Real Analysis

P. Ouwehand

Department of Mathematics and Applied Mathematics
University of Cape Town

Note to the Student

These notes are a *very rough first draft* for a short course in Real Analysis at the undergraduate level. This is the first time that I am teaching this particular course, and I'm still thinking hard about how to present the material; I'm likely to change my mind at short notice. At present, these notes are *unfinished*, i.e. still being written. There is no guarantee that you will be provided with a finished product by the end of this course, though I will try to do so. These notes are therefore meant as a *supplement* to the notes you take in class, and are *not* a *substitute*. Expect mistakes, but note that though all mistakes are *my* fault, it's *your* responsibility to find and correct them.

How will you do that? Go to the library, which houses many books on real analysis. Two books that you will find particularly useful are *Principles of Mathematical Analysis*, by Walter Rudin, and *The Elements of Real Analysis*, by Robert G. Bartle.

This course is but thirty lectures long. There is more material in the notes than can be covered in class, and many sections can be safely ignored. (Already the preliminary Chapter 0 is ridiculously long, and needs some serious editing.) What you need to know, and what you can omit, will be made clear in lectures.

The content of the course remains similar to what it has been in previous years, though the perspective has shifted slightly: More emphasis is placed on the the importance of *sets* of reals, and on *topological* notions. I'm also hoping to tackle some additional topics, such as the Riemann–Stieltjes integral, if time permits.

Peter Ouwehand
June 2004

Contents

Chapter 0

Preliminaries

0.1 What is Analysis?

Roughly speaking, analysis deals with numbers, sets of numbers, and operations on numbers. It is particularly concerned with what happens if certain operations are performed an arbitrarily large number of times, perhaps infinitely often.

These days we perform most calculations on a computer. Now a computer can handle only rational numbers: Each number is stored using only a finite number of bits, 0 and 1, and thus necessarily rational. For example,

$$101.11_{\text{binary}} = (1)(2^2) + (0)(2^1) + (1)(2^0) + (1)(2^{-1}) + (1)(2^{-2}) = \frac{23}{4}$$

It is clear, therefore, that any number expressed in finitely many bits is equal to $\frac{\text{integer}}{\text{power of 2}}$, and thus necessarily rational.

Since practically all our calculations are handled by computers, and since computers handle only rational numbers, it would seem that the set of rational numbers is sufficiently rich for all our calculations. However, *we* can imagine an operation being performed infinitely often, something that a computer cannot do. Allowing the infinite to creep into our operations results in the creation of something new, namely *irrational numbers*.

For example, start with 1 and perform the following operations over and over: add 1, invert the result, and then add 1, i.e.

$$\begin{aligned} x_0 &= 1 \\ x_{n+1} &= \frac{1}{x_n + 1} + 1 \qquad \text{for } n \geq 1 \end{aligned}$$

Each x_n is a *rational* number (i.e. a *ratio* of integers). If we perform this operation infinitely often, we "get" $\sqrt{2}$, i.e. the limit of the x_n is $\sqrt{2}$, an *irrational* number[1].

[1] A proof that $\sqrt{2}$ is irrational, i.e. not the ratio of two integers, will be provided shortly.

Also consider the following pseudo–code:

LET $X = 1$;
LET $Y = 1$;
FOR $N = 1$ TO ∞ {
    LET $Y = Y/N$;
    LET $X = X + Y$;
                }
PRINT X;

Of course, the output is just $\sum_{n=1}^{\infty} \frac{1}{n!} = e$ Thus this algorithm starts with two rational values for X and Y, and uses only the operations of addition and division. Both these operations preserve rational numbers, yet the output of this algorithm is an irrational number.

In the first example, we took the limit of a sequence of rational numbers, and in the second a limit of a sum of rational numbers. The concept of *limit* captures the notion performing an operation infinitely often. The rational numbers are not sufficiently rich to handle limits, forcing us to extend the number system to also include irrational numbers. Thus the set of *real* numbers is in essence obtained from the set of rational numbers by allowing the taking limits.

The notion of limit is fundamental to analysis, and many of the results we prove in these notes about the set of real numbers are simply not true for the set of rational numbers. Most of the fundamental concepts of calculus involve limits.

- A derivative is a limit:
$$\frac{df}{dx} = \lim_{h \to 0} \frac{f(x+h) - f(x)}{h}$$

- A Taylor series is a limit:
$$e^x = \sum_{k=1}^{\infty} \frac{x^k}{k!} = \lim_{n \to \infty} \sum_{k=1}^{n} \frac{x^k}{k!}$$
If we write $p_n(x) = \sum_{k=1}^{n} \frac{x^k}{k!}$, then each $p_n(x)$ is a polynomial. Thus we have here a sequence of polynomials whose limit is *not* a polynomial. Again, the taking of limits has created a new kind of object.
Similarly, every Fourier series is a limit of sums.

- A definite integral is a limit: If f is continuous on the interval $[a, b]$, then $\int_a^b f(x)\, dx$ is a limit of left–hand sums
$$\int_a^b f(x)\, dx = \lim_{\Delta x \to 0} \sum_{k=1}^{[\frac{b-a}{\Delta x}]} f(a + k\Delta x)\Delta x$$
Here $[y]$ denotes the greatest integer less than or equal to y.

