Chapter 1

Mathematics

What is mathematics? It is a strange irony that a subject which is generally accepted as being based on rigour and clarity lacks a commonly accepted definition amongst mathematicians. Mathematics is a uniquely ubiquitous discipline: mathematics has something useful to say about nearly every area of human endeavour, yet also produces questions far beyond the realms of anything but mathematics itself. Study of mathematics even raises questions *about* mathematics. Mathematics describes the way the universe works, and, since we are a part of this universe, how we as humans work. Mathematics is universal, yet may also be used powerfully and productively on small or local scales. Some aspects of mathematics appear incredibly remote and difficult, yet others are strangely naturally accessible to all people who, from the youngest of ages, learn to count and intuitively grasp the basic concepts of number, shape and space. In order to make sense of this rich set of ideas, mathematicians search for structure, relationship, consistency and implication. Whereas the physical world might provide motivation for the start of a mathematical journey, the path that develops can greatly transcend this. Mathematics as a body of knowledge and system of thought takes on a life and meaning of its own. It is a difficult and beautiful world to enter, but well worth the effort of the journey.

Of course, most adults know some mathematics, and the best students will have learned a good deal more mathematics than most people by the end of their school career. Have they not already entered the world of mathematics? By the measure of many the answer would definitely be 'yes', but by the measure of most mathematicians the answer would be 'barely'. However, these mathematicians would agree that such students stand at the gateway to a world of mathematics quite unlike that already encountered; a great deal more demanding and almost immeasurably more intellectually satisfying.

The purpose of this book is to help to provide a natural bridge between the mathematics which is typically encountered at school and the mathematics which is typically encountered in a mathematics degree, which we refer to as higher mathematics[1]. But why is a bridge to 'higher' mathematics needed at all? Surely university-level mathematics builds directly on school mathematics? No. Higher mathematics is an altogether much more refined and demanding affair than school mathematics. The level and precision of thinking required will be quite different from that required at school, even for many of the most talented school learners. In fact, a great part of the first year of university study will be spent demolishing the basic structures learnt at school and building new, far more carefully constructed structures in their place. Other parts of the first year of study will be spent creating structures which are likely to be utterly novel, and possibly initially deeply incomprehensible to the typical first year mathematics undergraduate. In both cases, the depth of mathematical thinking required will be quite demanding. Intuition and raw mathematical skill will no longer suffice. Carefully studied logic and long, hard training will be needed.

In short, these difficulties can result in a cognitive gap between school and higher mathematics. Throughout this book we shall endeavour to bridge this gap by starting with simple familiar mathematical concepts and building on these in a careful way to create and explore some of the fascinating basic mathematical structures of higher mathematics, with as much emphasis placed on the 'how' and 'why' as on the 'what'. Each chapter will build up to genuinely difficult and advanced results and theories. To begin this journey we need to spend some time asking ourselves the question: 'What actually *is* mathematics?'. We will then be in a better position to understand the elements of higher mathematics and higher mathematical thinking.

[1]Throughout this chapter, the nature of 'school mathematics' is presented somewhat in caricature to help to stress the points. In no way is it suggested that school mathematics is easy or unstimulating. It takes years of patient learning and practice to reach a stage at which higher mathematics will be comprehensible, in the same way that it takes years of reading and writing practice to be able to read, appreciate and critique Shakespeare.

1.1 What is Mathematics?

'What is mathematics?'. The answer to this question varies with the position of the traveller on their own personal journey; as a result, there are also various levels of detail and precision with which this question can be answered. At the most obvious level, mathematics learning begins at a very early age when children begin to count and play with shapes. Understanding of these concepts is refined throughout schooling, developing into arithmetic and symmetry and later algebra and geometry. However, at a more subtle level, mathematical thinking skills are also developed and refined through cause and effect and problem solving. These logical thinking skills develop quietly in the background as children grapple with the intricacies of the development of their sense of number and shape. Let us look at some of the developing complexities of these ideas; this will help us to form the basis of the transition from school to higher mathematics.

1.1.1 *Mathematical concepts at various levels*

To get a feel for mathematical progression, let us look at some of the various stages of understanding with the three particular topics of addition, circles and mathematical reasoning[2]. This will help us to understand the basic elements which will go into the creation of advanced mathematical theories; far from being simply 'mathematics', some of the ideas will involve definition, some the process of doing mathematics and some the consequences of both definition and process.

[2] Mathematical reasoning does not develop linearly; many of the ideas in the following lists can be accessed at various levels. Pattern recognition, for example, is developed from infancy. However, the examples presented here do fall roughly into a progession of mathematical sophistication.

Addition

Activity	*Examples*
Conservation laws	Order of counting does not change quantity
Adding real objects	Two bananas and one banana gives three bananas
Single digit abstract numbers	$2 + 7 = 9$
Multiple digits	$129 + 11 = 140$
Fractions	$\frac{6}{37} + \frac{1}{2} + \frac{2}{23} = \frac{1275}{1702}$
Decimals	$1.78 + 10.038 = 11.818$
Negative numbers	$-21 + (-34) = -55$
Clock arithmetic	7+9 $= 4$
Unknown numbers	$(2x - 1) + (6 - 9x) = 5 - 7x$
Vectors	$(1, 2) + (3, -6) + (-4, 4) = (0, 0)$
Complex numbers	$(2 + i) + \overline{(2 + i)} = 4$
Infinite sums	$\frac{1}{2} + \frac{1}{4} + \frac{1}{8} + \frac{1}{16} + \cdots = 1$

Circles

Activity	*Examples*
Shape	Circles are round
Description	Circles are shapes with fixed distance from a centre
Symmetries	Circles are symmetric about any fold or rotation through their centres
Length	Circles have circumference $2\pi r$
Area	Circles have area πr^2
Algebraic description	Circles have equations $x^2 + y^2 = r^2$
Complex description	Using complex numbers, circles have equations $\|z - c\| = r$
Transformations	$f(z) \to \frac{az+b}{cz+d}$ will transform a circle into another circle

Mathematical reasoning

Activity	*Examples*
Cause and effect	The tower falls over if I hit it
If	I can buy things if I have some money
Only if	I can buy a particular item only if I have enough money
Generalised thinking	$3+3=6, 17+57=74$ →two odds make an even
Abstraction	A circle is a line with no width
Therefore	The length is 2m and the width is 6m therefore the area is 12m ^{2}.
If – then	If the length is 2m and the width is 6m then the area is 12m ^{2}.
Using unknowns	If the length is a and the width is b then the area is ab.
Abstract reasoning	If all sprockets are blue and all widgets are sprockets then all widgets are blue.
If and Only If	A number is prime if and only if it has exactly two factors
Use of functions	If $\sin(x+27)=0$ then $x=180\times n-27$
Paradoxes	Consider a big book UltraRef which refers to all books which do not refer to themselves. Does UltraRef refer to itself?

These examples show that as mathematics becomes more advanced, it becomes more interrelated. It is this interrelated set of ideas which makes it difficult to grasp the key concepts of mathematics: what is a derived concept and what is a root concept? Furthermore, new complexities and questions creep in: what is space? what actually is a number? can we always be sure of our logic? Let us try to see how deconstructing some of these ideas raises many questions and ambiguities. In resolving and piecing back together the resulting ideas, we will be able to form a firm base on which to build our higher mathematical activity.

1.1.1.1 *Deconstructing pieces of mathematics*

Here we look at three simple examples of school mathematics. Whilst they might seem on the surface to be rather clear, under the surface a great many questions would need clarifying by the mathematician.

- Take the standard, algebraic definition of a circle:

$$x^2 + y^2 = r^2 \quad r > 0$$

  This simple equation is a shorthand representation of the statement that:

  - A point with Cartesian coordinates (x, y) will lie on the circle if and only if $x^2 + y^2 = r^2$

  What issues does this expression raise? Firstly, the definition requires us to have a firm concept of a point in space, the meaning of the coordinates of those points and also the meaning of the numbers x, y and their squares x^2 and y^2. What sort of numbers are they? Why does each point on the circle have a corresponding coordinate? Furthermore, we need to understand the meaning of an equation and also the logical meaning of the words 'if and only if'. In addition, why does this definition encapsulate the less formal requirement that a circle is 'round'? Would it make sense to say that the 'area' of the circle defined in this way is πr^2? Would this area include the edge of the circle? Would two intersecting circles share a common point?
- Consider the statement that 'Regular hexagons can cover a piece of paper with no gaps or overlaps.'
  Although this statement implies a strong intuitive meaning, as might be grasped even by very young children, this is a very vague statement mathematically. Various questions will arise: Does the paper have to be flat? What would we mean by flat? What is a regular hexagon? Does the size of the piece of paper matter? What if the piece of paper had holes in it? What do we mean by 'no gaps or overlaps'; Do the edges of the tesselating hexagons lie on top of each other?
- Fractions and decimals: $\frac{1}{8} = 0.125$ $\frac{1}{11} = 0.09090909\ldots$
  Whilst these conversions and representations might seem straightforward, fractions and decimals are troublesome if analysed closely. Do all numbers have a fraction representation? Do all numbers have a decimal representation? Is each number represented uniquely by a fraction or a decimal? Can we add these different types of representation together?

Whilst for simple, practical applications of mathematics these simple descriptions of concepts are intuitively clear enough to avoid any major conflict or confusion, higher mathematics asks more difficult and subtle questions, some of which we have alluded to in the lists of mathematical concepts at various level. At these levels of operation, the questions and uncertainties raised by loose definitions become too troublesome to overlook. This mix of ideas needs to be straightened out before further progress can be made with confidence: the basic foundations of mathematics needs to be analysed and made precise before any more real progress can be made.

Let us start by looking at a reasonably advanced piece of school mathematics. We can use this as a template to deconstruct and understand the key elements of any mathematical theory: axioms, theorems, proofs and logic.

1.1.2 *Two school-style mathematical theories*

1.1.2.1 *Defining sine and cosine from right angled triangles*

Definitions: For any right angled triangle with an angle θ we can define the *sine* and *cosine* of the angle θ as follows:

$$\sin\theta = \frac{O}{H} \quad \cos\theta = \frac{A}{H},$$

where H is the hypotenuse of the triangle and O and A are the lengths of the sides opposite and adjacent to the angle θ respectively (Fig. 1.1)

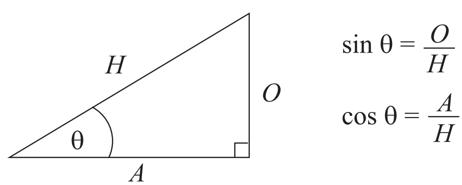

Fig. 1.1 A simple definition of sine and cosine.

Since the sine and cosine are defined using numerical lengths, it is assumed that $\sin\theta$ and $\cos\theta$ can be multiplied, added, subtracted and divided in the same way that any numbers can. Using these numerical operations, we can make logical deductions concerning our numbers $\sin\theta$ and $\cos\theta$, such as the following:

First result: $\sin^2\theta + \cos^2\theta = 1$.

Justification: To show this we can use the equation

$$\sin^2\theta + \cos^2\theta = \left(\frac{O}{H}\right)^2 + \left(\frac{A}{H}\right)^2 = \frac{O^2 + A^2}{H^2}$$

Since the triangle used in the definition is right-angled, Pythagoras's theorem tells us that $H^2 = O^2 + A^2$, which shows that the ratio is 1.

Second result: The sine and cosine of the angle θ are always between 0 and 1.

Justification: This fact is true because in any right angled triangle, the hypotenuse is always longer than each of the other two sides. Therefore the ratios O/H and A/H are always individually less than 1. Therefore, the sine and cosine are always between 0 and 1.

1.1.2.2 *Solution of quadratic equations*

Definition: A quadratic equation is an equation of the following form:

$$ax^2 + bx + c = 0 \quad \text{for any numbers } a, b, c$$

First result: The solutions to a quadratic equation are

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

Justification: Note that if $ax^2 + bx + c = 0$ then 'completing the square' shows that

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

This then shows that

$$x = -\frac{b}{2a} + \sqrt{\frac{b^2}{4a^2} - \frac{c}{a}}$$

Taking the positive and negative signs for the square root and taking out a factor of $\frac{1}{2a}$ provides the answer.

Second result: The two solutions to a quadratic equation sum to $-\frac{b}{a}$; the product of the two solutions is $\frac{c}{a}$.

Justification: Using the previous result we can see that the equations for two solutions x_1, x_2 to the quadratic equation are

$$x_1 = \frac{-b + \sqrt{b^2 - 4ac}}{2a} \quad \text{and} \quad x_2 = \frac{-b - \sqrt{b^2 - 4ac}}{2a}$$

If these are added, then the square roots cancel out leaving $\frac{-b}{2a} + \frac{-b}{2a} = -\frac{b}{a}$, as required. Multiplication of these two solutions and making use of the 'difference of two squares' identity $(p+q)(p-q) = p^2 - q^2$ provides $x_1 x_2 = \frac{c}{a}$.

1.1.3 *Components of a mathematical theory*

In looking at these school mathematical setups we see quite clearly the three key components of any mathematical theory: Definitions, results and justification of those results. For example, in the previous presentation the sine is *defined* to be O/H and a quadratic equation is *defined* to be $ax^2 + bx + c = 0$; a key result is that $\sin^2\theta + \cos^2\theta = 1$ or that the product of the solutions to a quadratic equation equals c/a; these results were justified with a mixture of algebra and logic. Higher mathematics reuses these ideas, but makes them far more precise, as follows:

Definitions are the basic starting point of any mathematical theory. They are used without justification: they are simply the rules of the game[3]. They must be made clear and precise. They must not refer to any quantity which has not already been clearly and precisely defined elsewhere. Not only must we define the objects used in the theory, we must also define clearly the allowed operations, such as addition or division of numbers. Together this set of definitions will be called the **axioms** of the theory.

Results must be clearly stated in terms of the objects and operations defined in the axioms. A fact stemming from the axioms which is known to be true will be called a **theorem**, whereas a speculative result, the truth of which is uncertain, will be called a **conjecture**.

Justifications of the theorems will be called *proofs*. Far from being a convincing argument, a proof must show the theorem to be true beyond all doubt[4]. There is no recipe for a proof: it can be any logically consistent

[3] Of course, coming up with a good set of rules which lead to a rich and interesting mathematical theory is a key part of mathematics.

[4] Of course, nothing can ever be *completely* beyond doubt: we could keep probing each step in a proof endlessly. However, the mathematical community seems to be in agreement with the nature of acceptable logical steps in a proof. We shall discuss these acceptable forms of logic and proof later. Note for now the important point that proof is

argument which derives the theorem from the axioms. Deciding on the difference between a proof and a convincing argument takes some experience and often much thought, but the key idea is to be sure that every step in the proof is 100% certain to be true. In order to facilitate this, mathematical proofs are often broken down into a sequence of small, uncontroversial steps.

How does the trigonometrical example stand up to the mathematicians glare? There are several obvious points that need refinement:

- What is a right-angled triangle?
- Why are the numbers O, A and H positive?
- Why is H longer than O for all triangles?
- Why can we be sure that for all positive numbers O and H that if $O < H$ then $\frac{O}{H} < 1$?

The quadratic equation example seems less controversial, but still raises several issues, particularly about the acceptable manipulation of numbers:

- What do we do if $b^2 - 4ac < 0$, in which case the square roots are not real numbers? Can we be sure that the algebra works for complex numbers?
- The quadratic equation solution can involve irrational numbers. Can we be sure that the difference of two squares formula used to derive the second result works for irrational numbers?
- Does our logic really imply that there are two solutions and only two solutions?

Whilst it might be tempting to say that these points are all obvious, mathematics builds heavily on its own results. Rigour in thinking will separate mathematical truth from woolly speculation. It is this rigour that has allowed mathematics to develop to the point at which mankind has been able to construct the complex, computerised and mechanised society in which we live today.

Let us now move forward and use these key ideas to understand the very foundations of mathematics.

---

a subtle concept; with experience you will gain an understanding of critically assessing the validity of proofs.

1.2 The Basic Structure of Mathematics

At the most basic level we may suppose that mathematics reduces to the study of the following ideas:

- Number
- Shape
- Space

Abstractions and idealisations of these concepts are explored using logic and reasoning. This division of mathematics into 'structure' and 'process' is very important. Of course, mathematics has developed far beyond this simple classification, with a rich interplay between all of these ideas. However, though there may be very many different branches of mathematics with different styles of analysis, all of modern mathematics still makes the clear distinction between *definition of structure* and *logical implication*. Definition of structure of a mathematical theory provides us with a series of statements which are not to be questioned: for the purposes of the theory in hand, they are simply facts or rules. Logical implication allows us to examine the consequences of these rules. It is assumed that in the context of the mathematical theory it is possible to make some *true* statements and some *false* statements. Put bluntly, mathematical endeavour is all about trying to find some true statements which are both interesting and not initially obvious from our definitions.

1.2.1 *What is a mathematical theory?*

Roughly speaking, the definition of a mathematical theory requires two parts:

(1) Definition of the *objects* in the theory (such as numbers, sets, shapes of a certain kind, functions).
(2) Definition of the *operations* in the theory allowing relationships between objects to be created (such as division, intersection, reflection, composition).

Together, these definitions are called the *axioms*. By logically analysing the axioms, interesting results, or *theorems* concerning the objects and operations can be deduced. Of course, mathematics tries to be as minimal as possible and the axioms from one area are used freely when exploring other areas. For example, once the properties of numbers are understood, numbers can be used as required in other areas of mathematics.

Let us look at these key ideas in turn.

1.2.2 *Axioms and definitions*

The axioms define the starting point for an established mathematical theory. There are various levels of clarity at which definition of this structure can be, and is, created. It is a subtle business, typically with many hidden assumptions: very frequently we will *implicitly* use various pieces of mathematics from elsewhere. In order to make progress in many fields of mathematics, some level of assumption in the definitions is often required; historically, key concepts in mathematics were used productively for centuries before formal grounding was provided[5]. In short, there are various levels of clarity at which useful mathematical theories can be learnt and developed. The concept of 'proof' in these cases tends to be tied in to the level of precision of the definitions; there is a balance to be struck between making progress and questioning every detail at every level. However, mathematicians hope that the details of any theory will eventually emerge with total clarity and precision; the first year of undergraduate mathematical study will typically involve a thorough grounding in the most fundamental of these details. Whilst the concepts can vary significantly in terms of style and content, they all share one common theme: the axioms defining them will be *self-consistent, independent and complete.*

1.2.2.1 *Self-consistency, independence and completeness*

Since the axioms of a theory form the basic truths from which to draw conclusions, it is clear that the axioms ought to be self-consistent. We cannot simply make a set of arbitrary definitions and hope to produce meaningful mathematics. In short, from the axioms we must not be able to logically deduce falsehoods or two contradictory statements. The question concerning this internal consistency of sets of axioms is one of the most subtle in all of mathematics.

Some sets of axioms would rather clearly be inconsistent. Assume, for example, that we wish to investigate an operation X which acts on any two whole numbers n, m as follows:

[5] A brief description of the key historical developments in mathematics is provided in an appendix.

- $X(n, m) = n + m$
- $X(n, m) = -X(m, n)$

Using the first of these axioms we could deduce that

$$X(1, 2) = 1 + 2 = 3 \quad \text{and} \quad X(2, 1) = 2 + 1 = 3$$

But, by using the second of these, we must also have

$$X(1, 2) = -X(2, 1), \text{ thus } 3 = -3$$

Clearly these axioms are inconsistent as, through correct logic, they imply a false statement.

Other sets of axioms are clearly consistent. Consider this example:

- A *bit* can take one of two values: 0 or 1
- Two bits a and b can be combined using operations of *addition* and *multiplication*
- Addition $+$ of two bits a and b is also a bit, taking the value $a + b = 1$ if a and b are not both 0 or $a + b = 0$ if a and b are both 0.
- *Multiplication* $\times$ of two bits a and b is also a bit, taking the value $a \times b = 0$ if a and b are not both 1 or $a \times b = 1$ if a and b are both 1.

It is hard to imagine how these axioms might lead to inconsistencies.

Other mathematical setups are far less clearly consistent or inconsistent. An important historical example centres around the introduction of the imaginary[6] number i, which proved initially to be helpful in the manipulation of algebraic expressions. Was the inclusion of this number to the 'real' numbers, leading to expressions $a+ib$ for real numbers a, b completely consistent? Could it ever lead to inconsistent real numerical results?

Another axiomatic issue is that of the *independence* of axioms. In short, is it possible to *deduce* one of the axioms in a theory from the other axioms? If so, then this axiom is not needed: it is a theorem, a derived result. We could make do with a smaller set of axioms. Consider, for example, a set N of whole numbers and an operation S which acts on whole numbers as follows:

(1) For any whole number n, $S(n)$ is also a whole number with $S(n) = n+1$
(2) For every number n in N, $S(n)$ is also in N

[6] The concept of i will be covered in detail in Chapter 2. Its main property is that it squares to -1.

(3) 1 is in N
(4) 1 does not equal $S(m)$ for any m in N
(5) 1 is the smallest member of N

In this list of axioms, the fifth point is not independent of the first four: it can be deduced as a logical consequence of the other axioms. Can you work out why?

In general, determining independence of a set of axioms is a very subtle idea. As a converse to the idea of independence, we could ask if a set of axioms were *complete*: can all theorems of interest be derived from them, or is more independent structure somewhere implicitly assumed? The most historically significant question of independence and completeness involves Euclid's formulation of geometry over 2000 years ago. He assumed as an axiom that parallel line segments do not intersect if the segments are extended indefinitely. The finest mathematicians for two millennia debated whether this axiom was a logical consequence of Euclid's other axioms of geometry before realising that geometries could exist in which this axiom was false.

Throughout this book we shall encounter several mathematical systems with clearly expressed sets of axioms that form the basis of modern mathematics, such as the axioms defining the real numbers and the axioms defining a mathematical group. On sets of axioms such as these mathematics is built. But how do we build the mathematics? In this discussion, it has been assumed that it makes sense to make logical deductions from the axioms. Why and how is this possible? Should the possible logical deductions be defined along with the axioms? No: it is generally agreed that logical implication exists outside of any particular mathematical system; it is the universal glue that binds mathematics together. We need to discuss this as a concept in its own right.

1.3 Mathematical Logic

Logical thinking is not exclusive to mathematics. Barristers and detectives are experts in creating logical arguments on the basis of some evidence to prove some particular point and we all intuitively know when someone's response to an event is 'illogical'. Let us look at some examples of logic in action. We can then refine these ideas to determine the components of a 'good' logical argument.

(1) You said that you were with Arthur last night but Arthur was with me last night. Therefore you are lying.
(2) Some of the suspect's fresh blood and footprints were found at the crime scene and neither were present 10 minutes before the time of the crime. Therefore the suspect was present at the time of the crime.
(3) If the Bank of England announces a rate cut then my mortgage payments will not go up.
(4) Joseph grew up in London, so he is familiar with the tube system.
(5) One thousand people took a test, scoring an average of 26%. Therefore some people scored more than 26%.
(6) If either your mother or father is a British citizen then you are a British citizen.
(7) The pass mark is 50%. You got 62 marks. Therefore, you won't have failed.
(8) There are two red balls and three green balls in a bag. Therefore the chance of randomly picking a green ball is 60%.
(9) If I miss the 7:23 train then I will catch a later train and I will be late for work.
(10) Although I missed the last minute of the football match I know that we won because we were 3 - 0 up.

These statements all seem quite reasonable and would probably be uncontroversial in normal conversation. But what makes an argument logically *certain* as opposed to merely rather *persuasive* or *likely*? Let us choose a defensive stance and argue against these statements as best we can:

(1) Maybe Arthur spent some time with me in the evening before going over to see you? Maybe I was also with you last night?
(2) Maybe the suspect was coincidentally injured at the crime scene just prior to the time of the crime and left the area to find a bandage?
(3) Maybe the cost of lending to individuals can change independently of that interest rate?
(4) Maybe Joseph has only ever travelled by car or bus?
(5) Maybe all 1000 people got the same mark 26%?
(6) Maybe your parents were American when you were born and several years later one of them became a British citizen?
(7) Maybe the test wasn't marked out of 100? You might have got 62 out of 130 marks which would be a fail.
(8) Maybe there are also some yellow, purple and orange balls in the bag?
(9) Maybe you miss the train and then, for some reason, all subsequent

trains are cancelled? Then you will not catch a later train.
(10) Maybe the other team scored 3 goals in the last minute.

If we are examining the logic of these statements critically, as mathematicians do, then we see that they are all somewhat lacking in logical certainty, even if our arguments against the statements seem improbable. However, as mathematicians we need to rule out all possible argument or confusion. Good, watertight examples of mathematical logic could be:

- If any Dog is a $Mammal$ and any $Mammal$ is an $Animal$ then either there are no dogs or at least one animal is a dog.
- A positive whole number n is exactly divisible by 7 or 5. Therefore n cannot be both a multiple of 29 and less than 100.
- If $a \geq 0$ and $a \leq 0$ then $a = 0$.

The reason that these statements are certain is that their inputs have been very clearly and precisely defined. There is no ambiguity and no confusion. In general we cannot reliably form a logical step 'If A then B' unless we are absolutely clear as to what constitutes A and B. This is the crucial first step in the development of mathematical logic[7].

1.3.1 *Constructing clear statements*

As a starting point to logic we assume that we can make mathematical statements with a clear meaning which will be either completely true or completely false, such as '$3 > -4$' (true) or '$3\times4 = -12.0923$' (false) or 'The 10^{1000}th digit of π is 6' (unknown at present, but will either be true or false). This binary true/false concept is very important to mathematicians, who are typically not interested in *ambiguous* situations in which a statement is not clearly either true or false. When statements are vaguely stated, mathematicians will first tidy up the statement until it says something clear and specific. For example 'the old days were the best' and 'there are lots of prime numbers' might be clarified into 'more than 95% of people surveyed over 70 years old in the UK said 'yes' in response to the question "do you believe that the quality of life of pensioners has deteriorated over the past 50 years?"' and 'the number of prime numbers not exceeding n is asymptotically equal to $n/\log(n)$'. The key to a good mathematical

[7] Reading the small print of an insurance document will give a good appreciation of the concept of 'clear' and 'precise' definition in a non-mathematical context. Insurers need to know if a certain event will or will not trigger a claim. This can only be done effectively in relation to a clearly and precisely defined policy.

statement is that it can, in principle, either be wholeheartedly accepted as true or wholeheartedly rejected as false. Put simply:

- A mathematical statement P must be either true or false. Some statements might obviously be true or obviously be false in the context of a given piece of mathematics. A statement might be in the form of an equation, an inequality, a sentence in written English or any other form such that the statement is clearly either true or false, although we might not currently be able to determine which way the statement will go.

This gives our analysis of logic a firm grounding. Let us continue to see how we might use logic in sentences.

1.3.2 *Constructing clear logical sentences*

Mathematical process begins when we are given a collection of statements from which we can deduce whether other statements are true or false. The way in which we make these deductions is with *logical implication.* As we saw in our previous examples there is a wide variety of applications of what one would intuitively call 'logic'. How might this break down into structured pieces? What do these applications have in common? The way forward is to notice that at the heart of logic is the following fundamental form:

- If X is true then Y is true

Thus, if we can determine that X is true then we can write

- X is true therefore Y is true

Let us see how to recast our examples into this form, in which the statements are bracketed and words connecting these statements written in capital letters:

(1) IF (you say that you were with Arthur last night) AND (Arthur was with me last night) THEN (you are lying)

(2) IF (some of the suspect's fresh blood and footprints were found at the crime scene) AND (it is NOT the case that (some of the suspect's fresh blood and footprints were present at the crime scene 10 minutes before the time of the crime)) THEN (the suspect was present at the time of the crime)

(3) IF (the Bank of England announces a rate cut) THEN (it is NOT the case that (my mortgage payments will go up))
(4) IF (Joseph grew up in London) THEN (he will be familiar with the tube system)
(5) IF (one thousand people took a test) AND (the average score was 26%) THEN (some people scored more than 26%)
(6) IF (your father is a British citizen) OR (your mother is a British citizen) THEN (you are a British citizen)
(7) IF (the pass mark is 50%) AND (you scored 62 marks) THEN (it is NOT the case that (you will have failed))
(8) IF (there are two red balls AND three green balls in a bag) THEN (the chance of randomly picking a green ball is 60%)
(9) IF (I miss the 7:23 train) THEN (I will catch a later train) AND (I will be late for work)
(10) IF (we are 3 - nil up) AND (we have 1 minute to play) THEN (we will win the match)

These sentences are now all much logically clearer: we can easily decide on whether each individual bracketed portion is true and then, using these facts, decide whether the implication is also true or whether we need more clarification to cover any queries and holes in the argument. As we shall see, this framework involving the key words IF, THEN, AND, OR and NOT is actually sufficient for building any logical argument, although the details might get somewhat involved[8], as in this example:

- A student will not get a grade *A* unless they are clever and work hard or are very clever and don't do no work.

We can clearly, in the sense of logical structure, write this as:

- IF [it is NOT the case that [((The student is clever) AND (The student works hard)) OR ((The student is very clever) AND (It is NOT the case that (The student does no work)))]] THEN (It is NOT the case that (The student will get a grade A))

[8]Be aware also of linguistic ambiguities. Mathematically P OR Q always means *either or both* of P and Q are true. P AND Q means that *both* P and Q are true.

Although difficult to disentangle, this statement is still of the form 'IF X THEN Y'. As this form is clearly so important, let us consider it further.

1.3.2.1 *Logical implication*

Consider any statement of the form 'IF X THEN Y'. Notice that it does not explicitly say anything about the truth of Y if X is false. For example, (If $x > 2$ then $x^2 > 4$) is certainly valid. However, if $x \leq 2$ then $x^2 > 4$ might or might not be true (consider the cases $x = 1$ and $x = -100$ to see this). We need to be fully aware of this subtlety of logical implication, and for this reason we often use the symbol $\Rightarrow$, called 'implies that', to reinforce this idea: $X \Rightarrow Y$ is the logical statement that 'If X is true then Y is true, but if X is false then Y can legitimately be true or false'[9]. In practice $X \Rightarrow Y$ means **The truth of Y follows from the truth of X.**

Obviously, our logic in writing down $P \Rightarrow Q$ might be correct or incorrect: $P \Rightarrow Q$ is itself a mathematical statement, and thus will itself either be true or false. For example, consider these two statements concerning a real number x: P states 'x has a real square root' and Q states '$x + 1$ has a real square root'. Then the statement $P \Rightarrow Q$ is true, whereas the statement $Q \Rightarrow P$ is false.

1.3.3 ***Propositional logic***

We have investigated some of the core ideas of logic and seen the importance of connecting statements with AND, OR, NOT and $\Rightarrow$. Now that we have a feel for the ideas involved, let us now explore the general question of how we might combine simple mathematical statements to create compound mathematical statements. Of course, we will still need to combine them in ways such that the combination is still either clearly true or false. The set of acceptable combinations forms the subject of *propositional logic*. Propositional logic allows us to combine mathematical statements[10] to produce other mathematical statements in particular ways so that the combination always forms a valid mathematical statement. Throughout we will find some conflict with the ambiguities of ordinary spoken language. Part of the job of the mathematician is to be sure that his or her spoken

[9] Note that if $X \Rightarrow Y$ and $Y \Rightarrow X$ then X is true if and only if Y is true, which we can write as $X \Leftrightarrow Y$.

[10] Statements are sometimes referred to as *propositions* or *assertions*. When the truth or falsehood of a statement P depends on the value of a variable then P is called a *propositional function*. An example would be the statement $(x < 3)$.

language is as free from ambiguity as possible.

1.3.3.1 *Negation*

As the simplest of examples, if we are given a single mathematical statement P then we can construct its *negation* which is false if P is true and true if P is false. We can use an obvious notation for the negation as NOT(P).

Although negation appears to be very simple, in practice, negating a sentence can be somewhat involved. Furthermore there are passive ways of negating a sentence, in which we say what *does not* happen, and active ways to negate a sentence in which we say what *does* happen. It is worth stating that *exactly one* of P and NOT(P) will be true; it is sometimes tempting to negate a statement by negating one of the words in the sentence, but this can lead to errors in logic. For example, the negation of '$P \equiv$ I won the match' is not 'NOT(P) $\equiv$ I lost the match', because in the case of a draw, neither are true. A passive negation would be 'NOT(P) $\equiv$ I did not win the match; an active negation would use the fact that a match can be won, lost or drawn to given 'NOT(P) $\equiv$ I lost or drew the match'.

Let us see how negation can develop in complexity:

- $x < 2 \leftrightarrow x \geq 2$
- $x^2 > 4 \leftrightarrow -2 \leq x \leq 2$
- All even numbers are the sum of two prime numbers $\leftrightarrow$ At least one even number is not the sum of two prime numbers.
- There exists a Fibonacci number greater than 1 which is a perfect 7th power. $\leftrightarrow$ All perfect 7th powers greater than 1 are not Fibonacci numbers.

1.3.3.2 *Compound statements*

The NOT operation is special because it acts on a single statement. How might we wish to combine multiple statements to make new statements? There are presumably very many ways in which statements could be combined? How are we to disentangle all of the possibilities? In mathematics it is often best to start with the simplest of cases and build from there: suppose that we wish to combine just two statements P and Q to form a new statement R. Since P and Q are either true or false and we require that we can determine the truth or falsehood of R directly from that of P and Q there are only actually four combinations of possibility for the truth or falsehood of R. We can tabulate these in a *truth table*:

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

Since R can take one of two possible values for each row of this truth table, there are 16 distinct completions of the truth table:

P	Q	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
T	T	T	T	T	T	T	T	T	T	F	F	F	F	F	F	F	F
T	F	T	T	T	F	T	F	F	F	F	F	F	T	F	T	T	T
F	T	T	T	F	T	F	T	F	F	F	F	T	F	T	F	T	T
F	F	T	F	T	T	F	F	T	F	F	T	F	F	T	T	F	T

Of these values it is possible to extract a set of particularly key distinct logical possibilities. To see why, note that of these 16 sets of possibilities, the second group of 8 correspond to the negations of the first 8; in this sense they are equivalent in logical content. Of the first eight only five are properly dependent on the values of *both* P and Q: 2, 3, 4, 7 and 8. Of the others, 1 is a statement which is always true, irrespective of the values of P and Q. The values of 5 and 6 only depend on the values of one of P or Q: they are logically equivalent to P and Q respectively. Rather nicely, the remaining combinations may neatly be described using the concepts of AND, OR and $\Rightarrow$[11]

P	Q	P AND Q	P OR Q	$P \Rightarrow Q$	$Q \Rightarrow P$	$P \Leftrightarrow Q$
T	T	T	T	T	T	T
T	F	F	T	F	T	F
F	T	F	T	T	F	F
F	F	F	F	T	T	T

Since we looked at all of the possibilities for the combination of two statements and all larger statements can be broken down into combinations of two statements, our logical classification is complete. We can use various

[11]In formal logic AND is often referred to as *conjunction* with the symbol $\wedge$; OR is likewise referred to as *disjunction* with the symbol $\vee$ and NOT is called *negation* and given the symbol $\sim$. In addition, they may all also be described using combinations of NOT and AND. For example, P OR Q is equivalent to NOT(NOT(P) AND NOT(Q)). However, despite this non-minimality the operations of AND, OR and $\Rightarrow$ have very clear intuitive meanings and these are the most regularly used operations in proofs.

combinations of true, false, AND, OR, NOT and implication applied to the axioms in the development of *proofs*. This, in a nutshell, is the process of mathematics.

1.3.4 *Proof*

Proving a result in mathematics can be quite similar to the process of a criminal trial. The mathematician wishes to prove a certain result to be true; the prosecution lawyer wishes to prove that the defendant is guilty of a certain crime. In both cases, evidence for the result is collected and a case for the result is then logically constructed. However, the meaning of proof in the two cases is very different: in an English trial of law a defendant is proven guilty of a crime if the evidence provided indicates guilt 'beyond all *reasonable* doubt'; in mathematics, a result is proved true if evidence is provided 'beyond *all* doubt'. A mathematical proof is simply any logical argument which demonstrates the truth of the result beyond doubt: there can be no logical holes, uncertainties or ambiguities in a mathematical proof[12]. Whilst mathematicians might believe that a result is true, until a proof is provided, the result remains *conjecture*. Sometimes interesting conjectures occupy mathematicians for centuries until proofs are found or the result is shown to be false. Many conjectures are described throughout this book; several play a key part in the history of mathematics.

1.3.4.1 *The elements of proof*

Whilst the mathematician is free to provide proof by any watertight means, there are five main methods of proof which are used frequently. In the same way that a complex chemical compound is created from groups of atoms connected in certain ways, a long proof will contain a sequence of logical steps taking the reader from the initial statement to the conclusion. The most fundamental ideas in mathematical proof are the following direct arguments:

- *Direct proof*[13]
  The first form of direct proof is as follows:

  P is true and $P \Rightarrow Q$ is true, therefore Q is true

  The second form of direct proof is:

[12] This absolute certainty, of course, requires absolute clarity of definition of the result.

[13] These direct proofs form the very basis of mathematical proof. They are sometimes referred to as *modus ponendo ponens* and *modus tollendo tollens* respectively.

$P \Rightarrow Q$ and Q is false, therefore P is false

These types of proof are often used in algebraic manipulations. Often a great deal of effort goes into justifying the truth of $P \Rightarrow Q$; once this is done we may reuse the result freely.

Example: For whole numbers N and M let P and Q be the statements

$$P \equiv N \text{ is a prime number} \qquad Q \equiv N \text{ divides exactly into } M^N - M$$

Then Fermat's little theorem[14] tells us that $P \Rightarrow Q$. Thus, *if* P is true *then* Q must also be true.
As an example, we can deduce that $(1000^{17} - 1000) \div 17$ is a whole number.

Example: Let P be the statement 'Napoleon is from London' and Q be the statement 'Napoleon is English'. Given that $P \Rightarrow Q$ and that Q is false then we can deduce that P is false.

Other forms of mathematical proof build on these direct ideas. Some of the other and most popular and productive proof forms are as follows:

- *Proof by contrapositive*
  Suppose that we wish to prove $P \Rightarrow Q$. If we can prove that NOT(Q) $\Rightarrow$ NOT(P) then we can deduce that $P \Rightarrow Q$.

  Example: To prove that (x^2 is even) $\Rightarrow$ (x is even) we need only prove that (x is odd) $\Rightarrow$ (x^2 is odd).
- *Proof by contradiction*
  Suppose that we wish to prove that a statement P is true. Suppose that we can deduce that NOT(P) $\Rightarrow$ (Q AND NOT(Q)), where Q is *any* statement, which may or may not be related to P. Notice that the right-hand side of this statement must be false: Q and NOT(Q) are contradictory statements. Therefore NOT(P) cannot be true in which case P must be true.

  Example: Let P be the statement 'There is no smallest positive fraction'. Then NOT(P) is the statement 'There is a smallest positive fraction'. If NOT(P) is true then there is a smallest positive fraction. Call this number ϵ and let Q be the statement 'ϵ is the smallest fraction'. For any positive fraction p the number $\frac{p}{2}$ is also a positive fraction. Therefore $\frac{\epsilon}{2}$ is also a positive fraction. But this

[14] This is proved in the Numbers chapter.

fraction is less than ϵ; therefore ϵ is not the smallest positive fraction, in which case NOT(Q) must be true. Thus, NOT(P) implies both Q and NOT(Q), which is a contradiction. Therefore P must be true: there is no smallest positive fraction.

- *Disproof by counterexample*
  Disproof by counterexample is used to prove that certain conjectured statements are, in fact, false in the following way. Suppose that we wish to prove that a statement $P(x)$ depending on an input x is true for each x. Supposing that we discover a single value x such that $P(x)$ is false. This proves that the statement is false.

  Example: $2^{2^n} + 1$ is a prime number for all whole numbers n.
  To disprove this conjecture, notice, as Euler did, the counterexample for $n = 5$: $2^{2^5} + 1 = 641 \times 6,700,417$, which is not a prime number.
- *Proof by induction*
  Suppose that we have a set of statements $P(n)$, one for each whole number n. Suppose that if $P(k)$ is true, then $P(k+1)$ is true and that $P(1)$ is true. Then[15] $P(n)$ is true for all whole numbers n.

  Example: Let $P(n)$ be the statement that $n^3 - n$ is a multiple of 3. Notice that

  $$(k+1)^3 - (k+1) = k^3 + 3k^2 + 3k + 1 - k - 1 = (k^3 - k) + 3(k^2 + k)$$

  If $k^3 - k$ is a multiple of 3, then $(k+1)^3 - (k+1)$ would also be a multiple of 3. Furthermore, $P(1)$ is the statement that $1^1 - 1$ is a multiple of 3, which is clearly true. Therefore, by induction, $P(n)$ is true for all whole numbers n.

1.3.4.2 *Problems with proofs*

Whilst the logical foundations of proof have remained unchanged for centuries, the practice of mathematics has altered significantly over recent decades to the point at which the question of proof has needed some reassessment. In essence, the only way in which a proof of a statement $P \Rightarrow Q$ can be tested is by human mathematicians examining the logic and agreeing that the logic is watertight. Until around the 1900s, it was possible for the greatest of minds to have a good understanding of almost all of

[15] This is not a strictly obvious step and requires the 'principle of mathematical induction', which we discuss in the development of number.

mathematics[16]. In addition, many of the mathematical structures under investigation at the time were, compared to the mathematics of today, very simple and could be very clearly expressed. A single mathematician could sit down and individually verify the details of a typical proof, which might run to the length of a few pages, in a short time.

At the start of the new millennium mathematics has become a very deep and complex affair, with global cooperation between thousands of mathematicians being used to probe the boundaries of mathematics. In most topic areas it takes years of study to reach the point at which a research contribution is possible, and then typically only in a very narrow area of expertise. Many proofs of results use theorems of other mathematicians which, for practical reasons of time, need to be taken on trust. Some proofs are literally thousands of pages long and incredibly complex. For example, the 'classification theorem for finite groups' is over 10,000 pages long and was constructed in a team effort by over 100 mathematicians. Can any individual claim to have understood *completely* this proof from start to end? Similarly, understanding Wiles's proof of *Fermat's last theorem*[17] is a massive undertaking. Whilst those few able to understand its details accept its truth, this acceptance does not hold quite the same certainty as the acceptance of much easier results. In short, very complex proofs now often take years to gain acceptance by the mathematics community. They need to be convincing in their strategy, lead to further sensible results which reinforce the mathematical structures and, of course, not be found over time to contain error. In brief, with many very complicated proofs there is still the possibility of error or inconsistency that no one has yet noticed. However, along with the vast length of some proofs often comes the upside of an enormous mathematical structure which is unlikely to collapse completely if a theorem turns out to be in error; mistakes which do turn up in widely accepted results tend to be minor details or holes which need patching up. Rarely will an accepted modern theorem later be found to be catastrophically lacking, and such problems will typically point to some new direction in which mathematics can be taken.

Another very modern facet of mathematics is the use of the computer. This has, of course, revolutionised the study of mathematics. Now, the power of rapid accurate calculation is available to all, not just the rare mathematical genius with a talent for computation. However, some *proofs*

---

[16]Some of the greatest of these mathematicians are described in the appendix 'Great mathematicians and their achievements'.

[17]See Chap. 2 and the appendix on great mathematicians for more details.

now rely on computers checking the details. The most famous example is that of the *four colour theorem*, which states that any map may be coloured with 4 or fewer distinct colours so that no two regions of the same colour share an edge. The 'proof' of this was completed in 1976 by a computer, which checked thousands of complicated configurations of key graphs. The computer was programmed by two human mathematicians Appel and Haken. Their plan seems to be mathematically sound, the computer program seemed correct and the proof has gained acceptance. But is this a proof in the traditional sense of the word? Not simply proof beyond reasonable doubt, but proof *beyond* doubt? Whilst purists may be concerned about such intricacies of proof, most mathematicians have been typically pragmatic and have embraced the power of the computer: computers undoubtedly have massively accelerated the development, discovery and application of new mathematics. All that has changed is the requirement that the computer programs used in proofs be scrutinised to the same level that a written theorem would be and the results verified in independent computer checks.

Logical implication and proof underlie the development of any mathematical system. Now let us consider how we can create and define the objects which make up any such mathematical system. At the most fundamental level, mathematical objects are built up from *sets*. In the description of sets we will make good use of the clarity of our development of mathematical logic.

1.3.5 *Sets*

A set S is simply a *collection of objects*. These objects are often numbers, but might be physical objects, words, functions or, indeed, other sets. Mathematics begins by assuming certain properties of sets which must, unarguably, be true. These natural properties stem from our everyday experience with physical sets of objects, but soon lead us into abstract territory. Indeed, each natural operation concerning manipulation of sets raises interesting structural issues.

As their most basic property, two sets S and T are the same if they contain the same objects, more formally known as *elements*:

- If an element x is found in S, then we write $x \in S$ and say 'x is a member of S'.

- $S = T$ if and only if for all elements x of S or $T, x \in S \Leftrightarrow x \in T$

Whilst the concept of a set transcends practical experience, to discover the natural properties of sets, it can be useful to have a simple physical mental picture of a set S as a sack containing a selection of objects. If we had two such sacks S and T, then we could tip their contents into a bigger sack U. Clearly U would also be a sack containing a selection of objects. This gives our first property of sets, but also our first query and deviation from physicality: although we can clearly join two sets to make another set, what do we do if the same object is found in both S and T? Although it might make sense to have two copies of a certain object, some objects are unique and cannot be copied. We will need to make the decision that we do not double-count when joining sets:

- Given two sets S and T, we can form a set U formed from all elements belonging to either or both S and T. U is called the *union* of S and T, and is written as $U = S \cup T$, with $x \in U \Leftrightarrow (x \in S \text{ OR } x \in T)$.

If we are comfortable with the idea that an object may be found in two or more sets, then we might compare sacks S and T to see which objects they have in common. These objects could be used to form another set:

- Given two sets S and T we can form a set I whose elements belong to both S and T. I is called the *intersection* of S and T, and is written as $I = S \cap T$, with $x \in I \Leftrightarrow (x \in S \text{ AND } x \in T)$.

This, again leads us to a deeper consideration. What happens if the sets S and T contain no objects in common? Our combined sack would be empty. It is natural to allow the existence of a set corresponding to such a sack:

- The *empty set* $\emptyset$ is the set containing no elements.

The empty set also raises questions, such as: is the set of words with both exactly three and exactly two letters simultaneously the same as the set of positive numbers which are simultaneously both even and odd? Both are empty sets, but are they the *same* empty set? Mathematicians, who like to be minimal in their definitions, agree that *the* empty set is unique.

Next imagine that instead of merging the contents of two sacks S and T to create the union of S and T we place the entire unopened sacks S and T into a larger sack U. In this case we could reach in and pull out either sack S or sack T; U is a set of two sets: it has two elements, which happen themselves to be sets. Whilst this concept might sound very reasonable, it raises the very subtle question: can a set be a member of itself?

1.3.5.1 *Sets of sets and Russell's paradox*

Clearly some sets do not contain themselves. To see an example, suppose that S is the set of whole numbers: S does not contain itself as an element, as S is a set of whole numbers, and a set of whole numbers is not a whole number. A set which contains itself must be a set which contains some sets. Unsurprisingly, this can be a tricky concept and often involves the idea of *recursion* in which the definition of a set refers to itself. Perhaps surprisingly, such sets need not be infinite. Consider, for example, the set W of all sets which can be defined using less than 20 standard English words. Clearly W is such a set, and must, therefore, be contained within itself[18]. Clearly this unusual set is very large, yet finite.

Set theory seems to be very reasonable when working with sets which do not contain themselves, and becomes difficult when working with sets which do contain themselves. Set theorists have labelled the former sets *ordinary* and the latter *extraordinary*. The mathematician and philosopher Bertrand Russell noticed that the concept of extraordinary sets lead to all sorts of complications. Consider, for example, the set C of all ordinary sets. Then each element of C is an ordinary set. Now, is C itself ordinary or is it extraordinary? If C is ordinary, then it must be found in C, in which case C must have been extraordinary. If, however, C is extraordinary, then it is not an element of C, in which case C is ordinary. We are thus led to a convoluted logical self-contradiction by the formation of the set C of all ordinary sets. This is *Russell's paradox*, and has no straightforward resolution in mathematics, other than the, in some sense unsatisfying, clause that it is unacceptable to form sets which lead to these contradictions.

Whilst fascinating, and deeply distressing to many of Russell's peers who were attempting to provide all of mathematics with a consistent logical foundation, these questions of the infinite have little consequence for the development of ordinary mathematics. However, it is worth bearing in mind that the deepest foundations of mathematics hide some difficult and troublesome issues.

[18] You might like to consider whether you feel that this set is precisely defined. Could such a set be defined more precisely?

1.3.5.2 *Union and intersection*

The most basic property of sets, beyond their definition, is that the order in which we take the union or intersection of two sets does not matter:

$$\begin{aligned} S \cup T &= T \cup S \\ S \cap T &= T \cap S \end{aligned}$$

Furthermore, since union and intersection of two sets give rise to other sets, we must be able to take sequences of unions and intersections. What properties will these sequences have? Clearly any finite string of unions and intersections can be built up by considering four simple cases: the union followed by a union, the union followed by an intersection, the intersection followed by a union and the intersection followed by an intersection. The operations of union and intersection on sets combine to give us a set of useful rules. They are very reasonable for finite sets of objects:

$$\begin{aligned} (S \cup T) \cup U &= S \cup (T \cup U) \\ (S \cup T) \cap U &= (S \cap U) \cup (T \cap U) \\ (S \cap T) \cup U &= (S \cup U) \cap (T \cup U) \\ (S \cap T) \cap U &= S \cap (T \cap U) \end{aligned}$$

It is helpful to use a *Venn diagram* to understand such statements (Fig. 1.2).

Whilst useful, Venn diagrams only help us by allowing us to imagine the interactions between sets: they are not to be used as proofs of logical statements. To prove that any pictorial representation has a watertight valid meaning requires a great deal of work: What are the 'circles'? What are the 'intersections'? What are the 'boundaries'? What does it mean for an object to be 'in' or 'out' of a circle? To use Venn diagrams to prove logical statements, we would need to carefully examine why they could be made equivalent in some way to logic. It is unclear even now whether this is possible or not[19].

For proofs, we need to use pure logic to help us with our manipulations. For example, to prove that $S \cap (T \cup U) = (S \cap T) \cup (S \cap U)$ for any imaginable sets S, T, U we would need to prove that the two sets on each side of the

---

[19] To see some of the intricacies, consider how we would represent a set which contains itself.

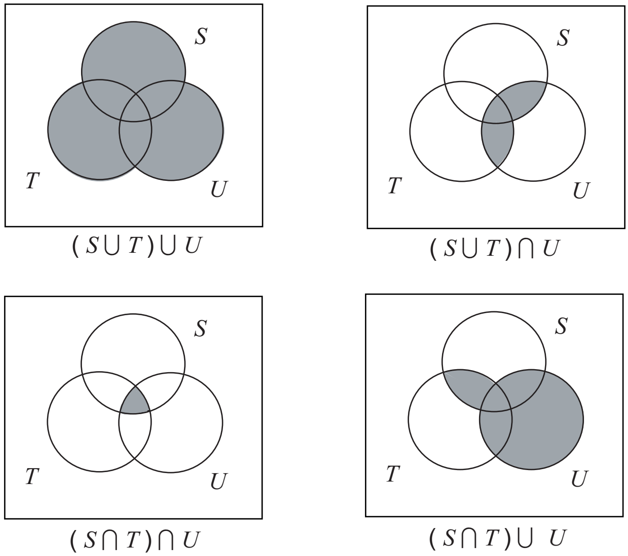

Fig. 1.2 Venn diagrams for combinations of two unions or intersections.

equality have the same elements. For any element x of $S \cap (T \cup U)$ we have

$$\begin{aligned} & x \in S \cap (T \cup U) \\ \Leftrightarrow\ & (x \in S) \text{ AND } (x \in T \cup U) \\ \Leftrightarrow\ & (x \in S) \text{ AND } \Big((x \in T) \text{ OR } (x \in U)\Big) \end{aligned}$$

We can use a truth table to show that for any statements A, B, C

$$A \text{ AND } (B \text{ OR } C) \equiv (A \text{ AND } B) \text{ OR } (A \text{ AND } C)$$

By selecting A, B and C to be the statements $x \in S, x \in T, x \in U$ we can see that

$$\begin{aligned} & x \in S \cap (T \cup U) \\ \Leftrightarrow\ & \Big((x \in S) \text{ AND } (x \in T)\Big) \text{ OR } \Big((x \in S) \text{ AND } (x \in U)\Big) \\ \Leftrightarrow\ & (x \in S \cap T) \text{ OR } (x \in S \cap U) \\ \Leftrightarrow\ & x \in (S \cap T) \cup (S \cap U) \end{aligned}$$

1.3.5.3 *Subsets*

We have investigated merging two sets and seeing what elements two sets have in common, giving us the operations of union and intersection. As a final set-theoretic exercise, let us suppose that we take our sack S and arrange its contents into two disjoint groups of objects T and U. Clearly $S = T \cup U$ and $T \cap U = \emptyset$. We can think of T and U as forming smaller sets contained within S: this leads us to the concept of a *subset* when neither T nor U is empty:

- T is a subset of S, written $T \subset S$, if and only if all the elements of T are also elements of S: $x \in T \Rightarrow x \in S$.

Creating subsets raises an interesting point: what if either of the two sets T and U were empty? We allow two special cases of subset: Any set is a subset of itself and the empty set is a subset of any set. Whilst the notion of a subset might seem to be quite straightforward, it can lead to the creation of very large, interesting sets in a very easy way. For example, consider the set T of subsets of the set $S = \{1, 2, 3\}$. T has $2^3 = 8$ elements:

$$\begin{array}{l}\{\emptyset\}, \{1\}, \{2\}, \{3\},\\ \{1,2\}, \{1,3\}, \{2,3\}\\ \{1,2,3\}\end{array}$$

Now consider the set T' of subsets of T. This would have $2^8 = 256$ elements. The set T'' of subsets of T' would have $2^{256} \approx 10^{77}$ elements, which is around the same as some estimates for the number of particles in the universe. Imagine attempting to form the set of subsets of all of the whole numbers!

1.3.5.4 *The axiom of choice and well-ordering*

The simple notion of a subset allows us to develop very easily the notion of massive sets, too enormous to comprehend. These can cause difficulties with the logical formulation of mathematics. Consider a set S of sets. Imagine the operation of choosing 1 element from each of these sets. Surely this simple process cannot be problematic? Mathematically, we could describe this activity using a function $f(s)$, acting on every $s \in S$, such that $f(s) \in s$. This is a *choice function* which selects an element from each element of S (Fig. 1.3).

It turns out that the theoretical existence of such functions cannot be determined from the other axioms of mathematics; the *axiom of choice*

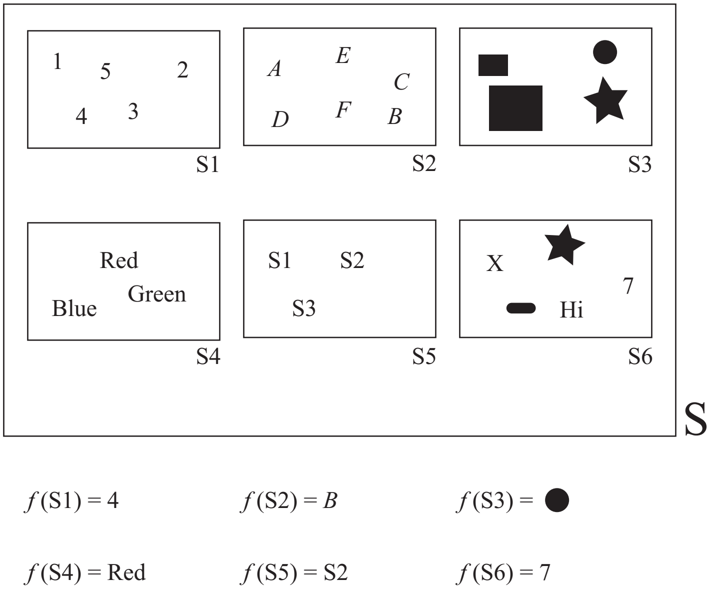

Fig. 1.3 Choosing from a set S of six sets.

states that such a function exists for any set. Typically, mathematicians are happy that the axiom of choice seems intuitively reasonable and most will use it in their proofs of other results. However, the axiom of choice is also logically equivalent to the *well-ordering axiom*, which, roughly speaking, says that you can put the elements of any set into an order so that any subset of that set has a smallest element. Whilst this might seem very reasonable for finite sets, to most mathematicians it seems intuitively entirely unreasonable for sets such as the real numbers. To get a feel for the problem which is raised, consider the set S_x of all real numbers greater than some particular real number x. In the usual way of ordering numbers, this set would have no smallest element: you can always get closer and closer to a real number x without actually reaching it. There is, therefore, no smallest number larger than x given the usual ordering of the real numbers. For example, consider $S_{\frac{1}{3}}$. The number $\frac{1}{3}$ has a recurring decimal expansion $0.33333333\ldots$. The sequence of numbers

$$0.4, 0.34, 0.334, 0.3334, 0.33334, \ldots, 0.\underbrace{3333\ldots333}_{n\ 3s}4, \ldots$$

is decreasing in size and is contained in S_x; yet each term is larger than $\frac{1}{3}$. The well ordering theorem says that we could rearrange the order of the real numbers such that problems like this vanish for all such sets.

To the human intuition of most mathematicians, it is rather surprising that the axioms of choice and well-ordering are logically equivalent. However, they represent study of the extreme boundaries of the foundations of mathematics. Whilst these deepest foundations of mathematics are extremely interesting and important, they do not concern the daily activity of most mathematicians who deal with less radical objects than those which cause set-theoretic concern. What is important to all mathematicians is the solid base of logic and ordinary set theory on which mathematics has been constructed. These are like the foundations of a large building. Mathematics is built on these foundations; the result is a complex and many layered structure, with each layer relying on several lower layers. But how do we go about constructing these layers? We now address the general, practical issue of how one goes about actually *doing* mathematics and solving problems.

1.4 Doing Mathematics

In this book we will explore a wide range of ideas in mathematics; we have begun this exploration with a look at the foundations. To develop as a mathematician, each of these ideas will require a great deal of thought, consideration and reflection. But mathematics is not simply about learning new theories: it is primarily about solving problems using existing mathematics if you are a student and inventing new theories if you have solved enough problems and become a researcher. Let us therefore discuss some of the important ideas surrounding the art of actually *doing mathematics*. Whilst mathematics can be performed at an enormous range of levels of sophistication, here we shall focus on the shift in style required to begin undergraduate mathematical study; this complexity will be developed throughout each subsequent chapter. It is important to realise that the practice of mathematics *is* an art-form; learning it is a slow and subtle business with no set 'answer' or 'recipe', as such there is no well-defined 'accepted way' of doing mathematics. This section provides an overview of some of the key ideas in this area[20].

[20] Please be aware that this material is not standard and contains many personal viewpoints; the views of other mathematicians in this field might vary in places from

1.4.1 *Mathematics from school to university*

Though school mathematics often involves application of a fairly routine set of ideas in fairly routine setting, higher mathematics will be more challenging. Proactively applying problem solving skills will be required to make progress at university and in real-world applications of mathematics. As with any art-form, there is no 'right approach' or 'method' to problem solving, and good problems may yield their solutions in a variety of ways. However, it is very important to realise that the process of problem solving is a discipline which can be considered independently of a particular mathematical context, in the same way that logic exists outside of any particular piece of mathematics. Students who have not sufficiently practised their careful, questioning, tenacious and patient mathematical thinking will find things more difficult than is strictly necessary. Let us start by discussing some of the ideas surrounding problem solving and mathematical reasoning through a series of examples.

1.4.2 *Starting to deconstruct mathematical problem solving*

A typical example of a basic result in higher mathematics might be presented and proved thus:

- For any whole number n, $1^2 + \cdots + n^2 = \frac{1}{6}n(n+1)(2n+1)$

Proof: Suppose that the result is true for the special case $n = N$. Then $1^2 + \cdots + N^2 = \frac{1}{6}N(N+1)(2N+1)$. Adding $(N+1)^2$ to both sides of the equation yields

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

The right-hand side of this expression can be rearranged to give

$$\begin{aligned}\tfrac{1}{6}(N+1)\big(N(2N+1) + 6(N+1)\big) &= \tfrac{1}{6}(N+1)(2N^2 + 7N + 6)\\ &= \tfrac{1}{6}(N+1)(N+2)(2N+3)\end{aligned}$$

But this is equal to $\frac{1}{6}M(M+1)(2M+1)$, where $M = N + 1$. Thus, *if* the result is true when $n = N$ *then* it is also true when $n = N + 1$. Clearly, by substitution, when $n = 1$ the result is true. The result is therefore true for all n by the principle of mathematical induction. □

those of the author. However, it is hoped that the section will provide an engaging and stimulating set of ideas to consider as the reader progresses through the book and on his or her wider mathematical journey.

Whilst this proof is interesting and instructive, it is a finished, polished article, containing a series of fully formed, logical steps. The *procedural* details can be used in attacking similar problems, but the main issue of *why* anyone even attempted to prove the theorem in the first place, and how they initially came up with the identity is left a mystery; it gives little help in the solution of a related problem for which the answer is not known:

- Can you find an expression for the sum $1^3 + \cdots + n^3$?

Doing mathematics is not simply about making a step-by-step series of logical deductions until a problem is solved. It also involves guesswork, trial and error, following hunches, leaps of intuition, trying out a variety of approaches, making lots of mistakes and taking lots of false turns. However, once a solution to a problem is found, mathematicians like to tidy up their work into a neat, compact, clean, logical solution or proof. Whilst this final complete clarification is essential to be mathematically certain of our results, simply looking at the final presentation fails to give an appreciation of the creative process of actually doing the mathematics and the struggle and toil which went in to creating the proof, which can take days, weeks, months or even years to complete. Without an appreciation of the process, only the very tip of the iceberg of mathematical activity will be exposed. Whilst there is no substitute for actually working through mathematical problems, it can be helpful to understand that there are a variety of ways in which mathematics can be done. Let us now look at some of these ways.

1.4.3 *Mathematical thinking*

How do we deconstruct 'mathematical thinking'? At high school, students who are 'good' at mathematics tend to be good at *all* of the mathematics that high school can throw at them. More so than experts at other subjects, a student who is very talented at mathematics tends to be able to perform in their strongest area with ease and without apparent effort. All problems suitable for whole class teaching will be solved with minimal working and effort of thought, regardless of the particular piece of the syllabus being studied. At the other end of the spectrum, in high school there are very large numbers of students, many of whom will be perfectly intelligent and may excel in all manner of other fields, for whom *algebra* is almost totally impenetrable. Given these, commonly seen, extremes, it is not surprising that talented high-school mathematicians are declared to be 'excellent at mathematics'. Nothing more, nothing less. Such statements implicitly bundle up mathematics into a single, indivisible whole, which does not happen in other disciplines: Whereas it would be unsurprising to hear that a talented sportsman might be said to be 'a good all-rounder and an excellent soccer player', or a gifted writer might be 'a good novelist who writes outstanding poetry' it would perhaps be more surprising to hear that Jane was a 'talented algebraist with particularly exceptional skills of visualisation'. In short, mathematics falls foul of the problem that the main development of mathematics only really begins at university. Even excellent high-school students of mathematics might not realise that mathematics breaks down into a set of highly individual disciplines requiring very different mathematical skills and abilities, in the same way that the study and activity of 'writing' breaks down into various categories: fiction, prose, comedy, play-writing, grammar, structure of language, etymology and so on. All of these areas of writing activity require the basic skills of reading, writing and grammar and yet develop in their own unique way. The mathematics learnt at high school (basic number, algebra, and shape) in a very real sense parallels the development of these basic reading, writing and grammar skills. It is at university-level that these basic skills in mathematics will be developed, refined, specialised and put into action; it is at this point that young mathematicians will discover where their real mathematical skills lie.

1.4.3.1 *Types of thinking in mathematics*

Let us begin to understand how there might be different types of mathematical thinking. From the outset it is essential to stress that none of these is 'the best' and none of these is 'the most difficult'. Most mathematicians will have all of these skills developed to a high level, but most will also have a 'strong suit' in which they truly excel. It is also important to be fully aware at the outset that mathematics is *vastly* bigger as a structure than any one person: your journey into mathematics will be individual, and the more mathematics that you do the more your knowledge and particular skills will diverge from other mathematicians. How will your journey develop?

Let us take some steps on this journey by considering how different mathematicians might think their way through a simple high-school problem: We shall use three famous mathematicians to help us: Gauss, Euler and Archimedes[21].

[21] These are three of the greatest mathematicians of all time. Their thinking skills

- Ahmed and Ben each have the same number of sweets. Ben swaps one of his sweets for three of Ahmed's sweets. Ben now has twice as many sweets as Ahmed. How many sweets did each boy start off with?

There are multiple ways in which this problem can be approached: algebraically, experimentally and by using the properties of prime numbers.

Gauss: I see that this problem has one variable, the number of sweets the boys each start with, so I ought to be able to use algebra to solve this problem. Let us suppose that Ahmed and Ben each have n sweets to start off with. After the swap Ben will have $(n - 1 + 3)$ sweets and Ahmed will have $(n - 3 + 1)$ sweets. Therefore

$2(n - 3 + 1) = n - 1 + 3$
Therefore, $2(n - 2) = n + 2$
Therefore, $2n - 4 = n + 2$
Therefore, $n = 6$
I conclude that each must have had 6 sweets to start off with.

Euler: By looking at the numbers in the statement of the problem, I can see that the answer must be a low number of sweets, so let us look at some examples. They must have at least three sweets each. With the swap Ahmed loses 2 and Ben gains 2. Suppose that they start with 3 sweets each. Then they would end up with 5 and 1 sweets so that is wrong. What if they started with 4 sweets each? Then they would end up with 6 and 2 sweets, which is also wrong. 5 sweets goes to 7 and 3, 6 sweets goes to 8 and 4, which are in the ratio 2:1. Thus the boys had 6 sweets to begin with.

Archimedes: Since the two boys have the same number of sweets to begin with, the total must be an even number. Furthermore, since one boy ends up with twice as many sweets as the other, and none are eaten or lost, the sweets can be split into the ratio 2:1. The total must, therefore, also be a multiple of 3. Thus the total must be a multiple of 6. By inspection, there were 12 sweets in total.

Which of these methods are correct? All of them, as they all reach the answer without any incorrect logic along the way. Which of these methods is mathematically the best? None: they are all mathematically sound and each has their place in the repertoire of mathematical techniques and approaches. Of course, you might prefer one method over the others

were almost miraculously well developed, but we shall use each to highlight a particular mode of mathematical operation. Their names are pronounced thus: Gauss rhymes with 'mouse', for Euler say 'oiler' and Archimedes is said as 'are-key-me-dease'.

and in more difficult situations one of the approaches might clearly lead to a more efficient solution. Part of the skill of the mathematician is to be able to assess which approach is likely to be more helpful. One of the problems that the gifted high-school mathematician faces is that high-school mathematics problems almost always have a nice algebraic solution: there is little need to consider other methods, because the algebra will always work, and it will be clear from the setting of the problem that this will be the case[22]. However, advanced mathematics will not be so simple and routine algebraic methods will often fail to yield a solution in any reasonable time. We are led to the following important points:

- There are usually several different ways in which a mathematical problem can be solved.
- Do not simply apply a known method or procedure without some thought as to its suitability

Consider this variation of the problem:

Extension: Ahmed and Ben want to buy some more sweets. Ahmed cannot afford to buy the packet of sweets costing £1.99. However, Ahmed has four times as much money as Ben. Ben gives Ahmed some money until Ahmed has six times as much money as Ben. Ahmed can now afford the sweets. How much money did the two boys start off with in total?

How would you try to solve this variant? Is any one method more effective than the others in this case?

1.4.3.2 *Applying different styles to different problems*

Let us now explore in more detail the possibilities offered by different mathematical thinking styles. Gauss, Euler and Archimedes will each solve the following problems using the approaches they demonstrated in the previous example. First consider this 'standard' problem, for which three different methods offer powerful approaches:

Problem 1: Prove that $1 + 2 + \cdots + n = \frac{1}{2}n(n+1)$

Gauss: I can use induction to prove this expression. Supposing that it is

[22] It is, of course, possible to provide a rich, mathematical problem solving experience for school students, but culturally the algebraic approach for high achievers has been favoured in our schools.

true for $n = N$, I can add $(N+1)$ to each side to give

$$1 + 2 + \cdots + N + (N+1) = \frac{1}{2}N(N+1) + (N+1) = \frac{1}{2}(N+1)\big((N) + (2)\big)$$

This is the same expression with n replaced by $N+1$. Therefore, if the result is true for $n = N$ then it is also true for $n = N+1$. Clearly the result is true for $n = 1$; therefore the result is true by using the principle of induction.

Euler: Since the sum is very regular, I might be able to rearrange the left-hand side to good effect. 1 plus n equals 2 plus $n-1$. Clearly I can continue this process. Since I am doubling up the numbers, I will need to consider the two cases even n and odd n separately. If n is even then I will get $\frac{n}{2}$ lots of $(n+1)$, which immediately gives the result. If n is odd, then I will get $\frac{n-1}{2}$ lots of $(n+1)$ added to the middle value, which equals $\frac{n+1}{2}$. Adding these together gives

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

The result is therefore also true when n is odd.

Archimedes: Let me consider the left-hand side of the expression. Imagine adding up strips of width 1 and of length $1, 2, 3, \ldots, n$. These can be arranged to cover half of an n by n square grid, including the diagonal (Fig. 1.4). Two lots of this would be equal to the area of the n by n square grid with the diagonal counted twice. Since the diagonal is of length n, I find that

$$2(1 + 2 + \cdots + n) = n^2 + n = n(n+1)$$

This proves the result.

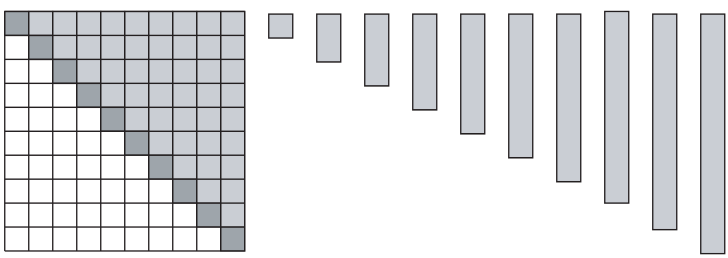

Fig. 1.4 Partially covering a set of squares with a set of strips.

Next, consider this numerical question:

Problem 2: A positive number exceeds its reciprocal by 7. Find a fraction which closely approximates this number.

Gauss: I see that I will need the quantities x, the number, $\frac{1}{x}$, its reciprocal, and the equation $x - \frac{1}{x} = 7$. Clearly x cannot be zero, so this equation makes sense. Multiplying through by x gives me $x^2 - 7x - 1 = 0$. This is a quadratic equation in x which I know how to solve. There are two possible answers; I will need the positive one. This is

$$x = \frac{7 + \sqrt{49+4}}{2} = \frac{7 + \sqrt{53}}{2}$$

Now, this expression involves the square root of 53. I need to approximate this by a fraction. Clearly the whole part of this fraction will be 7, and I can use the binomial theorem to expand the square root:

$$\begin{aligned}\sqrt{53} &= (49+4)^{\frac{1}{2}} \\ &= 7 \times \left(1 + \tfrac{4}{49}\right)^{\frac{1}{2}} \\ &= 7\left(1 + \tfrac{1}{2}\cdot\tfrac{4}{49} + \tfrac{1}{2}\cdot\tfrac{-1}{2}\cdot\left(\tfrac{4}{49}\right)^2 + \cdots\right) \\ &= 7 + \tfrac{2}{7} - \tfrac{4}{343} + \cdots\end{aligned}$$

This gives my approximation as

$$x = \frac{7 + \left(7 + \frac{2}{7} - \frac{4}{343} + \ldots\right)}{2} \approx 7 + \frac{1}{7} - \frac{2}{343}$$

Euler: Let me trial a few numbers:

$$\begin{aligned} 1 - \tfrac{1}{1} &< 7 \\ 2 - \tfrac{1}{2} &< 7 \\ 3 - \tfrac{1}{3} &< 7 \\ \vdots\quad & \ \vdots \\ 7 - \tfrac{1}{7} &< 7 \\ 8 - \tfrac{1}{8} &> 7 \end{aligned}$$

This shows clearly that x must be larger than 7 and smaller than 8 in order that the difference between it and its reciprocal can equal 7. Let us try some fractions; based on these inequalities, the first natural choice is to start with $7\frac{1}{7} = \frac{50}{7}$

$$\frac{50}{7} - \frac{7}{50} = \frac{50\cdot 50 - 7\cdot 7}{7\cdot 50} = \frac{2451}{350} = \frac{350\cdot 7 + 1}{350} = 7 + \frac{1}{350}$$

So, the number $7\frac{1}{7}$ is slightly too large, but yields the required answer of 7 to a good accuracy. I could improve this approximation by subtracting the small remainder $\frac{1}{350}$: $x = 7 + \frac{1}{7} - \frac{1}{350} = 7\frac{7}{50}$.

Archimedes: This problem presents us with a recursive relationship: x equals 7 more than its own reciprocal:

$$x = 7 + \frac{1}{x}$$

I can use this expression to feed x into the right-hand side, to get

$$x = 7 + \frac{1}{7 + \frac{1}{x}}$$

Repeating this process I get an infinite 'continued fraction'

$$x = 7 + \frac{1}{7 + \frac{1}{7 + \frac{1}{7 + \dots}}}$$

I can truncate this to find an approximate answer

$$x \approx 7 + \frac{1}{7 + \frac{1}{7}} = 7 + \frac{1}{\frac{49+1}{7}} = 7 + \frac{7}{50} = 7 + \frac{1}{7} - \frac{1}{350}$$

Thus, the approximation of $x = 7\frac{1}{7}$ is a good one.

These different approaches highlight three basic ways in which mathematicians attack problems: *algorithmic*, *experimental* and *inspirational*.

(1) *Algorithmic*: A known method or approach is applied to a problem. The method is applied in a sequence of steps, each of which moves the solver closer to the answer.
(2) *Experimental*: Various forms of the question and answer are tested out in an attempt to find patterns, structure and, finally, the solution.
(3) *Inspirational*: The structure of the problem is analysed and probed in a variety of ways. Alternative representations and links to other areas of mathematics are considered until a good idea arrives.

Mathematicians will not operate solely in one mode or another, but many will have a preferred style: Some mathematicians naturally spot links to other areas of mathematics or may simply intuit solutions almost from thin air; some prefer to play or experiment with a problem and tidy up the details at the end; others are most comfortable following clear, well-defined procedures on the way to the solution of a problem or, if such a procedure is lacking, absolute clarity at each step of the process. At high-school level, problems are most typically algorithmic: from the context of the problem it will often be clear which piece of mathematics is required to reach the solution; the process reduces to setting up the algebra correctly and solving the equations. Many students feel very uncomfortable without this clear signposting of a routine or example to follow in a mathematics problem. Luckily, higher mathematics is rarely so obvious: the more difficult and involved problems become in mathematics, the more a combination of all three modes will be required to find solutions: experimentation leads to links, links lead to possible algorithms or simplification; repeating this cycle and discarding or amending false leads will hopefully end up with a solution. The more skilled the mathematician is at the art of problem solving, the sooner he or she will arrive at solution. The skill of the mathematician depends a great deal on practice and experience, but also depends greatly on their particular type of mathematical *intelligence*.

1.4.3.3 *Types of intelligence in mathematics*

Higher mathematics breaks down into several rather distinct areas of activity, each with their own style and flavour. Mathematicians who find it easy to operate in one area might find it difficult to operate in another and vice-versa; the most commonly cited division highlighting this fact is the perceived split between 'pure' and 'applied' mathematics. However, even within these areas there are great differences in the mathematical style of the various subjects and most mathematicians find themselves naturally drawn to a few particular areas in which they excel. There thus appears to exist a real concept of mathematical intelligence: particular aspects of intelligence which relate to mathematical thought. Within the area of mathematical intelligence, we can determine a variety of sets of different skills. Two obvious skills within mathematics are algebraic skill and visualisation skill, yet there are many others. It is helpful to realise that mathematical intelligence can be broken down in this way[23]:

- *Abstract thinking:* The ability to work with objects or processes lacking any physical or obvious geometrical representation.

[23] There are many ways in which to break down 'mathematical skills' or 'mathematical intelligence'. This list is a simple indication of the possibilities and does not represent a 'rigid' or 'complete' categorisation. However, it should serve to make you more thoughtful about your own practice. By becoming more aware of your personal style, you will become a better mathematician.

- *Algebraic skill:* The ability to manipulate correctly symbols according to fixed, pre-determined rules.
- *Memory and knowledge:* A great many areas of mathematics make use of a large body of theoretical results. Simply being able to remember these results in *exact* detail is a highly useful mathematical skill.
- *Estimation skill:* The ability to assess relative magnitude, make or guess approximate answers and to judge relative importance of effects.
- *Mental calculation:* The ability to perform and keep track of calculations mentally.
- *Visualisation skill:* The ability to 'see' and manipulate mathematical figures and expressions in the mind. The ability to relate physical objects in the real world to diagrammatic representations and to understand these representations.
- *Logical reasoning:* The ability to create and follow logical proof.
- *Mathematical intuition:* The ability to make mathematical jumps and connections between various areas of mathematics and to guess answers correctly without calculation.

In the same way that an athlete who fully understands his or her own strengths and weaknesses can more effectively train to become a better competitor, a mathematician who understands clearly his or her own skills profile will, with focus and effort, become a better mathematician. In the following chapters of this book we discuss a very wide range of ideas in mathematics, using a variety of skills and approaches as required. Sometimes the presentation will be working in one of your preferred modes and other times in one of your less developed modes. It is important to note that none is 'the best way'; they all have their place and their power in the mathematical universe.

1.5 Mathematical Problem Solving

Solving a mathematical problem is rather similar to completing a journey: one arrives at a destination after a period of travelling from a starting point. On a journey, the travelling may be simple or arduous, long or short; the start or end might be familiar or unfamiliar, very specific or quite general. In the same way that it is possible to prepare for a journey between two places it is possible to prepare for mathematical problem solving and to use certain general problem solving techniques to help you to reach the solution. In this section we shall explore these ideas. The philosophy is that by focussing explicitly on problem solving skills in a structured way, mathematicians are able to achieve a higher order of mathematical thought than might otherwise be possible. We shall break this procedure down into questions to ask *before*, *during* and *after* the problem solving journey. We shall highlight the points in question with some simple examples.

1.5.1 *Before starting to solve the problem*

Before embarking on any big journey in life, preparation is key. Do you know exactly where you are going? Do you need to make any special preparations before starting? Have you studied and planned your route? Solving problems in mathematics is no different: preparing properly before attempting a solution will undoubtedly save time and effort. Don't start writing immediately; first you need to stop and think[24] :

- Do you understand precisely the meaning of *all* the information presented in the problem?
- Do you know precisely what you are trying to solve?

If the answer to either of these questions is 'no', then you might need to look up definitions, refresh your knowledge of various concepts or add extra constraints to the problem. In more difficult areas of mathematics, you might not at the outset be able to obtain complete clarity; this might only emerge following engagement with the problem. In this case, you need to be as clear as possible where the holes in your understanding are. With this in mind, you might be able to fill in gaps along the way. In short, before starting to solve a problem you need to be able to answer this question:

- *What do I know? What do I need to find out? What am I uncertain about?*

Let us see this in action with three examples.

Example: Show that the solutions (x, y) to these simultaneous equations are pairs of complex numbers:

$$y = x^2 - 3x + 27$$
$$y = -x^2 + x - 16$$

[24]As a rule you can expect every single word used by a mathematician to have been carefully selected. A good technique is to read a question or piece of difficult mathematics several times, each time putting the emphasis on a different word. Why, specifically, has that word been chosen? What does it mean in the context of the mathematical scenario?

What do we know? We are provided with two **simultaneous equations** involving the letters x and y. We are asked to show that the **solutions** are **pairs of complex numbers**. We need to know the meaning of all the terms in bold face before we can understand properly the meaning of the question. Now, what are we asked to find out? We need to *show that the solutions are complex numbers.* Notice that the question is not asking us to find the solutions, merely to demonstrate that they are complex.

Example: A is a matrix with integer elements and negative determinant. $A^2 = \begin{pmatrix} 7 & 0 \\ 0 & 7 \end{pmatrix}$. Find the determinant of A^3.

In this problem I am told that A is a **matrix**; that A has **integer elements** with **negative determinant** and that A^2 has a certain form. The technical points here are highlighted in bold. If I do not know what these mean then I need to look them up. The question asks me to find the *determinant* of A^3. Note that it does not actually ask me to find the matrix A, although finding this might be a first instinctive response to starting the solution to the problem.

Example: You are provided with a large quantity of numerical data recorded from a measuring device and asked if it shows bias.

In this problem, although the inputs are a clear set of data, the required output is not clear. In order to answer the question, we would need to define mathematically some concept of the word **bias**. There are many uncertainties facing us in this question. This might require some work investigating the nature of the physical measurements and why and how these measurements were to be used. Some concept of errors in measurement would need to be developed. Appropriate statistical tools would need to be chosen, at which point a full analysis of the data could take place.

Once you have assessed the start and end points of the problem solving journey, you will need to give a preliminary assessment of the journey itself:

- *What are the key features of my problem?*

The assessment of the key features of a problem is very likely to help to determine the mathematics that will be required to solve the problem. Deciding upon the key features is more than simply noting the information given in the question. It is about weighing up aspects of the problem which are likely to be significant in some way during the solution of the problem. Determining the key features of a problem can be something of an art-form, rather like a detective at a messy crime scene deciding which pieces of evidence are important and which pieces are irrelevant to the case. In short, assessing the key features of a problem involves separating the *relevant* from the *irrelevant* and the *important* from the *minor details*. Making these distinctions will allow us to focus our attention on the heart of the problem.

Example: A bag contains 10 very long pieces of string. Brian is asked to tie the ends together to create one long piece of string. Instead, he reaches into the bag, finds two loose ends randomly and ties these together. He repeats this until there are only two loose ends left. What is the probability that he has created one very long piece of string as asked?

As this question is expressed in words, there is a great deal of 'irrelevant' information. We might imagine a bag of strings as a complicated mass of cord, but mathematically we are pointed to the fact that the problem involves **10 line segments** which we are to **join at the ends**. We might also try to imagine the problem in its entirely, but need to realise that the whole process breaks down into a sequence of smaller stages. At each stage the only decision involves **linking two randomly selected end-points**. To end up with one long piece of string **we must never randomly join the two end-points on the same line segment**. The key feature of this problem is therefore the number of **free end-points** at each stage of the process.

Example: Factorise this equation:

$$x^5 - x^4 + x^3 - x^2 + x^1 - 1 = 0$$

There are two key features to this equation. The first is that instead of being any polynomial it is a **fifth order** polynomial, and will therefore have **up to five distinct factors**. The second key fact in this question is that the **coefficients are all either** $+1$ **or** -1, suggesting an important regularity to the problem. It would be prudent to begin by considering the values $x = \pm 1$.

Once the problem has been assessed, the problem solving process should begin.

1.5.2 *During the problem solving process*

Long journeys are often broken down into a series of smaller sections. During your journey you will need to stop regularly to check that you are still on course for your destination. You will need to be sure that you are following all of the signs and landmarks available to you. Sometimes you will need to make a U-turn; on occasion you will need to take a detour and at other times you will meet impassable obstacles that might require you to start your journey over again in a completely different direction. Sometimes you will simply be lost and have no idea in which direction to go.

At high-school, mathematical problems often involve just one key idea and application: it is often possible to see through a problem from start to finish immediately. More advanced problems typically involve multiple steps and will be solved in several smaller stages, each with their own individual goal. In this case, it is important to set clear signposts to lead you through the problem solving process by keeping very clear and distinct the concept of the overall goal from the intermediate goals along the way. When stuck on a problem, repeated reference to the following list of ideas is important:

(1) *What mathematics will fit my problem?* Can you represent your problem in an alternative geometrical or algebraic way?
(2) *What solutions will fit my problem?* Can you work backwards by guessing what sort of solutions might fit the problem?
(3) *What pieces will my problem split into?* Can you break the problem up into simpler pieces, and try to solve one of these pieces?
(4) *What simplifications, extensions or comparisons can be made?* Can you substitute, change, extend or alter the parameters or setup of the problem to create an alternative version of the problem to gain intuition? Is it a special or general case of some other problem?

As before the start of the journey, you will still need to keep on asking the following questions:

(5) *What do I know? What do I need to find out? What am I uncertain about?*

The answers to these questions will evolve and update throughout the problem solving process.

(6) *What are the key features of the problem?*

Eventually, you will hopefully arrive at a solution, which takes us to the final part of the journey.

1.5.3 *After the problem solving process*

Once you have reached your solution to a problem, it is very important to check that your solution works properly. This, to the advanced mathematician, is a fundamental part of the problem solving process:

(1) *Does my solution make sense?* Can you give a variety of checks of your solution? Does the solution work for key limits and values of the parameters?
(2) *Is my solution fully consistent?* Is the solution fully consistent with all constraints and requirements laid out in the question?
(3) *Does my solution fully solve the problem?* Are there multiple solutions? Have you satisfied all constraints and requirements?

It is not always possible to check each of these points or to pursue them in detail, but any solution will benefit from as much investigation as possible. This is not simply a matter going over the steps in your existing calculation, it is about probing the solution in other, independent ways. Imagine a prosecution lawyer attempting to poke holes in a solid defence: this is your job in checking the solution; you must test every possible aspect for weaknesses. Here we show two examples of the checking process; one for a numerical problem and one for an algebraic problem.

Example: Suppose that I have solved numerically the equation

$$x + \cos(x) = 0\,,$$

where x is an angle in radians. I have the answer $x = -0.99985$ to 5 decimal places, which appears to work in my calculator.

Checking process: For numerical problems it can be difficult to give an independent verification, but can I check to see whether the solution makes sense? I am suspicious of an answer so close to -1, so let me investigate this. Since x is very close to -1, this means, from my equation, that $\cos(x)$ must be very close to 1. From the graph of $\cos(x)$, I know that this would mean that x must be very close to 0 or $\pm 2\pi$, which is around ± 6. Neither 0 nor ± 6 are close to -1; the answer, therefore, cannot be correct. Further checking shows that I have been using my calculator in the 'degree mode'. After rectifying this error, I can ask whether my single solution fully solves the problem. I can use a graphical representation of the equation to see that there ought to be only one solution to the equation; the problem is therefore fully solved (Fig. 1.5).

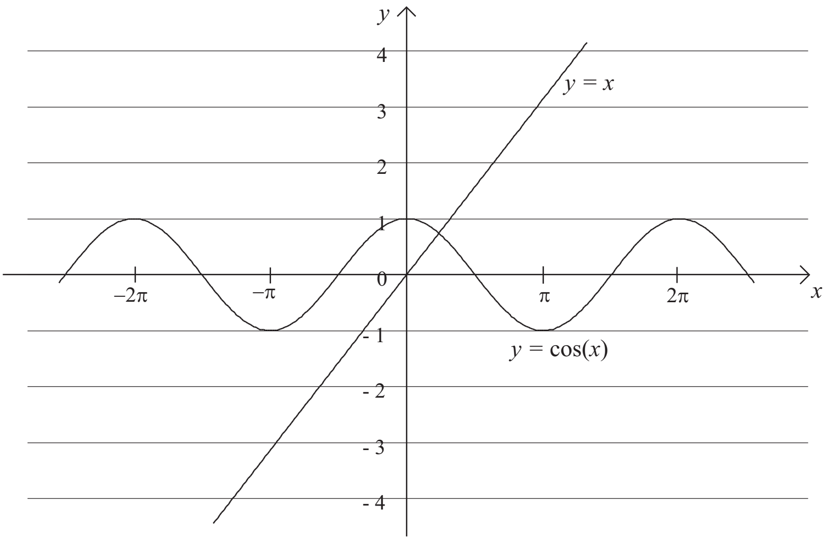

Fig. 1.5 Graphs of $y = x$ and $y = \cos x$ in radians.

Example: Suppose that I have factorised the polynomial

$$p(x) = x^4 - 13x^2 + 36$$

as

$$p(x) = (x + 2)(x + 2)(x + 3)(x + 3)$$

Checking process: Can I find some quick, sensible checks of this factorisation other than expanding out the brackets? Clearly, the factorisation gives the correct constant $36 = 2 \times 2 \times 3 \times 3$ and the correct coefficient for x^4. What other quick checks could I make? The most notable fact about the polynomial is that all of the powers of x are even. This means that $p(x) = +p(-x)$, which is clearly not the case for my factorisation, so it must be wrong.

Example: Suppose that I have been asked to work out a probability $P(n)$ that n buses will arrive at my bus stop over the next hour given that the buses arrive, on average, at a rate of λ per hour. I have found the answer

$$P(n) = e^{-n\lambda}\frac{n\lambda^n}{n!}$$

Checking process: As this is a probability I should certainly have that $P(n)$ is not negative and tends to zero for large values of n. This is certainly true for my expression, so the first check is passed. Now, let me check the $n = 1$ case. This gives $P(1) = \lambda e^{-\lambda}$, which is positive, less than 1 and small for large λ, so this is a sensible probability. What about the other simple value $n = 0$? This gives $P(0) = 0$. There cannot be *zero* chance of no buses arriving, so there must be a mistake in my formula.

1.5.4 *Solving the Fibonacci series*

Let us see the application of these problem solving ideas in action in the full solution of a difficult problem: that of deducing a closed form for the famous *Fibonacci sequence*. Let us imagine following the progress of an exceedingly capable solver as he or she approaches the problem. Whilst the content is clearly fascinating and rich, here we can focus on the process by which the content is naturally discovered in a sequence of steps and ideas.

- The *Fibonacci sequence* is a sequence of numbers $1, 1, 2, 3, 5, 8, 13, \ldots$. The nth term in this sequence is defined to be the sum of the previous two terms. Can you make a conjecture for a closed algebraic expression for the nth term?

Let us see how to break down the solution of this problem into steps:

(1) What do I know? I am told the definition of the sequence in terms of a recursive formula. Notice that the first two terms in this sequence are special, as they cannot be defined in terms of the 'previous two' terms. The first two terms thus start the sequence off. I could imagine another sequence defined in essentially the same way starting with, for example, 1 and 3 or 0 and 2:

$$1, 3, 4, 7, 11, 18, 29, \ldots \qquad 0, 2, 2, 4, 6, 10, 16, 26, \ldots$$

In general, for our sequence we have

$$a_n = a_{n-1} + a_{n-2} \text{ for } (n > 2) \quad a_1 = 1, a_2 = 1$$

(2) What do I need to find out? The problem asks for a 'closed algebraic expression'. I feel confident about the meaning of 'algebraic' expression, but need to clarify the part 'closed'. I notice that in the equation above a_n is expressed algebraically in terms of a_{n-1} and a_{n-2}. Closed in this sense must mean to write a_n as a function of n instead of other terms such as a_{n-1}. I therefore need to find a function $f(n)$ for which

$$a_n = f(n) \quad n = 1, 2, 3, \ldots$$

(3) I will need somehow to 'solve' the recursive formula. I could try to solve the recurrence relation algebraically, but feel that it might be prudent first to get a 'feel' for how the sequence behaves numerically. The first obvious step is to solve explicitly for a few low values of n. I can use my calculator to help here. Let me inspect some of the first few numbers in the sequence and compare them to simple functions of n:

n	1	2	3	4	5	7	10	15	20	30
a_n	1	1	2	3	5	13	55	610	6765	832040
n^2	1	4	9	16	25	49	100	225	400	900
n^3	1	8	27	64	125	343	1000	3375	8000	27000

Can I spot any patterns in the way the numbers grow to give me an idea of a possible form of the function $f(n)$? Roughly speaking, a_n appears to start small and then grow more and more quickly. This suggests that $f(n)$ grows faster than a straight-forward polynomial expression in n.

(4) Since comparison of the sequence with powers of n did not help, other than to eliminate likely possibilities, let us consider next the rate of growth of the series. To do this let us take some ratios of successive terms in the sequence:

n	2	3	4	5	6	7	8	9	10
$\frac{a_n}{a_{n-1}}$	1	2	1.5	1.667	1.6	1.625	1.615	1.619	1.618

The ratios of the terms appear to be settling down to a limit[25] of $1.618\ldots$. A quick check with a larger value of n appears to confirm this. I would conjecture that the limit of the ratio is indeed a fixed number $k = 1.618\ldots$; if true, this would help me to determine the possible form of $f(n)$.

[25] Limits will be properly covered in the Analysis chapter. For now, think of it as the ratios settling down to, or tending to, a fixed value for very large values of n.

(5) Let me try the approach of algebraically solving the relationship. I'll first try to reduce a few terms to try to spot any patterns emerging:

$$\begin{aligned} a_n &= a_{n-1} + a_{n-2} \\ &= (a_{n-2} + a_{n-3}) + (a_{n-3} + a_{n-4}) \\ &= a_{n-2} + a_{n-3} + (a_{n-4} + a_{n-5}) + (a_{n-5} + a_{n-6}) \\ &= a_{n-2} + a_{n-3} + a_{n-4} + a_{n-5} + (a_{n-6} + a_{n-7}) + (a_{n-7} + a_{n-8}) \end{aligned}$$

In writing these expressions I bracketed off the terms where I used the relationship defining the sequence to make the process seem clearer. Looking at these, I notice that this process can be continued down to the term a_1

$$a_n = a_{n-2} + a_{n-3} + a_{n-4} + \cdots + a_4 + a_3 + 2a_2 + a_1 \quad (n \geq 6)$$

Interestingly, the term a_2 is singled out in this expression as the only one with a coefficient not equal to 1; furthermore this expression holds for generalisations of the sequence with any initial values of a_1 and a_2.

$$a_n = a_2 + \sum_{k=1}^{n-2} a_k$$

Now, let me review where I have got to. I have determined an expression for a_n in terms of a sum of previous terms in the sequence. I am trying to find $a_n = f(n)$ for some closed function. I do not yet have a_n in this form; in fact this and other algebraic manipulations do not seem to be helping me to determine $f(n)$. I need to try a different approach.

(6) I determined that the ratio of terms in the sequence appeared to tend to a fixed number, which I called k. Can I find any more details about this number? As I conjectured that the limit of the ratio of successive terms in the sequence *tended* to k I shall therefore have to look at very large values of n to try to glean any information about the value of k. In these cases, we must have $a_n \approx ka_{n-1}$ and $a_{n-1} \approx ka_{n-2}$. Therefore, the relationship $a_n = a_{n-1} + a_{n-2}$ becomes

$$k^2 a_{n-2} \approx ka_{n-2} + a_{n-2}$$

Cancelling out the a_{n-2} provides us with a quadratic equation for k:

$$k^2 - k - 1 \approx 0$$

This has solutions

$$k \approx \frac{1 \pm \sqrt{5}}{2}$$

Numerically I can see that $k \approx 1.61803\ldots$ or $k \approx -0.61803\ldots$, which I recognise from my initial numerical investigation; the positive value is the one I require. This adds evidence in favour of my conjecture. Therefore, if the ratio of the terms tend to a fixed number, then I conjecture that the number will be exactly equal to

$$k = \frac{1 + \sqrt{5}}{2}$$

The plan appears to be building momentum. Let us use this conjectured form of k to try to build a form for $f(n)$.

(7) Since the ratio of subsequent terms tends to k, let us speculate that the nth term in the sequence will look something like

$$a_n = A(n)\left(\frac{1+\sqrt{5}}{2}\right)^n + B(n)\,,$$

where $A(n)$ and $B(n)$ are functions of n which tend to a constant and zero respectively for very large values of n. Can I make progress with this form of the a_n?

(8) Let me recap my position. I have suggested a particular candidate for a closed form for a_n on the assumption that the ratio of subsequent terms tends to a fixed number k for large n. From the defining relationship $a_n = a_{n-1} + a_{n-2}$ and $a_1 = 1$, $a_2 = 1$ it is clear that each term a_n must be a *whole number*. However, the speculative form for a_n involves the term $(1+\sqrt{5})^n$, which is never a whole number; somehow we will need to arrange a cancellation of the $\sqrt{5}$ terms. Expanding out these brackets using the binomial theorem I will get the sum of a whole number and a multiple of $\sqrt{5}$:

$$(1+\sqrt{5})^n = 1+n\sqrt{5}+\frac{n(n-1)}{2}(\sqrt{5})^2+\cdots+n(\sqrt{5})^{n-1}+(\sqrt{5})^n = a+b\sqrt{5}\,,$$

for two whole numbers a and b. Therefore, we deduce that

$$a_n = \frac{A(n)}{2^n}(a + b\sqrt{5}) + B(n)$$

For this form of a_n to work it must provide a whole number for every value of n; therefore, the part proportional to $\sqrt{5}$ must vanish for every value of n. The problem in doing this is that the coefficients a and b of the sum in the binomial expansion are complicated expressions which vary for each n. How are we to deal with these? After some thought it seems that the best way to proceed is to use $B(n)$ to add or subtract a similar binomial sum, leaving an expression which is either a whole number or a mutiple of $\sqrt{5}$. The resulting possible form for a_n is

$$\begin{aligned} a_n = A_+(n)\left[\left(\tfrac{1+\sqrt{5}}{2}\right)^n + \left(\tfrac{1-\sqrt{5}}{2}\right)^n\right] \\ +A_-(n)\left[\left(\tfrac{1+\sqrt{5}}{2}\right)^n - \left(\tfrac{1-\sqrt{5}}{2}\right)^n\right] \end{aligned}$$

(9) This form for a_n appears to be making some progress. If it is to work then it must, of course, work for the particular values $a_1 = 1$ and $a_2 = 1$. Let us substitute these values to see how they constrain the functions $A_\pm(n)$:

$$\begin{aligned} 1 &= A_+(1)\left[\left(\tfrac{1+\sqrt{5}}{2}\right) + \left(\tfrac{1-\sqrt{5}}{2}\right)\right] \\ &\quad +A_-(1)\left[\left(\tfrac{1+\sqrt{5}}{2}\right) - \left(\tfrac{1-\sqrt{5}}{2}\right)\right] \\ &= 1A_+(1) + \sqrt{5}A_-(1) \end{aligned}$$

and

$$\begin{aligned} 1 &= A_+(2)\left[\left(\tfrac{1+\sqrt{5}}{2}\right)^2 + \left(\tfrac{1-\sqrt{5}}{2}\right)^2\right] \\ &\quad +A_-(2)\left[\left(\tfrac{1+\sqrt{5}}{2}\right)^2 - \left(\tfrac{1-\sqrt{5}}{2}\right)^2\right] \\ &= 3A_+(2) + \sqrt{5}A_-(2) \end{aligned}$$

Let us look at these expressions. I cannot yet see any obvious patterns in the numbers, so let us look at a few more terms in the sequence. Calculation for $a_3 = 2$, $a_4 = 3$, $a_5 = 5$ and $a_6 = 8$ give

$$\begin{aligned} a_1 &= 1 = A_+(1) + \sqrt{5}A_-(1) \\ a_2 &= 1 = 3A_+(2) + \sqrt{5}A_-(2) \\ a_3 &= 2 = 4A_+(3) + 2\sqrt{5}A_-(3) \\ a_4 &= 3 = 7A_+(3) + 3\sqrt{5}A_-(4) \\ a_5 &= 5 = 11A_+(5) + 5\sqrt{5}A_-(5) \\ a_6 &= 8 = 18A_+(5) + 8\sqrt{5}A_-(6) \end{aligned}$$

Now I can see two patterns: the coefficients in front of $\sqrt{5}A_-(n)$ appear to be equal to a_n, whereas the coefficients in front of $A_+(n)$ appear to increase like a Fibonacci sequence starting with 1 and 3. Focussing on the simplicity of the coefficients of the terms $A_-(n)$, I would feel confident in making the conjecture that $A_-(n) = \frac{1}{\sqrt{5}}$ and $A_+(n) = 0$ for every n. This would provide me with this form for a_n:

$$a_n = \frac{1}{\sqrt{5}}\left[\left(\frac{1+\sqrt{5}}{2}\right)^n - \left(\frac{1-\sqrt{5}}{2}\right)^n\right]$$

(10) I now feel that I am on firm mathematical ground: I have created a clear, precise, well-justified conjecture for an algebraic form for a_n. It looks like I would be able to test this conjecture using an induction argument. This will be a difficult challenge in itself, but is a clear, well-defined mathematical problem.[26]

1.5.5 *Mathematical problem solving summary*

As we have seen in the previous extended example, there are many different facets of mathematical problem solving. Sometimes you will see your way through to a solution or process by which you know the problem will be solved almost immediately; at other times the direction will be less obvious. It is on these occasions that focussing on explicit problem solving skills will be of great assistance. For convenience, here we summarise the set of key ideas from this section:

(1) *What do I know? What do I need to find out? What am I uncertain about?*
(2) *What are the key features of my problem?*
(3) *What mathematics will fit my problem?*
(4) *What solutions will fit my problem?*
(5) *What pieces will my problem split into?*
(6) *What simplifications, extensions or comparisons can be made?*
(7) *Does my solution make sense?*
(8) *Is my solution fully consistent?*
(9) *Does my solution fully solve the problem?*

Engaging in the problem solving process requires a mixture of experimentation, inspiration and application of algorithms or procedures. We shall use these ideas at all times throughout the rest of the book as we explore the world of higher mathematics.

As a final comment before the journey continues into the territory of numbers, note that curiosity is one of the main characteristics of the best

[26]Note that this form is indeed correct, and can be proved by induction.

mathematicians. You should continually challenge and reflect as you learn and explore:

- *Do I believe this piece of mathematics? How does it fit with other mathematics I know? What have I learnt? What can I do next? What questions has this mathematics raised? Would parts warrant further investigation? Can I extend or generalise the problem I have just solved or theory I have just encountered?*

Engagement in mathematics in this rich way will provide rich intellectual rewards. Enjoy the journey.