nation based on regurgitating extracts from the book. (Indeed, in my own examinations I gave a supplemental sheet listing the key definitions and theorems which were relevant to the examination problems.) Making the examinations similar to the homework assigned in the course will also help motivate the students to work through and understand their homework problems as thoroughly as possible (as opposed to, say, using flash cards or other such devices to memorize material), which is good preparation not only for examinations but for doing mathematics in general.

Some of the material in this textbook is somewhat peripheral to the main theme and may be omitted for reasons of time constraints. For instance, as set theory is not as fundamental to analysis as are the number systems, the chapters on set theory (Chapters 3, 8) can be covered more quickly and with substantially less rigour, or be given as reading assignments. The appendices on logic and the decimal system are intended as optional or supplemental reading and would probably not be covered in the main course lectures; the appendix on logic is particularly suitable for reading concurrently with the first few chapters. Also, Chapter 16 (on Fourier series) is not needed elsewhere in the text and can be omitted.

For reasons of length, this textbook has been split into two volumes. The first volume is slightly longer, but can be covered in about thirty lectures if the peripheral material is omitted or abridged. The second volume refers at times to the first, but can also be taught to students who have had a first course in analysis from other sources. It also takes about thirty lectures to cover.

I am deeply indebted to my students, who over the progression of the real analysis course corrected several errors in the lectures notes from which this text is derived, and gave other valuable feedback. I am also very grateful to the many anonymous referees who made several corrections and suggested many important improvements to the text.

Terence Tao

Chapter 1

Introduction

1.1 What is analysis?

This text is an honours-level undergraduate introduction to *real analysis*: the analysis of the real numbers, sequences and series of real numbers, and real-valued functions. This is related to, but is distinct from, *complex analysis*, which concerns the analysis of the complex numbers and complex functions, *harmonic analysis*, which concerns the analysis of harmonics (waves) such as sine waves, and how they synthesize other functions via the Fourier transform, *functional analysis*, which focuses much more heavily on functions (and how they form things like vector spaces), and so forth. *Analysis* is the rigourous study of such objects, with a focus on trying to pin down precisely and accurately the qualitative and quantitative behavior of these objects. Real analysis is the theoretical foundation which underlies *calculus*, which is the collection of computational algorithms which one uses to manipulate functions.

In this text we will be studying many objects which will be familiar to you from freshman calculus: numbers, sequences, series, limits, functions, definite integrals, derivatives, and so forth. You already have a great deal of experience of *computing* with these objects; however here we will be focused more on the underlying theory for these objects. We will be concerned with questions such as the following:

1. What is a real number? Is there a largest real number? After 0, what is the "next" real number (i.e., what is the smallest positive real number)? Can you cut a real number into pieces infinitely many times? Why does a number such as 2 have a square root, while a number such as -2 does not? If there are infinitely many reals and infinitely many rationals, how come there are "more" real numbers than rational numbers?

2. How do you take the limit of a sequence of real numbers? Which sequences have limits and which ones don't? If you can stop a sequence from escaping to infinity, does this mean that it must eventually settle down and converge? Can you add infinitely many real numbers together and still get a finite real number? Can you add infinitely many rational numbers together and end up with a non-rational number? If you rearrange the elements of an infinite sum, is the sum still the same?

3. What is a function? What does it mean for a function to be continuous? differentiable? integrable? bounded? can you add infinitely many functions together? What about taking limits of sequences of functions? Can you differentiate an infinite series of functions? What about integrating? If a function $f(x)$ takes the value 3 when $x = 0$ and 5 when $x = 1$ (i.e., $f(0) = 3$ and $f(1) = 5$), does it have to take every intermediate value between 3 and 5 when x goes between 0 and 1? Why?

You may already know how to answer some of these questions from your calculus classes, but most likely these sorts of issues were only of secondary importance to those courses; the emphasis was on getting you to perform computations, such as computing the integral of $x \sin(x^2)$ from $x = 0$ to $x = 1$. But now that you are comfortable with these objects and already know how to do all the computations, we will go back to the theory and try to *really* understand what is going on.

1.2 Why do analysis?

It is a fair question to ask, "why bother?", when it comes to analysis. There is a certain philosophical satisfaction in knowing *why* things work, but a pragmatic person may argue that one only needs to know *how* things work to do real-life problems. The calculus training you receive in introductory classes is certainly adequate for you to begin solving many problems in physics, chemistry, biology, economics, computer science, finance, engineering, or whatever else you end up doing - and you can certainly use things like the chain rule, L'Hôpital's rule, or integration by parts without knowing why these rules work, or whether there are any exceptions to these rules. However, one can get into trouble if one applies rules without knowing where they came from and what the limits of their applicability are. Let me give some examples in which several of these familiar rules, if applied blindly without knowledge of the underlying analysis, can lead to disaster.

Example 1.2.1 (Division by zero). This is a very familiar one to you: the cancellation law $ac = bc \implies a = b$ does not work when $c = 0$. For instance, the identity $1 \times 0 = 2 \times 0$ is true, but if one blindly cancels the 0 then one obtains $1 = 2$, which is false. In this case it was obvious that one was dividing by zero; but in other cases it can be more hidden.

Example 1.2.2 (Divergent series). You have probably seen geometric series such as the infinite sum

$$S = 1 + \frac{1}{2} + \frac{1}{4} + \frac{1}{8} + \frac{1}{16} + \ldots.$$

You have probably seen the following trick to sum this series: if we call the above sum S, then if we multiply both sides by 2, we obtain

$$2S = 2 + 1 + \frac{1}{2} + \frac{1}{4} + \frac{1}{8} + \ldots = 2 + S$$

and hence $S = 2$, so the series sums to 2. However, if you apply the same trick to the series

$$S = 1 + 2 + 4 + 8 + 16 + \ldots$$

one gets nonsensical results:

$$2S = 2 + 4 + 8 + 16 + \ldots = S - 1 \implies S = -1.$$

So the same reasoning that shows that $1 + \frac{1}{2} + \frac{1}{4} + \ldots = 2$ also gives that $1 + 2 + 4 + 8 + \ldots = -1$. Why is it that we trust the first equation but not the second? A similar example arises with the series

$$S = 1 - 1 + 1 - 1 + 1 - 1 + \ldots;$$

we can write

$$S = 1 - (1 - 1 + 1 - 1 + \ldots) = 1 - S$$

and hence that $S = 1/2$; or instead we can write

$$S = (1 - 1) + (1 - 1) + (1 - 1) + \ldots = 0 + 0 + \ldots$$

and hence that $S = 0$; or instead we can write

$$S = 1 + (-1 + 1) + (-1 + 1) + \ldots = 1 + 0 + 0 + \ldots$$

and hence that $S = 1$. Which one is correct? (See Exercise 7.2.1 for an answer.)

Example 1.2.3 (Divergent sequences). Here is a slight variation of the previous example. Let x be a real number, and let L be the limit

$$L = \lim_{n \to \infty} x^n.$$

Changing variables $n = m + 1$, we have

$$L = \lim_{m+1 \to \infty} x^{m+1} = \lim_{m+1 \to \infty} x \times x^m = x \lim_{m+1 \to \infty} x^m.$$

But if $m + 1 \to \infty$, then $m \to \infty$, thus

$$\lim_{m+1 \to \infty} x^m = \lim_{m \to \infty} x^m = \lim_{n \to \infty} x^n = L,$$

and thus

$$xL = L.$$

At this point we could cancel the L's and conclude that $x = 1$ for an arbitrary real number x, which is absurd. But since we are already aware of the division by zero problem, we could be a little smarter and conclude instead that either $x = 1$, or $L = 0$. In particular we seem to have shown that

$$\lim_{n\to\infty} x^n = 0 \text{ for all } x \neq 1.$$

But this conclusion is absurd if we apply it to certain values of x, for instance by specializing to the case $x = 2$ we could conclude that the sequence $1, 2, 4, 8, \ldots$ converges to zero, and by specializing to the case $x = -1$ we conclude that the sequence $1, -1, 1, -1, \ldots$ also converges to zero. These conclusions appear to be absurd; what is the problem with the above argument? (See Exercise 6.3.4 for an answer.)

Example 1.2.4 (Limiting values of functions). Start with the expression $\lim_{x\to\infty} \sin(x)$, make the change of variable $x = y + \pi$ and recall that $\sin(y + \pi) = -\sin(y)$ to obtain

$$\lim_{x\to\infty} \sin(x) = \lim_{y+\pi\to\infty} \sin(y + \pi) = \lim_{y\to\infty} (-\sin(y)) = -\lim_{y\to\infty} \sin(y).$$

Since $\lim_{x\to\infty} \sin(x) = \lim_{y\to\infty} \sin(y)$ we thus have

$$\lim_{x\to\infty} \sin(x) = -\lim_{x\to\infty} \sin(x)$$

and hence

$$\lim_{x\to\infty} \sin(x) = 0.$$

If we then make the change of variables $x = \pi/2 - z$ and recall that $\sin(\pi/2 - 2) = \cos(z)$ we conclude that

$$\lim_{x\to\infty} \cos(x) = 0.$$

Squaring both of these limits and adding we see that

$$\lim_{x\to\infty} (\sin^2(x) + \cos^2(x)) = 0^2 + 0^2 = 0.$$

On the other hand, we have $\sin^2(x) + \cos^2(x) = 1$ for all x. Thus we have shown that $1 = 0$! What is the difficulty here?

Example 1.2.5 (Interchanging sums). Consider the following fact of arithmetic. Consider any matrix of numbers, e.g.

$$\begin{pmatrix} 1 & 2 & 3 \\ 4 & 5 & 6 \\ 7 & 8 & 9 \end{pmatrix}$$

and compute the sums of all the rows and the sums of all the columns, and then total all the row sums and total all the column sums. In both cases you will get the same number - the total sum of all the entries in the matrix:

$$\begin{array}{cccc} \begin{pmatrix} 1 & 2 & 3 \\ 4 & 5 & 6 \\ 7 & 8 & 9 \end{pmatrix} & & & \begin{matrix} 6 \\ 15 \\ 24 \end{matrix} \\ 12 & 15 & 18 & 45 \end{array}$$

To put it another way, if you want to add all the entries in an $m \times n$ matrix together, it doesn't matter whether you sum the rows first or sum the columns first, you end up with the same answer. (Before the invention of computers, accountants and book-keepers would use this fact to guard against making errors when balancing their books.) In series notation, this fact would be expressed as

$$\sum_{i=1}^{m} \sum_{j=1}^{n} a_{ij} = \sum_{j=1}^{n} \sum_{i=1}^{m} a_{ij},$$

if a_{ij} denoted the entry in the i^{th} row and j^{th} column of the matrix.

Now one might think that this rule should extend easily to infinite series:

$$\sum_{i=1}^{\infty} \sum_{j=1}^{\infty} a_{ij} = \sum_{j=1}^{\infty} \sum_{i=1}^{\infty} a_{ij}.$$

Indeed, if you use infinite series a lot in your work, you will find yourself having to switch summations like this fairly often. Another way of saying this fact is that in an infinite matrix, the sum of the row-totals should equal the sum of the column-totals.

However, despite the reasonableness of this statement, it is actually false! Here is a counterexample:

$$\begin{pmatrix} 1 & 0 & 0 & 0 & \dots \\ -1 & 1 & 0 & 0 & \dots \\ 0 & -1 & 1 & 0 & \dots \\ 0 & 0 & -1 & 1 & \dots \\ 0 & 0 & 0 & -1 & \dots \\ \vdots & \vdots & \vdots & \vdots & \ddots \end{pmatrix}$$

If you sum up all the rows, and then add up all the row totals, you get 1; but if you sum up all the columns, and add up all the column totals, you get 0! So, does this mean that summations for infinite series should not be swapped, and that any argument using such a swapping should be distrusted? (See Theorem 8.2.2 for an answer.)

Example 1.2.6 (Interchanging integrals). The interchanging of integrals is a trick which occurs in mathematics just as commonly as the interchanging of sums. Suppose one wants to compute the volume under a surface $z = f(x,y)$ (let us ignore the limits of integration for the moment). One can do it by slicing parallel to the x-axis: for each fixed value of y, we can compute an area $\int f(x,y)\ dx$, and then we integrate the area in the y variable to obtain the volume

$$V = \int \int f(x,y) dx dy.$$

Or we could slice parallel to the y-axis for each fixed x and compute an area $\int f(x,y)\ dy$, and then integrate in the x-axis to obtain

$$V = \int \int f(x,y) dy dx.$$

This seems to suggest that one should always be able to swap integral signs:

$$\int \int f(x,y)\ dx dy = \int \int f(x,y)\ dy dx.$$

And indeed, people swap integral signs all the time, because sometimes one variable is easier to integrate in first than the other. However, just as infinite sums sometimes cannot be swapped, integrals are also sometimes dangerous to swap. An example is with the integrand $e^{-xy} - xye^{-xy}$. Suppose we believe that we can swap the integrals:

$$\int_0^\infty \int_0^1 (e^{-xy} - xye^{-xy})\, dy\, dx = \int_0^1 \int_0^\infty (e^{-xy} - xye^{-xy})\, dx\, dy.$$

Since

$$\int_0^1 (e^{-xy} - xye^{-xy})\, dy = ye^{-xy}|_{y=0}^{y=1} = e^{-x},$$

the left-hand side is $\int_0^\infty e^{-x}\, dx = -e^{-x}|_0^\infty = 1$. But since

$$\int_0^\infty (e^{-xy} - xye^{-xy})\, dx = xe^{-xy}|_{x=0}^{x=\infty} = 0,$$

the right-hand side is $\int_0^1 0\, dx = 0$. Clearly $1 \neq 0$, so there is an error somewhere; but you won't find one anywhere except in the step where we interchanged the integrals. So how do we know when to trust the interchange of integrals? (See Theorem 19.5.1 for a partial answer.)

Example 1.2.7 (Interchanging limits)**.** Suppose we start with the plausible looking statement

$$\lim_{x\to 0} \lim_{y\to 0} \frac{x^2}{x^2+y^2} = \lim_{y\to 0} \lim_{x\to 0} \frac{x^2}{x^2+y^2}. \tag{1.1}$$

But we have

$$\lim_{y\to 0} \frac{x^2}{x^2+y^2} = \frac{x^2}{x^2+0^2} = 1,$$

so the left-hand side of (1.1) is 1; on the other hand, we have

$$\lim_{x\to 0} \frac{x^2}{x^2+y^2} = \frac{0^2}{0^2+y^2} = 0,$$

so the right-hand side of (1.1) is 0. Since 1 is clearly not equal to zero, this suggests that interchange of limits is untrustworthy. But are there any other circumstances in which the interchange of limits is legitimate? (See Exercise 13.2.9 for a partial answer.)

Example 1.2.8 (Interchanging limits, again). Consider the plausible looking statement

$$\lim_{x\to 1^-} \lim_{n\to\infty} x^n = \lim_{n\to\infty} \lim_{x\to 1^-} x^n$$

where the notation $x \to 1^-$ means that x is approaching 1 from the left. When x is to the left of 1, then $\lim_{n\to\infty} x^n = 0$, and hence the left-hand side is zero. But we also have $\lim_{x\to 1^-} x^n = 1$ for all n, and so the right-hand side limit is 1. Does this demonstrate that this type of limit interchange is always untrustworthy? (See Proposition 14.3.3 for an answer.)

Example 1.2.9 (Interchanging limits and integrals). For any real number y, we have

$$\int_{-\infty}^{\infty} \frac{1}{1+(x-y)^2}\, dx = \arctan(x-y)|_{x=-\infty}^{\infty} = \frac{\pi}{2} - (-\frac{\pi}{2}) = \pi.$$

Taking limits as $y \to \infty$, we should obtain

$$\int_{-\infty}^{\infty} \lim_{y\to\infty} \frac{1}{1+(x-y)^2}\, dx = \lim_{y\to\infty} \int_{-\infty}^{\infty} \frac{1}{1+(x-y)^2}\, dx = \pi.$$

But for every x, we have $\lim_{y\to\infty} \frac{1}{1+(x-y)^2} = 0$. So we seem to have concluded that $0 = \pi$. What was the problem with the above argument? Should one abandon the (very useful) technique of interchanging limits and integrals? (See Theorem 14.6.1 for a partial answer.)

Example 1.2.10 (Interchanging limits and derivatives). Observe that if $\varepsilon > 0$, then

$$\frac{d}{dx}\left(\frac{x^3}{\varepsilon^2+x^2}\right) = \frac{3x^2(\varepsilon^2+x^2) - 2x^4}{(\varepsilon^2+x^2)^2}$$

and in particular that

$$\frac{d}{dx}\left(\frac{x^3}{\varepsilon^2+x^2}\right)|_{x=0} = 0.$$

Taking limits as $\varepsilon \to 0$, one might then expect that

$$\frac{d}{dx}\left(\frac{x^3}{0+x^2}\right)|_{x=0} = 0.$$

But the right-hand side is $\frac{d}{dx}x = 1$. Does this mean that it is always illegitimate to interchange limits and derivatives? (See Theorem 14.7.1 for an answer.)

Example 1.2.11 (Interchanging derivatives). Let[1] $f(x,y)$ be the function $f(x,y) := \frac{xy^3}{x^2+y^2}$. A common maneuvre in analysis is to interchange two partial derivatives, thus one expects

$$\frac{\partial^2 f}{\partial x \partial y}(0,0) = \frac{\partial^2 f}{\partial y \partial x}(0,0).$$

But from the quotient rule we have

$$\frac{\partial f}{\partial y}(x,y) = \frac{3xy^2}{x^2+y^2} - \frac{2xy^4}{(x^2+y^2)^2}$$

and in particular

$$\frac{\partial f}{\partial y}(x,0) = \frac{0}{x^2} - \frac{0}{x^4} = 0.$$

Thus

$$\frac{\partial^2 f}{\partial x \partial y}(0,0) = 0.$$

On the other hand, from the quotient rule again we have

$$\frac{\partial f}{\partial x}(x,y) = \frac{y^3}{x^2+y^2} - \frac{2x^2y^3}{(x^2+y^2)^2}$$

and hence

$$\frac{\partial f}{\partial x}(0,y) = \frac{y^3}{y^2} - \frac{0}{y^4} = y.$$

[1] One might object that this function is not defined at $(x,y) = (0,0)$, but if we set $f(0,0) := (0,0)$ then this function becomes continuous and differentiable for all (x,y), and in fact both partial derivatives $\frac{\partial f}{\partial x}$, $\frac{\partial f}{\partial y}$ are also continuous and differentiable for all (x,y)!

Thus

$$\frac{\partial^2 f}{\partial x \partial y}(0,0) = 1.$$

Since $1 \neq 0$, we thus seem to have shown that interchange of derivatives is untrustworthy. But are there any other circumstances in which the interchange of derivatives is legitimate? (See Theorem 17.5.4 and Exercise 17.5.1 for some answers.)

Example 1.2.12 (L'Hôpital's rule). We are all familiar with the beautifully simple L'Hôpital's rule

$$\lim_{x \to x_0} \frac{f(x)}{g(x)} = \lim_{x \to x_0} \frac{f'(x)}{g'(x)},$$

but one can still get led to incorrect conclusions if one applies it incorrectly. For instance, applying it to $f(x) := x$, $g(x) := 1 + x$, and $x_0 := 0$ we would obtain

$$\lim_{x \to 0} \frac{x}{1+x} = \lim_{x \to 0} \frac{1}{1} = 1,$$

but this is the incorrect answer, since $\lim_{x \to 0} \frac{x}{1+x} = \frac{0}{1+0} = 0$. Of course, all that is going on here is that L'Hôpital's rule is only applicable when both $f(x)$ and $g(x)$ go to zero as $x \to x_0$, a condition which was violated in the above example. But even when $f(x)$ and $g(x)$ do go to zero as $x \to x_0$ there is still a possibility for an incorrect conclusion. For instance, consider the limit

$$\lim_{x \to 0} \frac{x^2 \sin(x^{-4})}{x}.$$

Both numerator and denominator go to zero as $x \to 0$, so it seems pretty safe to apply L'Hôpital's rule, to obtain

$$\begin{aligned} \lim_{x \to 0} \frac{x^2 \sin(x^{-4})}{x} &= \lim_{x \to 0} \frac{2x \sin(x^{-4}) - 4x^{-3} \cos(x^{-4})}{1} \\ &= \lim_{x \to 0} 2x \sin(x^{-4}) - \lim_{x \to 0} 4x^{-3} \cos(x^{-4}). \end{aligned}$$

The first limit converges to zero by the squeeze test (since the function $2x \sin(x^{-4})$ is bounded above by $2|x|$ and below by $-2|x|$,

both of which go to zero at 0). But the second limit is divergent (because x^{-3} goes to infinity as $x \to 0$, and $\cos(x^{-4})$ does not go to zero). So the limit $\lim_{x\to 0} \frac{2x\sin(x^{-4})-4x^{-2}\cos(x^{-4})}{1}$ diverges. One might then conclude using L'Hôpital's rule that $\lim_{x\to 0} \frac{x^2\sin(x^{-4})}{x}$ also diverges; however we can clearly rewrite this limit as $\lim_{x\to 0} x\sin(x^{-4})$, which goes to zero when $x \to 0$ by the squeeze test again. This does not show that L'Hôpital's rule is untrustworthy (indeed, it is quite rigourous; see Section 10.5), but it still requires some care when applied.

Example 1.2.13 (Limits and lengths). When you learn about integration and how it relates to the area under a curve, you were probably presented with some picture in which the area under the curve was approximated by a bunch of rectangles, whose area was given by a Riemann sum, and then one somehow "took limits" to replace that Riemann sum with an integral, which then presumably matched the actual area under the curve. Perhaps a little later, you learnt how to compute the length of a curve by a similar method - approximate the curve by a bunch of line segments, compute the length of all the line segments, then take limits again to see what you get.

However, it should come as no surprise by now that this approach also can lead to nonsense if used incorrectly. Consider the right-angled triangle with vertices $(0,0)$, $(1,0)$, and $(0,1)$, and suppose we wanted to compute the length of the hypotenuse of this triangle. Pythagoras' theorem tells us that this hypotenuse has length $\sqrt{2}$, but suppose for some reason that we did not know about Pythagoras' theorem, and wanted to compute the length using calculus methods. Well, one way to do so is to approximate the hypotenuse by horizontal and vertical edges. Pick a large number N, and approximate the hypotenuse by a "staircase" consisting of N horizontal edges of equal length, alternating with N vertical edges of equal length. Clearly these edges all have length $1/N$, so the total length of the staircase is $2N/N = 2$. If one takes limits as N goes to infinity, the staircase clearly approaches the hypotenuse, and so in the limit we should get the length of the

hypotenuse. However, as $N \to \infty$, the limit of $2N/N$ is 2, not $\sqrt{2}$, so we have an incorrect value for the length of the hypotenuse. How did this happen?

The analysis you learn in this text will help you resolve these questions, and will let you know when these rules (and others) are justified, and when they are illegal, thus separating the useful applications of these rules from the nonsense. Thus they can prevent you from making mistakes, and can help you place these rules in a wider context. Moreover, as you learn analysis you will develop an "analytical way of thinking", which will help you whenever you come into contact with any new rules of mathematics, or when dealing with situations which are not quite covered by the standard rules, For instance, what if your functions are complex-valued instead of real-valued? What if you are working on the sphere instead of the plane? What if your functions are not continuous, but are instead things like square waves and delta functions? What if your functions, or limits of integration, or limits of summation, are occasionally infinite? You will develop a sense of *why* a rule in mathematics (e.g., the chain rule) works, how to adapt it to new situations, and what its limitations (if any) are; this will allow you to apply the mathematics you have already learnt more confidently and correctly.

Chapter 2

Starting at the beginning: the natural numbers

In this text, we will review the material you have learnt in high school and in elementary calculus classes, but as rigourously as possible. To do so we will have to begin at the very basics - indeed, we will go back to the concept of *numbers* and what their properties are. Of course, you have dealt with numbers for over ten years and you know how to manipulate the rules of algebra to simplify any expression involving numbers, but we will now turn to a more fundamental issue, which is: *why* do the rules of algebra work at all? For instance, why is it true that $a(b+c)$ is equal to $ab+ac$ for any three numbers a, b, c? This is not an arbitrary choice of rule; it can be proven from more primitive, and more fundamental, properties of the number system. This will teach you a new skill - how to prove complicated properties from simpler ones. You will find that even though a statement may be "obvious", it may not be easy to prove; the material here will give you plenty of practice in doing so, and in the process will lead you to think about *why* an obvious statement really is obvious. One skill in particular that you will pick up here is the use of *mathematical induction*, which is a basic tool in proving things in many areas of mathematics.

So in the first few chapters we will re-acquaint you with various number systems that are used in real analysis. In increasing order of sophistication, they are the *natural numbers* $\mathbf{N}$; the *integers* $\mathbf{Z}$;

the *rationals* $\mathbf{Q}$, and the *real numbers* $\mathbf{R}$. (There are other number systems such as the *complex numbers* $\mathbf{C}$, but we will not study them until Section 15.6.) The natural numbers $\{0, 1, 2, \ldots\}$ are the most primitive of the number systems, but they are used to build the integers, which in turn are used to build the rationals. Furthermore, the rationals are used to build the real numbers, which are in turn used to build the complex numbers. Thus to begin at the very beginning, we must look at the natural numbers. We will consider the following question: how does one actually *define* the natural numbers? (This is a very different question from how to *use* the natural numbers, which is something you of course know how to do very well. It's like the difference between knowing how to use, say, a computer, versus knowing how to *build* that computer.)

This question is more difficult to answer than it looks. The basic problem is that you have used the natural numbers for so long that they are embedded deeply into your mathematical thinking, and you can make various implicit assumptions about these numbers (e.g., that $a+b$ is always equal to $b+a$) without even aware that you are doing so; it is difficult to let go and try to inspect this number system as if it is the first time you have seen it. So in what follows I will have to ask you to perform a rather difficult task: try to set aside, for the moment, everything you know about the natural numbers; forget that you know how to count, to add, to multiply, to manipulate the rules of algebra, etc. We will try to introduce these concepts one at a time and identify explicitly what our assumptions are as we go along - and not allow ourselves to use more "advanced" tricks such as the rules of algebra until we have actually proven them. This may seem like an irritating constraint, especially as we will spend a lot of time proving statements which are "obvious", but it is necessary to do this suspension of known facts to avoid *circularity* (e.g., using an advanced fact to prove a more elementary fact, and then later using the elementary fact to prove the advanced fact). Also, this exercise will be an excellent way to affirm the foundations of your mathematical knowledge. Furthermore, practicing your proofs and abstract thinking here

will be invaluable when we move on to more advanced concepts, such as real numbers, functions, sequences and series, differentials and integrals, and so forth. In short, the results here may seem trivial, but the journey is much more important than the destination, for now. (Once the number systems are constructed properly, we can resume using the laws of algebra etc. without having to rederive them each time.)

We will also forget that we know the decimal system, which of course is an extremely convenient way to manipulate numbers, but it is not something which is fundamental to what numbers are. (For instance, one could use an octal or binary system instead of the decimal system, or even the Roman numeral system, and still get exactly the same set of numbers.) Besides, if one tries to fully explain what the decimal number system is, it isn't as natural as you might think. Why is 00423 the same number as 423, but 32400 isn't the same number as 324? Why is $123.4444\ldots$ a real number, while $\ldots 444.321$ is not? And why do we have to carry of digits when adding or multiplying? Why is $0.999\ldots$ the same number as 1? What is the smallest positive real number? Isn't it just $0.00\ldots001$? So to set aside these problems, we will not try to assume any knowledge of the decimal system, though we will of course still refer to numbers by their familiar names such as 1,2,3, etc. instead of using other notation such as I,II,III or 0++, (0++)++, ((0++)++)++ (see below) so as not to be needlessly artificial. For completeness, we review the decimal system in an Appendix (§B).

2.1 The Peano axioms

We now present one standard way to define the natural numbers, in terms of the *Peano axioms*, which were first laid out by Guiseppe Peano (1858–1932). This is not the only way to define the natural numbers. For instance, another approach is to talk about the cardinality of finite sets, for instance one could take a set of five elements and define 5 to be the number of elements in that set. We shall discuss this alternate approach in Section 3.6.