- *Continuity* is defined in terms of a limit: A function f is continuous at a point x_0 if and only if $\lim_{h \to 0} f(x_0 + h) = f(x_0)$.

0.2 Basic Set Theory

Because it became accepted in the 20th century that, in principle, mathematical objects should be sets and mathematical notions should be expressible as relationships between sets, every mathematician needs just a little set theory. The material in this section is not difficult, and no doubt you have seen it all before. We include it merely as a reminder and to fix notation.

Intuitively, a *set* is just a collection of objects.

If A is a set and x is some mathematical object, we say that

$$x \in A \qquad (x \text{ is an } \textbf{element} \text{ of } A)$$

if x is amongst the objects collected in A, and we write

$$x \notin A$$

if it isn't.

The idea is that a set is *characterized entirely by its elements*. Thus if two sets A and B have exactly the same elements, then we must have $A = B$. For example, the sets $A = \{a\}$ and $B = \{a, a\}$ have the same elements, namely only a. Thus $A = B$. The fact that B seems to have two copies of A is immaterial.

For the philosophically minded: This means, for example that

$$\{\text{Evening Star}\} = \{\text{Morning Star}\}$$

as both sets are equal to the {planet *Venus*}. Yet the Evening Star is seen only in the evening, whereas the Morning Star is seen only in the morning...

Instead of *set*, we will also sometimes say *class*, *collection* or *family*; instead of saying *x is an element of A* we will sometimes say *x is a member of A* or *x belongs to A*.

There are two ways to represent sets: (i) by *listing* its elements, and (ii) by some defining *property*. For example, if a set A has finitely many elements $a_1, \ldots, a_n$ then it can be represented by $A = \{a_1, a_2, \ldots, a_n\}$. On the other hand if A is the set of all x having a certain property $P(x)$, then A can be denoted by $A = \{x : P(x)\}$.

Example 0.2.1 The set A of all integers between -1 and 3 can be represented in two ways:

(i) $A = \{-1, 0, 1, 2, 3\}$

(ii) $A = \{n : n \text{ is an integer and } -1 \leq n \leq 3\}$

□

In analysis, the following sets are important:

- The set of natural numbers $\mathbb{N} = \{0, 1, 2, 3, \ldots\}$
- The set of integers or whole numbers $\mathbb{Z} = \{\ldots, -2, -1, 0, 1, 2, \ldots\}$
- The set of rational numbers $\mathbb{Q} = \{\frac{n}{m} : n, m \in \mathbb{Z}, m \neq 0\}$

- The set of real numbers $\mathbb{R}$, and the set of non–negative real numbers is denoted by $\mathbb{R}^+$.

- The set of complex numbers $\mathbb{C} = \{a + ib : a, b \in \mathbb{R}\}$

Another way to represent a set is by *indexing* its elements using another set. This is actually just a method of listing the elements of the set in some coherent way. We write $A = \{a_i : i \in I\}$. Here a_i are the elements of A indexed by the set I. Basically, the set I can be thought of as a set of labels attached in some way to the elements of A.
For example, define $x_n = 2n$. Then $\{x_n : n \in \mathbb{N}\}$ is the set of even numbers, indexed by $\mathbb{N}$.
Or define, for $r \in \mathbb{R}^+$, I_r to be the interval $(-r, r)$. Then $\{I_r : r \in \mathbb{R}^+\}$ is the set of all open intervals centered at zero.

A set doesn't even have to have any elements:

Definition 0.2.2 We define the *empty set* to be the set with no members, and denote it by the symbol $\emptyset$.

□

For example, $\{x : x \in \mathbb{R} \text{ and } x^2 < 0\} = \emptyset$. One could also define the empty set by $\emptyset = \{x : x \neq x\}$. The empty set plays roughly the same role in set theory that the number zero plays in ordinary mathematics.

Definition 0.2.3 We say that a set A is a *subset* of another set B, and write

$$A \subseteq B$$

if and only if every element of A is also an element of B.

We say that A is a *proper subset* of B if A is subset of B, but $A \neq B$.

□

We may also write $B \supseteq A$ instead of $A \subseteq B$; they mean the same thing (just as $x \leq y$ and $y \geq x$ mean the same thing).

Remarks 0.2.4 Note that $A = B$ if and only if $A \subseteq B$ *and* $B \subseteq A$.
Further note that

$$\mathbb{N} \subseteq \mathbb{Z} \subseteq \mathbb{Q} \subseteq \mathbb{R} \subseteq \mathbb{C}$$

□

Exercises 0.2.5 (1) Prove that $\emptyset$ is a subset of every set.
[Hint: Give a proof by contradiction. Assume that there is a set A such that $\emptyset \not\subseteq A$.]

(2) Show formally that if $A \subseteq B$ and if $B \subseteq C$, then $A \subseteq C$.

□

0.2.1 Operations on sets

There are several ways of combining sets to form new sets. In this section we define and give some examples of the set–operations *union, intersection, difference, complementation, cartesian product* and *power set formation.*

Definition 0.2.6 (Union, intersection and difference of two sets)
Suppose that A, B are sets.

(a) The *union* of A and B is the set of all elements which are either in A or in B (or both).

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

(b) The *intersection* of A and B is the set of all elements which belong to *both* A and B.

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

(c) The *set difference* of A and B is the set of all elements which belong to A, but not to B.

$$A - B = \{x : x \in A \text{ and } x \notin B\}$$

□

Two sets A, B are said to be *disjoint* if they have no members in common, i.e. if $A \cap B = \emptyset$. In that case, $A - B = A, B - A = B$.

Often we work within some *universe*, which is just the set of all objects under consideration at that time. The sets that we deal with are then typically subsets of the universe. *Which* set is the universe depends very much on context. If one is dealing with real numbers, the obvious choice of universe is $\mathbb{R}$, but if one is dealing with complex numbers as well, then it would be $\mathbb{C}$. If one is trying to find the solution of an n^{th} order differential equation, then the universe will generally be the set of all n–times differentiable functions.

Given a universe, we also have a unary operation on sets, called *complementation.*

Definition 0.2.7 Let the universe be Ω, and let $A \subseteq \Omega$. The *complement* of A is the set of all elements in the universe which are not in A.

$$A^c = \{x \in \Omega : x \notin A\}$$

□

Note that $A^c = \Omega - A$. Also note that $A - B = A \cap B^c$.

Exercise 0.2.8 Show that A, B are disjoint if and only if $A \subseteq B^c$.

□

Here are some standard identities involving the operations:

Proposition 0.2.9 *Suppose that A, B, C are subsets of some universe Ω.*

(a) Idempotent laws:
$$A \cup A = A; \qquad A \cap A = A$$

(b) Commutative laws:
$$A \cup B = B \cup A; \qquad A \cap B = B \cap A$$

(c) Associative laws:
$$(A \cup B) \cup C = A \cup (B \cup C); \qquad (A \cap B) \cap C = A \cap (B \cap C)$$

(d) Distributive laws:
$$A \cap (B \cup C) = (A \cap B) \cup (A \cap C); \qquad A \cup (B \cap C) = (A \cup B) \cap (A \cup C)$$

(e) Absorption laws:
$$A \cup (A \cap B) = A; \qquad A \cap (A \cup B) = A$$

(f) Complementation laws:
$$A \cup A^c = \Omega; \qquad A \cap A^c = \emptyset$$
$$(A^c)^c = A$$

(g) De Morgan's laws:
$$(A \cap B)^c = A^c \cup B^c; \qquad (A \cup B)^c = A^c \cap B^c$$

□

Note that each of the identities remains true if

- $\cap$ and $\cup$ are interchanged, and
- $\emptyset$ and Ω are interchanged.

Proof: We show how to prove one of the above laws, and leave the remainder as an exercise. Let us prove that $A \cap (B \cup C) = (A \cap B) \cup (A \cap C)$.

First suppose that $x \in A \cap (B \cup C)$. Then $x \in A$ *and* $x \in B \cup C$, by definition of $\cap$. Thus $x \in A$ and *either* (1) $x \in B$, *or* (2) $x \in C$ (or both), by definition of $\cup$. Thus either (1) $x \in A$ and $x \in B$, *or* (2) $x \in A$ and $x \in C$. It follows that either (1) $x \in A \cap B$ *or* (2) $x \in A \cap C$, and thus that $x \in (A \cap B) \cup (A \cap C)$. We have now shown that if $x \in A \cap (B \cup C)$, then also $x \in (A \cap B) \cup (A \cap C)$, i.e. that

$$A \cap (B \cup C) \subseteq (A \cap B) \cup (A \cap C) \qquad (*)$$

Next, assume that $x \in (A \cap B) \cup (A \cap C)$. Then either (1) $x \in A \cap B$, *or* (2) $x \in A \cap C$. In either case, it follows that $x \in A$. Also we must have either (1) $x \in B$, *or* (2) $x \in C$, and thus $x \in B \cup C$. We see, therefore, that we have *both* $x \in A$ *and* $x \in B \cup C$, so that

$x \in A \cap (B \cup C)$. It follows that whenever $x \in (A \cap B) \cup (A \cap C)$, then also $x \in A \cap (B \cup C)$, i.e. that

$$(A \cap B) \cup (A \cap C) \subseteq A \cap (B \cup C) \qquad (\dagger)$$

Putting $(*)$ and $(\dagger)$ together, we obtain

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

as required.

⊣

Exercise 0.2.10 Prove the remaining identities in the proposition above.
(By the way, drawing a Venn diagram does **not** constitute a proof! Venn diagrams work only when you are dealing with a small number of sets.)

□

A set is completely determined by its elements. The order in which those elements are arranged does not matter. For example, $\{a, b\} = \{b, a\}$. When we want the order to matter, we have to deal with ordered tuples. An *ordered pair* is denoted by (a, b), and should be thought of as a collection containing a and b, *in that order*. Thus $(a, b) \neq (b, a)$. Note that

$$(a, b) = (c, d) \iff a = c \text{ and } b = d$$

Generally, an *ordered n–tuple* is denoted by $(a_1, a_2, \ldots, a_n)$, and should be thought of as a collection containing $a_1, a_2, \ldots, a_n$, *in that order*.

The pair (a, b) is often defined to be the *set* $\{\{a\}, \{a, b\}\}$. You can check that this definition yields the required property that $(a, b) = (c, d)$ iff $a = c$ and $b = d$.
(a, b, c) is then defined to be $(a, (b, c))$ (which is just the set $\{\{a\}, \{a, \{\{b\}, \{b, c\}\}\}\}$), etc. This is in keeping with the notion that all mathematical objects should be sets. On first encounter, however, you might find this arbitrary, clumsy, and unnecessary, and you wouldn't be far wrong: The *main* thing that you need to keep in mind is that *an ordered tuple is a collection in which the order matters.*

Using ordered tuples, we can define one more way of making new sets from old:

Definition 0.2.11 (Cartesian product) Suppose that $A_1, A_2, \ldots, A_n$ are sets. The *cartesian product* of $A_1, \ldots, A_n$ is the set of *all* n–tuples $(a_1, \ldots, a_n)$, with each $a_k \in A_k$.

$$A_1 \times A_2 \times \cdots \times A_n = \{(a_1, a_2, \ldots, a_n) : a_k \in A_k \text{ for } k = 1, 2, \ldots, n\}$$

□

Example 0.2.12 If $A = \{a, b\}$ and $B = \{1, 2, 3\}$, then their product is the 6–element set

$$A \times B = \{(a, 1), (a, 2), (a, 3), (b, 1), (b, 2), (b, 3)\}$$

□

Proposition 0.2.13 *If A has n elements and B has m elements, then $A \times B$ has $n \times m$ elements.*

⊣

Exercises 0.2.14 (1) Prove the preceding theorem by induction.
[Hint: Let B be a fixed set with m elements, and proceed by induction on the number of elements in A. First show that if A has 0 elements, then $A \times B$ has $0 \cdot m$ elements. Now *assume* that whenever A has $n = k$ elements, $A \times B$ has km elements. Show that this implies that if A has $n = k + 1$ elements, then $A \times B$ has $(k + 1)m$ elements.]

(2) Prove that $A \times (B \cap C) = (A \times B) \cap (A \times C)$

(3) Is it true that $(A \times B) \cup (C \times D) = (A \cup C) \times (B \cup D)$?

□

We will identify the sets $(A \times B) \times C$ and $A \times (B \times C)$ with $A \times B \times C$, although, strictly speaking, they are not equal.

For example, $((a, b), c))$ is an element of the first set, but not of the second or third. $(a, (b, c))$ belongs to the second, but not to the first or third. (a, b, c) belongs to the third, but not to the first two. However, we shall simply *identify* $(a, (b, c)), ((a, b), c)$ and (a, b, c), i.e. we shall not distinguish between them. After all, all that matters is the order of a, b, c and that is the *same* in each of these tuples.

Example 0.2.15 The n–dimensional Euclidean space, denoted by $\mathbb{R}^n$, is just the n–fold cartesian product of $\mathbb{R}$:

$$\mathbb{R}^n = \underbrace{\mathbb{R} \times \cdots \times \mathbb{R}}_{n \text{ times}} = \{(x_1, x_2, \ldots, x_n) : x_i \in \mathbb{R}\}$$

We will identify the sets $\mathbb{R}^n \times \mathbb{R}^m$ and $\mathbb{R}^{n+m}$.

Strictly speaking, the first is the set

$$\mathbb{R}^n \times \mathbb{R}^m = \{((x_1, \ldots, x_n), (y_1, \ldots, y_m)) : x_i, y_j \in \mathbb{R}\}$$

whereas the second set is

$$\mathbb{R}^{n+m} = \{(x_1, \ldots, x_n, y_1, \ldots, y_m) : x_i, y_j \in \mathbb{R}\}$$

but the two sets clearly have the same basic structure, in that they are made up of tuples with x_1 followed by x_2, …, followed by y_m.

□

Exercises 0.2.16 (1) Draw the following sets in the xy–plane (i.e. $\mathbb{R}^2$):

(i) $\{-1, 2, 3\} \times \{3, 4, 5\}$

(ii) $\{1\} \times [0, 1]$

(iii) $[0, 1] \times \{1\}$

(iv) $(0, 1] \times [2, 3)$

(2) Describe the set $[0,1] \times [0,1] \times [0,1]$.

(3) Consider the cylinder of unit radius about the z–axis in $\mathbb{R}^3$:

$$\mathcal{C} = \{(x,y,z) : x^2 + y^2 = 1\}$$

Represent $\mathcal{C}$ as a product of two sets.

□

Thus far, we have considered union, intersection and cartesian product as *binary operations*, involving just two sets. Frequently, however, we may need to consider these as *infinitary* operations: We can, for example, take the union of infinitely many sets. We define the union, intersection and cartesian product of a family of sets as follows:

Definition 0.2.17 (Union, intersection and product of a family of sets)
If $\mathcal{A} = \{A_i : i \in I\}$ is a family of sets, we may define

(a) the *union*

$$\bigcup \mathcal{A} = \bigcup_{i \in I} A_i = \{x : x \in A_i \text{ for some } i \in I\}$$

(b) the *intersection*

$$\bigcap \mathcal{A} = \bigcap_{i \in I} A_i = \{x : x \in A_i \text{ for all } i \in I\}$$

(c) the *cartesian product*

$$\prod \mathcal{A} = \prod_{i \in I} A_i = \{(a_i)_I : a_i \in A_i \text{ for all } i \in I\}$$

Here $(a_i)_I$ is a generalized tuple, indexed by I.
In essence, $(a_i)_I$ is a function with domain I and range $\bigcup_{i \in I} A_i$. We will return to this later.

We will frequently write $\bigcup_I A_i$ or $\bigcup_i A_i$ instead of $\bigcup_{i \in I} A_i$. We will also write $\bigcup_{n=1}^{\infty} A_n$ instead of $\bigcup_{n \in \mathbb{N}} A_n$. The same holds for $\bigcap$ and $\prod$.

□

Remarks 0.2.18 Note that

(i) $\bigcup\{A, B\} = A \cup B$

(ii) $\prod\{A, B, C\} = A \times B \times C$

(iii) $\bigcap\{X_1, X_2, \ldots, X_n\} = X_1 \cap X_2 \cap \cdots \cap X_n$

etc.

□

Exercises 0.2.19 (1) Define $A_n = (\frac{1}{n+1}, 1]$ for $n \in \mathbb{N}$. Calculate $\bigcup_{n=1}^{\infty} A_n$ and $\bigcap_{n=1}^{\infty} A_n$.

(2) Let $B_r = \{\vec{x} \in \mathbb{R}^3 : |x| \leq r\}$. Calculate $\bigcup_{r \in (0,1]} B_r$ and $\bigcap_{r \in (0,1]} B_r$.

□

Definition 0.2.20 Let $\mathcal{A} = \{A_n : n \in \mathbb{N}\}$ be a family of sets.

(a) We define the **limit superior** of the sets (A_n) by

$$\limsup_n A_n = \bigcap_{n=1}^{\infty} \bigcup_{m \geq n} A_m$$

(b) We define the **limit inferior** of the sets (A_n) by

$$\liminf_n A_n = \bigcup_{n=1}^{\infty} \bigcap_{m \geq n} A_m$$

□

Note that $a \in \limsup A_n$ if and only if $a \in \bigcup_{m \geq n} A_m$ for all n, i.e. if and only if for all n there is $m \geq n$ such that $a \in A_m$. Thus $a \in \limsup_n A_n$ if and only if *a belongs to infinitely many of the sets A_n*.

Exercises 0.2.21 (1) Show that $a \in \liminf A_n$ if and only if *a belongs to almost all of the A_n*. ("Almost all" mean "all except possibly finitely many". Thus we are claiming that $a \in \liminf A_n$ if and only if there are at most finitely many n such that $a \notin A_n$.)

(2) Let $A_n = [0, \frac{1}{n}]$ if n is even, and let $A_n = [-\frac{1}{n}, 0]$ if n is odd. Calculate $\limsup_n A_n$ and $\liminf_n A_n$.

(3) Let $A_n = [0, n]$ if n is even, and let $A_n = [-n, 0]$ if n is odd. Calculate $\limsup_n A_n$ and $\liminf_n A_n$.

(4) Let $A_n = (-1, 1 + \frac{1}{n})$ if n is even, and let $A_n = [-1 - \frac{1}{n}, 1]$ if n is odd. Calculate $\limsup_n A_n$ and $\liminf_n A_n$.

(5) Given a sequence of sets A_n, and a positive integer N, define $B_n = A_{N+n}$. Show that $\limsup_n A_n = \limsup_n B_n$ and that $\liminf_n A_n = \liminf_n B_n$. This shows that lim sup and lim inf are determined by the "tail" of the sequence A_n only.

□

Here is another way of making new sets from old: Given a particular set, one should be able to collect all of its subsets together into a new set, called the *power set*.

Definition 0.2.22 (Power set)
If A is a set, then the *power set* of A is the set of all subsets of A.

$$\mathcal{P}(A) = \{B : B \subseteq A\}$$

□

Note that $\emptyset, A \in \mathcal{P}(A)$. They are, respectively, its smallest and biggest members.

Example 0.2.23 Let $A = \{1, 2, 3\}$. Then the powerset of A is the 8–element set

$$\mathcal{P}(A) = \{\emptyset, \{1\}, \{2\}, \{3\}, \{1, 2\}, \{1, 3\}, \{2, 3\}, \{1, 2, 3\}\}$$

□

Proposition 0.2.24 *If A has n elements, then $\mathcal{P}(A)$ has 2^n elements.*

Proof: We prove this proposition by *mathematical induction*. Firstly, the proposition is true for all sets with 0 elements (i.e. if a set has 0 elements, then it has 2^0 subsets. To see this, note that the only set with 0 elements is $\emptyset$, and that $\emptyset$ has only one subset, namely itself. Since $2^0 = 1$, the proposition holds for all sets with 0 elements.)

Next *assume* that the proposition holds for all sets with n elements. We now want to prove that the proposition is also true for sets with $n+1$ elements. So let A be a set with $n+1$ elements, and pick $a \in A$. Let $B = A - \{a\}$. Then B has n elements, and so by assumption, 2^n subsets.

Now reason as follows: The subsets of A can be divided into two classes, namely (1) those which have a as element, and (2) those which do not. It is obvious that no subset of A belongs to both classes, and that every subset of A belongs to one of them.

(1): If $C \subseteq A$ does not have a as element, then $C \subseteq B$. The former are just subsets of B, and there are 2^n of them.

(2): If $C \subseteq A$ does have $a \in C$, then $C = C' \cup \{a\}$, where $C' \subseteq B$. Since to each such C there corresponds a C', there are as many subsets of A containing a as there are subsets of B, i.e. 2^n.

Hence A has $2^n + 2^n = 2^{n+1}$ subsets. We have therefore proved the following:

(i) Every set with 0 elements has 2^0 subsets;

(ii) If every set with n elements has 2^n subsets, then every set with $n+1$ elements has 2^{n+1} subsets

Thus since every set with 0 elements has 2^0 subsets, we deduce that every set with 1 element has 2^1 subsets. From *that* we deduce that every set with 2 elements has 2^2 subsets, and from *that*, that every set with 3 elements has 2^3 subsets, etc.

⊣

Exercises 0.2.25 Prove the above proposition again, using the **binomial theorem**.
[Hint: Recall that the binomial coefficient $\binom{n}{m} = \frac{n!}{m!(n-m)!}$ describes the number of ways in which m objects can be chosen from a collection of n objects.For example , there are $\binom{20}{11}$ ways of choosing a soccer team from a group of twenty individuals. Also recall the binomial theorem:

$$(a+b)^n = \sum_{m=0}^{n} \binom{n}{m} a^m b^{n-m}$$

Use these facts to prove that if A is a set with n elements, then $\mathcal{P}(A)$ is a set with 2^n elements.]

□

0.2.2 Functions

Originally, a function was regarded as a *rule* (or a *formula*, or an *algorithm*) for associating one real number with another. For example,

$$f(x) = 2x^3$$

explicitly shows how to calculate a number $f(x)$ which is to be associated with x: First cube x, and then multiply the resultant by 2. However, this original formulation proved to be unduly restrictive. For one thing, Fourier showed that practically any continuous curve of finite length could be give a "formula" as an infinite trigonometric series. For another, we may want to associate numbers with other mathematical objects, or one kind of mathematical object with another — there is no reason to restrict ourselves solely to numbers.

For example, we may want to associate with each rectangle its area. Thus we have a function which assigns a number to each rectangle.
Or, we may want to assign to each subset of $\mathbb{R}$ its power set. This yields a function which assigns a set to each set.

Thus a general definition of function dispenses with the idea that it is a rule, but keeps the idea of associating one object with another:

Definition 0.2.26 Let A, B be sets. A *function* (or *map*) f from A to B, written

$$f : A \to B \quad \text{or} \quad A \xrightarrow{f} B$$

is a subset of the cartesian product $A \times B$ with the following property:

$$\text{for each } a \in A \text{ there exists } \textit{exactly one } b \in B \text{ such that } (a, b) \in f$$

In that case write

$$f(a) = b \quad \text{instead of } (a, b) \in f$$

We call b the *image* (or *value*) of a under f, and call a a *preimage* of b. We also say that a *maps to* b under f.
The set A is called the *domain* of f, and the set B is called the *codomain* of f

$$A = \text{dom}(f) \qquad B = \text{codom}(f)$$

The *range* of f is the set of all possible values of f, and denoted $\text{ran}(f)$.

□

Essentially, this concept of function is arrived at by deliberately confusing a function with its graph. For example, the graph of the function $f : \mathbb{R} \to \mathbb{R} : x \mapsto 2x^3$ is a curve in the cartesian plane. This curve is therefore a set of ordered pairs:

$$\text{Graph}(f) = \{(x, y) : y = 2x^3\}$$

For example, the points $(0,0),(1,2),(2,16),(3,54)$ belong to the graph. Now we assert that a function *is* its graph. Thus the function $f(x) = 2x^3$ is nothing but the set $\{(x,y) : y = 2x^3\} \subseteq \mathbb{R} \times \mathbb{R}$.

You've already met more than just a few functions in your mathematical education up to date. The most obvious ones are functions from $\mathbb{R}^n$ to $\mathbb{R}^m$, such as $f(x) = x^2, g(x,y) = \sin(x^3 + y), h(x,y,z) = (xy, x \ln z)$, etc. Here are a few more that you might not yet have considered as functions:

Examples 0.2.27 (a) Define $\mathbb{Z} \xrightarrow{f} \mathcal{P}(\mathbb{Z})$ by: $f(n) = \{m : m \text{ divides } n\}$. Then f is a function which maps a number to a set. For example,

$$f(12) = \{\pm 1, \pm 2, \pm 3, \pm 4, \pm 6, \pm 12\} = f(-12)$$

(b) Let $\mathcal{C}^0(\mathbb{R},\mathbb{R}) = \{f : f \text{ is a continuous map from } \mathbb{R} \text{ to } \mathbb{R}\}$, and let $a \leq b \in \mathbb{R}$. Then $\int_a^b : \mathcal{C}^0(\mathbb{R},\mathbb{R}) \longrightarrow \mathbb{R}$ is a function which assigns to every continuous map its definite integral.

(c) Let $\mathcal{C}^1(\mathbb{R},\mathbb{R})$ be the set of all maps from $\mathbb{R}$ to $\mathbb{R}$ which have continuous first derivatives. Then the derivative operator is a map $D : \mathcal{C}^1(\mathbb{R},\mathbb{R}) \longrightarrow \mathcal{C}^0(\mathbb{R},\mathbb{R})$.

(d) **curl** is a map from the set of vector fields on $\mathbb{R}^3$ to itself. **div** is a map from the set of vector fields on $\mathbb{R}^3$ to the set of functions on $\mathbb{R}^3 \to \mathbb{R}$. **grad** is a map from the set of differentiable functions $\mathbb{R}^3 \to \mathbb{R}$ to the set of vector fields on $\mathbb{R}^3$.

(e) An $n \times m$ matrix A can be regarded as a map from $A : \mathbb{R}^m \longrightarrow \mathbb{R}^n$.

(f) Addition and multiplication are functions from $\mathbb{R}^2$ to $\mathbb{R}$. Addition can, in fact, be described by the 1×2–matrix $(1\ \ 1)$, for $(1\ \ 1)\binom{a}{b} = a + b$.

(g) If Ω is a universal set, then union and intersection can be regarded as functions from $\mathcal{P}(\Omega) \times \mathcal{P}(\Omega)$ to $\mathcal{P}(\Omega)$, which map the ordered pair (A,B) to $A \cup B$ and $A \cap B$ respectively.

(h) We can also regard the bigger version $\bigcup$ of union as a map, but this time we have $\bigcup : \mathcal{P}(\mathcal{P}(\Omega)) \longrightarrow \mathcal{P}(\Omega)$. It assigns to any family of subsets of Ω its union. (Note that a family of subsets of Ω is just a set of elements of $\mathcal{P}(\Omega)$, i.e. it is a subset of $\mathcal{P}(\Omega)$, and therefore an element of $\mathcal{P}(\mathcal{P}(\Omega))$.) The same goes for intersection.

□

For any set A, there is an important function on A called the *identity function*. It is denoted by id_A, and is defined by

$$\mathrm{id}_A : A \longrightarrow A \qquad \mathrm{id}_A(a) = a$$

Thus $\mathrm{id}_A = \{(a,a) : a \in A\}$.

Examples 0.2.28 (a) The identity function on $\mathbb{R}$ is just the function $y = x$.

(b) The identity function on $\mathbb{R}^n$ is the identity matrix

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

□

Definition 0.2.29 Let $f : A \to B$. If $A' \subseteq A$, we can define the *restriction* of f to A' as follows:

$$f|A' \text{ is a map from } A' \text{ to } B, \text{ such that } (f|A')(a) = f(a) \text{ for all } a \in A'$$

□

Definition 0.2.30 Let $A \xrightarrow{f} B$ be a function.

(a) f is said to be *one-to-one* (or 1-1, or *injective*) if and only if the following condition holds:

$$\text{If } f(a_1) = f(a_2), \text{ then } a_1 = a_2.$$

(b) f is said to be *onto* (or *surjective*) if and only if

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

(c) f is said to be a *bijection* (or a *one-to-one correspondence*) if it is both an injection and a surjection.

□

Remarks 0.2.31 A function $f : A \to B$ is injective if no two distinct members of a map to the same $b \in B$, i.e. if every $b \in B$ has *at most one* preimage.
f is surjective if and only if every b in B gets mapped onto by some $a \in A$, i.e. if every $b \in B$ has *at least one preimage*. In that case B is the range of f, i.e. $\text{ran}(f) = \text{codom}(f)$.
f is a bijection if and only if every $b \in B$ has *exactly one* preimage.

It should be clear that there is a bijection from a finite set A to another set B if and only if A and B have the same number of elements.

□

Examples 0.2.32 (a) Let $f(x) = x^2$. We would generally regard f as a function with domain $\mathbb{R}$ and codomain $\mathbb{R}$. The range of f is $[0, +\infty)$, since f takes no negative values. f is not injective, because, for example $f(1) = f(-1)$. f is not surjective either, since -1 is not in the range of f.

(b) If we define $g(x) : [0, 1] \longrightarrow [0, 1]$ by $g(x) = x^2$, then we may regard g as the restriction of f to $[0, 1]$, i.e. $g = f|[0, 1]$. Now g is clearly a bijection.

(c) $x^3 : \mathbb{R} \longrightarrow \mathbb{R}$ is a bijection.

(d) Let $\mathbb{Q}^+$ denote the set of all non–negative rational numbers. The map $h : \mathbb{Z} \times \mathbb{N} \longrightarrow \mathbb{Q}^+$ defined by $h(n, m) = \frac{n}{m}$ is surjective, but not injective.

(e) If $A \subseteq B$, then the **inclusion** $f : A \to B$ defined by: $f(a) = a$ is an injection. It is a bijection if and only if $A = B$.

(f) Let A be an $n \times n$–matrix, regarded as a map from $\mathbb{R}^n$ to $\mathbb{R}^n$. Then A is injective if and only if $\det(A) \neq 0$.

□

Next, we discuss how functions can be combined:

Definition 0.2.33 If $f : A \to B$ and $g : B \to C$, then $g \circ f$ is a function from A to C, defined by

□

Note that the composition does in one step what f and g do in two:

$$A \xrightarrow{f} B \xrightarrow{g} C \qquad a \overset{f}{\mapsto} f(a) \overset{g}{\mapsto} g(f(a))$$
$$A \xrightarrow{g\circ f} C \qquad a \overset{g\circ f}{\mapsto} g(f(a))$$

Also note that $g \circ f$ means:

Do f first, then g

i.e. the last shall be first.

An often used fact is that *composition is an associative operation* on functions, i.e.

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

By this equation we mean that: one side is defined if and only if the other side is defined, and in that case they are equal.
For if $A \xrightarrow{f} B, B \xrightarrow{g} C$, and $C \xrightarrow{h} D$, then $h \circ (g \circ f)$ is a function from A to D which works as follows: First do $g \circ f$, then do h. But to do $g \circ f$, you must first do f, then g. The combined result is

First do f, then g, and then h: $(h \circ (g \circ f))(a) = h(g(f(a)))$

Similarly, $(h \circ g) \circ f$ is a function from A to D which works as follows: First do f, then $h \circ g$. But to do $h \circ g$, you must first do g, then h. The combined result is therefore

First do f, then g, and then h: $((h \circ g) \circ f)(a) = h(g(f(a)))$

and thus $h \circ (g \circ f) = (h \circ g) \circ f$, as claimed.

Example 0.2.34 Consider the following functions (note their domains and codomains):

$$\mathbb{R} \xrightarrow{f} \mathbb{R}^+ : x \mapsto x^2 + 1$$
$$\mathbb{R}^+ \xrightarrow{g} \mathbb{R}^+ : y \mapsto \sqrt{y}$$
$$\mathbb{R}^+ \xrightarrow{h} [-1, 1] : z \mapsto \sin(z)$$

Then

$$\mathbb{R} \xrightarrow{g\circ f} \mathbb{R}^+ : x \mapsto \sqrt{x^2 + 1}$$
$$\mathbb{R}^{+1} \xrightarrow{h\circ g} [-1, 1] : y \mapsto \sin(\sqrt{y})$$

and thus

$$\mathbb{R} \xrightarrow{h\circ(g\circ f)} [-1, 1] : x \mapsto \sin(\sqrt{x^2 + 1})$$
$$\mathbb{R} \xrightarrow{(h\circ)g\circ f} [-1, 1] : x \mapsto \sin(\sqrt{x^2 + 1})$$

□

Exercises 0.2.35 (1) Let $f : \mathbb{N} \to \mathbb{N} : n \mapsto n^2$, and let $g : \mathbb{N} \to \mathbb{N} : n \to n + 2$. Calculate $(f \circ g)(5)$ and $g \circ f(5)$.
Write down formulas for $f \circ g$ and $g \circ f$.

(2) Suppose that $f(x) = x^2$ and $g(x) = x + 3$. Calculate $g \circ f(x)$ and $f \circ g(x)$. Note that $g \circ f \neq f \circ g$.

(3) If A is an $n \times m$–matrix, and B is an $m \times r$–matrix, then we can regard them as functions $\mathbb{R}^m \xrightarrow{A} \mathbb{R}^n$, $\mathbb{R}^r \xrightarrow{B} \mathbb{R}^m$. The composition $A \circ B$ is therefore a map $\mathbb{R}^r \to \mathbb{R}^n$. It is not hard to show that the composition is just the matrix product, i.e. that $A \circ B = AB$. Do so!

(4) Suppose that $g \circ f_1 = g \circ f_2$. Prove that if g is injective then we can "cancel" g to conclude $f_1 = f_2$. Give an example to show that left–cancellation may fail if g is not injective.

(5) Suppose that $g_1 \circ f = g_2 \circ f$. Prove that if f is surjective then we can "cancel" f to obtain $g_1 = g_2$. Show that right–cancellation may fail if f is not surjective.

□

Note that if $f : A \longrightarrow B$, then $f \circ \mathrm{id}_A = f$, and $\mathrm{id}_B \circ f = f$. Thus the identity function behaves like an identity element for the operation of composition.

The number 0 is an identity element for the operation of addition, because $x + 0 = x$.
The number 1 is an identity element for the operation of multiplication, because $x \cdot 1 = x$.

Next, we tackle the idea of *inverting* (or *reversing*) the effect of a function. Take the function $f(x) = 3x$. It transforms the number x into the number $3x$. To *undo* this transformation, you just multiply $3x$ by $\frac{1}{3}$. The function $g(x) = \frac{1}{3}x$ inverts the effect of f, in that

$$g \circ f(x) = x \qquad f \circ g(y) = y$$

Thus applying first f, and then g gets you back to the starting point x. The same holds true if you apply g first, and then f.

Can every function be inverted? No, as is easy to see: Consider the function $f(x) = x^2$. Then $f(2) = 4 = f(-2)$. