was introduced in Section 2.2 above, and a typical example of proof by induction was given there. It is so important as a tool for proving theorems that it is worth thinking of variants of it. In this subsection we treat two other versions.

THEOREM 6.4 [**Strong Induction**]. *Let P be a property of natural numbers, that is, a statement $P(x)$ about natural numbers x that may be true for some natural numbers x and false for others. Suppose that for all natural numbers n, if $P(m)$ is true for all natural numbers $m < n$ then $P(n)$ is true. Then $P(n)$ is true for all natural numbers n.*

Although this looks superficially weaker than Theorem 2.2 because its hypothesis is stronger while its conclusion is the same, the two are in fact equivalent. We can deduce it from Theorem 2.2 as follows. Suppose that P is as described. Let $Q(x)$ be the property that for all natural numbers $m < x$, $P(m)$ is true. Then $Q(0)$ is true, because there are no natural numbers $m < 0$, so $P(m)$ is true for every $m < 0$. If $Q(n)$ is true, then for all $m < n$, $P(m)$ holds. So by assumption, $P(n)$ holds also. Thus $P(m)$ holds for all $m < n + 1$, that is, $Q(n + 1)$ holds. Therefore by Mathematical Induction, $Q(n)$ holds for all natural numbers n. It follows, of course, that for all natural numbers n, $Q(n+1)$ holds; that is, for all $m < n+1$, $P(m)$ holds; in particular, $P(n)$ holds. So $P(n)$ is true for all natural numbers n.

Here is an example of how strong induction may be used.

THEOREM 6.5. *Every natural number greater than 1 may be expressed as a product of one or more prime numbers.*

Proof. Let $P(x)$ be the assertion that either $x \leqslant 1$ or x may be expressed as a product of prime numbers. Let n be a natural number and suppose that $P(m)$ holds for all $m < n$. If $n \leqslant 1$ then $P(n)$ certainly holds. If $n \geqslant 2$ then either n is prime or n is not prime. If n is prime then it is the 'product' of the single prime number n. If n is not prime then there exist $r, s > 1$ such that $n = rs$. Then $r,\ s < n$, so by the induction hypothesis, r and s may each be written as a product of prime numbers. Therefore rs is a product of prime numbers, that is, n is a product of prime numbers. Now by Strong Induction, $P(n)$ is true for all natural numbers n, that is, every natural number greater than 1 may be expressed as a product of one or more prime numbers.

Mathematical Induction may be expressed in a very different way:

THEOREM 6.6 [**Well-ordering of the natural numbers**]. *If S is any non-empty subset of $\mathbb{N}$ then S has a least member.*

We say that $\mathbb{N}$ is *well ordered*. The principle of Mathematical Induction may be proved from the fact that $\mathbb{N}$ is well ordered in the following way. Let $P(x)$ be a property of natural numbers such that $P(0)$ is true and whenever $P(n)$ is true then so also is $P(n + 1)$. Suppose, seeking a contradiction, that $P(n)$ does not hold for all natural numbers n. Define $S := \{x \in \mathbb{N} \mid P(x) \text{ is false}\}$. By assumption $S \neq \emptyset$ and so S must have a least member m. Now $m > 0$ since $P(0)$ *is* true. Therefore $m = k + 1$ for some $k \in \mathbb{N}$. Since $k < m$, $k \notin S$ and therefore $P(k)$ does hold. But then, by assumption,

$P(k+1)$ also holds, that is, $P(m)$ holds. This contradicts the fact that $m \in S$. Therefore $P(n)$ must hold for all natural numbers n.

Conversely, we may prove Theorem 6.6 by Strong Induction. Let $P(x)$ be the statement that every subset S of $\mathbb{N}$ such that $x \in S$ has a least member. Now suppose that $P(m)$ holds for every natural number $m < n$. Let S be any subset of $\mathbb{N}$ with $n \in S$. If there exists $m < n$ such that $m \in S$ then, since $P(m)$ holds, S has a least member. Otherwise n itself is the least member of S. Either way, S has a least member, so $P(n)$ holds. By Strong Induction $P(n)$ holds for every natural number n. But now, to say that S is non-empty is to say that there exists $x \in S$. Since we know that $P(x)$ holds, S must have a least member.

Thus all of induction, strong induction and well-ordering are equivalent. They are different ways of saying the same thing. Sometimes one is more convenient, sometimes another. In applications of well-ordering one usually seeks what Reinhold Baer [1902–1979] used to call 'the least criminal'. That is to say, we are trying to prove some statement involving natural numbers, and we do it by contradiction: if the statement is false then there will be a least natural number for which it is false. This he called *the least criminal*. The leastness of the criminal gave extra information about it which, suitably exploited, could give a contradiction. Indeed, this is how the proof by Feit & Thompson quoted above goes: they suppose their theorem false and they study a least criminal. That least criminal will be a non-solvable group of least possible odd order (whatever that may mean). The leastness of its order gave them a huge amount of information about it, information which led, in the end, to the killer contradiction. Here, to illustrate the idea, is an alternative proof of Theorem 6.5.

EXAMPLES 6.7. *An alternative proof that if $n \geqslant 2$ then n may be expressed as a product of prime numbers.*

Suppose the statement were false. Then there would be a least natural number $n > 1$ that is not expressible as a product of prime numbers. It cannot itself be prime, so $n = rs$ where $1 < r < n$ and $1 < s < n$. Being smaller than n, each of r, s must be expressible as a product of prime numbers, whence n is such a product after all. This contradiction proves the theorem.

6.4 Refutation

To refute: to prove a statement to be false or incorrect; to disprove.

Here is an observation. Let $f(x) := x^2 + x + 41$. Then

$$\begin{array}{lll} f(-1) = f(0) = 41, & f(-2) = f(1) = 43, & f(-3) = f(2) = 47, \\ f(-4) = f(3) = 53, & f(-5) = f(4) = 61, & \ldots, \\ f(20) = f(-21) = 461, & \ldots, & f(-31) = f(30) = 973, \quad \ldots \end{array}$$

As far as the eye can see, $f(k)$ is prime for $k \in \mathbb{Z}$. It is not unreasonable therefore to make the following

CONJECTURE: *if $f(x) := x^2 + x + 41$ then $f(k)$ takes only prime values, that is, $f(k)$ is a prime number for every $k \in \mathbb{Z}$.*

Nonsense! Although, as it happens, $f(k)$ is prime for $-40 \leqslant k \leqslant 39$, it is clear that $f(40) = 41 \times 41$, so is composite. One counterexample is enough to refute a conjecture.

7 Problem-solving in mathematics

There cannot be rules and recipes for problem-solving in mathematics. If there were then research would be routine (and perhaps not much fun). To solve problems you need to work out what they are about, think about what mathematics might be relevant, experiment, pursue ideas that might work and recognise when they show clear signs of not working, think laterally, be devious. Very roughly, the problems that one meets in an undergraduate course are either 'closed' or 'open-ended'. The former are of the form 'prove this', 'calculate that', where you might have a pretty good idea of the goal, but the problem is to find a good way to get there. An open-ended problem might be a question of the form 'is it true that ...?' or 'when is it true that ...?' or 'what can you discover about ...?' which is a bit like a piece of research.

7.1 Some techniques and examples

What techniques do we have? Well, if we are asked to prove an assertion of the form *if P then Q* then there are several possibilities. We could try making deductions from the assumption that P is true and seek to close in on Q. Or (hoping to use proof by contradiction) we could assume that Q is false, pursue a line of reasoning, and seek to discover that then P would have to be false. Or we could do both, and hope that our two lines meet somewhere in the middle. Or, if P and Q depend on a parameter, as they often will, we could experiment with special values of that parameter and see if they give us some insight. If they depend on a parameter $n \in \mathbb{N}$ we might try induction. The first steps in problem-solving are experimentation, pattern-spotting, conjecture. They are followed by refutation, in which case go back and start again, or verification, writing-up and celebration. In that order: never celebrate until you have completed writing up your solution as a fair copy.

Here is an example.

PROBLEM A. *Let $a_1, a_2, \ldots, a_n$ and $b_1, b_2, \ldots, b_n$ be real numbers and suppose that $a_1 \leqslant a_2 \leqslant \cdots \leqslant a_n$. Let $c_1, c_2, \ldots, c_n$ be the numbers $b_1, b_2, \ldots, b_n$ in some order. Prove that $a_1c_1 + a_2c_2 + \cdots + a_nc_n$ is maximised when $c_1 \leqslant c_2 \leqslant \cdots \leqslant c_n$ and is minimised when $c_1 \geqslant c_2 \geqslant \cdots \geqslant c_n$.*

It is a closed question. We know what we are aiming for and the problem is to find an appropriate line of reasoning. Where can we start? We start with rough scribbling, back-of-an-envelope experimentation. One common starting point is trying small values of n. What does 'small' mean here? Well, it might mean $n = 1$, $n = 2$, $n = 3$, possibly $n = 4$. If we have not spotted what matters and what does not matter by the time we get to $n = 5$ then we should probably try a different strategy.

Back-of-an-envelope experimentation. In this case, for $n = 1$ there is nothing to prove—the problem is non-trivial only for $n > 1$. For $n = 2$ the situation is relatively simple because either $b_1 \leqslant b_2$ or $b_2 \leqslant b_1$. Re-numbering if necessary we may suppose that $b_1 \leqslant b_2$. So now the question is, which is the larger of $a_1b_1 + a_2b_2$ and $a_1b_2 + a_2b_1$ (given that $a_1 \leqslant a_2$ and $b_1 \leqslant b_2$)? What strategy do we have for testing an inequality? Well, to discover whether or not $A < B$ we could examine $A - B$ and seek to discover whether it is positive, negative or zero; or we could try A/B and seek to discover whether it is greater or smaller than 1. Either technique would replace the two numbers A, B with just one number. Division, though, has to be used with great care—we need to

know beforehand that $B \neq 0$, and indeed, since we are working with inequalities we need to know whether or not B is positive: if so then division by B preserves inequalities, whereas if B is negative then division by B reverses inequalities. It is usually (though not always) safer and easier to use subtraction.

So let's look at the difference $D := (a_1b_1 + a_2b_2) - (a_1b_2 + a_2b_1)$. Aha! It factorises: $D = (a_1 - a_2)(b_1 - b_2)$. Since $a_1 - a_2 \leqslant 0$ and $b_1 - b_2 \leqslant 0$, in fact $D \geqslant 0$. Thus in this case $a_1c_1 + a_2c_2$ is maximised when $c_1 \leqslant c_2$ and minimised when $c_2 \leqslant c_1$, just as the examiner predicted.

Moving to the case $n = 3$ we find we are in different territory because we cannot simply assume that $b_1 \leqslant b_2 \leqslant b_3$. There are six different possibilities; or perhaps three insofar as any one of the b_i could be between the other two. But can we use the case that we have already dealt with? After all, if $c_1 > c_2$ then (as the case $n = 2$ tells us) $a_1c_1 + a_2c_2 \leqslant a_1c_2 + a_2c_1$, so interchanging c_1 and c_2 would increase (or at least, not decrease) the sum $a_1c_1 + a_2c_2 + a_3c_3$. And now we have the key. So now we write out our fair-copy answer.

Fair copy answer. Consider the ordering $c_1, c_2, \ldots, c_n$ of the numbers $b_1, b_2, \ldots, b_n$ and suppose that for some r in the range $1 \leqslant r < n$ we have $c_r > c_{r+1}$. Let $S := \sum a_ic_i$ and let T be the same sum except that c_r, c_{r+1} are interchanged. All except two of the summands of T are the same as those of S, so they cancel on subtraction, and $S - T = (a_rc_r + a_{r+1}c_{r+1}) - (a_rc_{r+1} + a_{r+1}c_r)$. Thus $S - T = (a_r - a_{r+1})(c_r - c_{r+1})$, so $S - T \leqslant 0$ since $a_r \leqslant a_{r+1}$ and $c_r > c_{r+1}$. Hence interchanging c_r and c_{r+1} would increase $\sum a_ic_i$, or at least, not decrease it. Therefore $\sum a_ic_i$ is as large as possible when $c_r \leqslant c_{r+1}$ for $1 \leqslant r < n$, that is, $c_1 \leqslant c_2 \leqslant \cdots \leqslant c_n$. The same argument shows that $\sum a_ic_i$ is as small as possible when $c_1 \geqslant c_2 \geqslant \cdots \geqslant c_n$.

Here is another example.

Problem B. *How many solutions are there of the equation $x_1+x_2+\cdots+x_m = 2012$ in which m and every x_i are positive integers?*

It is an open question in that it asks us to find a number. But however can we get started? Well, as always, we start by jotting on scrap-paper.

Back-of-an-envelope experimentation. Can we spot some solutions? Well, yes: $m = 1$, $x_1 = 2012$ is a solution; so is $m = 2012$, $x_i = 1$ for $1 \leqslant i \leqslant 2012$. Don't laugh! Boundary values, very special cases, can sometimes help us on our way. In this case, though, perhaps all they do is draw attention to the hugeness of the problem—they suggest how to find all solutions with $m = 2$ and $m = 2011$, perhaps, but give no real insight into what happens in midstream.

This has told us very little. It has told us two things, however: first that of course $m \leqslant 2012$; secondly that 2012 is a very big number. And after all, though a constant, 2012 could be varied. What about trying the same problem, but with 2012 replaced by a small number? Sometimes such variants of a problem are known as 'toy versions' of it. Let's ask for the number of solutions of the equation $x_1 + x_2 + \cdots + x_m = n$ for small values of n.

The case $n = 1$ is a bit too trivial to give us any insight. We see immediately that there is just the 1 solution, $m = 1$, $x_1 = 1$. The case $n = 2$ is hardly less trivial—there are just the 2 solutions $m = 2$, $x_1 = x_2 = 1$ and $m = 1$, $x_1 = 2$. The cases $n = 3$,

$n = 4$ are still small enough that we can enumerate all solutions by hand:

$$1+1+1,\ \ 1+2,\ \ 2+1,\ \ 3,$$

4 solutions for $n = 3$;

$$1+1+1+1,\ \ 1+1+2,\ \ 1+2+1,\ \ 2+1+1,\ \ 2+2,\ \ 1+3,\ \ 3+1,\ \ 4,$$

8 solutions for $n = 4$.

Is there a pattern here? The numbers 1, 2, 4, 8 look familiar. Would it be fair to conjecture that the number of solutions of the given equation with 2012 replaced by n is 2^{n-1}? Looking at the solutions for $n = 1, 2, 3, 4$ we see that we can get solutions for the next number up by either adding 1 to the value of the last variable or increasing m by 1 and giving to the new variable the value 1.

Perhaps try using this insight to get up to $n = 5$.

Does this work in general?

Think a bit.

Yes!

So we have reached the writing-up stage

Fair copy answer. We generalise and prove that for $n \geqslant 1$ the number of solutions to the equation $x_1 + x_2 + \cdots + x_m = n$ is 2^{n-1}. This is certainly true for $n = 1$ so we use induction. Suppose that the statement is true for n. For each equality $a_1 + a_2 + \cdots + a_m = n$ we make the two equalities $b_1 + b_2 + \cdots + b_m = n + 1$ where $b_i := a_i$ for $1 \leqslant i < m$, $b_m := a_m + 1$, and $c_1 + c_2 + \cdots + c_m + c_{m+1} = n + 1$, where $c_i := a_i$ for $1 \leqslant i \leqslant m$, $c_{m+1} := 1$. Thus each solution of the equation for n gives rise to two (obviously different) solutions for $n+1$. Does every solution for $n+1$ arise in this way? Consider the solution $d_1+d_2+\cdots+d_k = n+1$. If $d_k = 1$ then it arises from the solution $d_1 + d_2 + \cdots + d_{k-1} = n$ for n, and from no other; whereas if $d_k > 1$ then it arises from the solution $d_1 + d_2 + \cdots + d_{k-1} + (d_k - 1) = n$ for n, and from no other. Thus every solution for n gives rise to two solutions for $n + 1$, and every solution for $n + 1$ arises from exactly one solution for n. Therefore the number of solutions for $n + 1$ is exactly twice as big as the number for n, which, by inductive hypothesis is 2^{n-1}, so the number of solutions for $n+1$ is 2^n. Hence by induction, for every positive integer n, the number of solutions of the given equation is 2^{n-1}.

Returning to the problem posed we see that the number of solutions of the equation $x_1 + x_2 + \cdots + x_m = 2012$ in which m and every x_i are positive integers is 2^{2011}.

Commentary. Several points about this are worth noting. First, that the general problem turned out to be tractable, whereas the given problem, a special case of it, looked out of reach. This happens quite frequently. Here we could use induction to handle the problem with 2012 replaced by n, whereas for the problem as posed there were difficulties with the horrible hugeness of 2012.

Secondly, the question is about what are called *compositions* of an integer. A composition of n is a sequence $(a_1, a_2, \ldots, a_m)$ of positive integers adding to n. The numbers a_i are known as the 'parts' of the composition. An unordered sum $a_1 + a_2 + \cdots + a_m = n$ (where again the a_i are positive integers) is known as a *partition* of n. The number of partitions of n, known as the *partition function* of n, is a far more tricky matter.

Third, there is another way to see why the answer to Problem B is 2^{2011}. Write $2012 = (1 + 1 + 1 + \cdots + 1)$ (recall the White Queen's questioning of Alice in *Through*

the Looking Glass). In this expression the number 1 occurs 2012 times, the symbol + occurs 2011 times. Take any subset of the set of occurrences of +, and for each + in the chosen subset insert a closing bracket before it and an opening bracket after it: thus, for example, $2012 = (1+1)+(1+\cdots)+(1)$. This gives us a composition of 2012; conversely, given a composition we replace each part a_i by $(1+1+\cdots+1)$ of appropriate length, and get a subset of the original + signs. Thus the number of compositions of 2012 is the same as the number of subsets of a set of size 2011, hence it is 2^{2011}.

EXERCISE 7.1. How many solutions of the equation $x_1+x_2+\cdots+x_{100} = 2012$ are there, in which $x_1, x_2, \ldots, x_{100}$ are positive integers?

EXERCISE 7.2. Let's write $n = `d_m d_{m-1} \cdots d_2 d_1 d_0`$ where $0 \leqslant d_i \leqslant 9$ for $0 \leqslant i \leqslant m$ and $d_m \neq 0$, to mean that n is an $(m+1)$-digit natural number and the given string of digits is its decimal representation. Prove that n and $d_0 + d_1 + \cdots + d_{m-1} + d_m$ leave the same remainder when divided by 9.

EXERCISE 7.3. Does there exist a positive integer N which is a power of 2, and a different positive integer M obtained from N by permuting its digits (in the usual base 10 representation), such that M is also a power of 2? Note that we do not allow the base 10 representation of a positive integer to begin with 0. [UK Maths Trust, Mathematical Olympiad for Girls, 20 September 2012.]

EXERCISE 7.4. (a) Is it true that every natural number may be expressed as a sum of three square numbers?

(b) Does there exist $N \in \mathbb{N}$ such that for all $n \geqslant N$, n may be expressed as a sum of three square numbers?

EXERCISE 7.5. You know that if $S_r(n) := \sum_{k=1}^{n} k^r$ then $S_0(n) = n$, $S_1(n) = \frac{1}{2}n(n+1)$, and $S_2(n) = \frac{1}{6}n(n+1)(n+2)$. What can be said about $S_3(n)$? Prove that in general $S_r(n)$ may be expressed as a polynomial of degree $r+1$ in n. What is its leading coefficient?

7.2 Problem-solving: a summary

I wrote above that there cannot be rules and recipes for problem-solving in mathematics. I also offered some suggestions: experimentation, pattern-spotting, conjecture, followed by refutation and return to square 1, or verification, writing-up and celebration. You are on your own. Learning the definitions and theorems, learning to appreciate the theories that mathematicians have developed over hundreds of years, learning how to apply them—these are the most important ingredients. Add native wit, low cunning and initiative. Enhance the mixture with experience—that is to say, much practice. Persevere. If at first you don't succeed, pause, do something else, have a coffee (remember the Hungarian view that a mathematician is a machine for turning coffee into theorems), come back to the problem and try again. Remember Samuel Beckett's words:

> No matter. Try again. Fail again. Fail better.

That's the way to learn. That's the way to success.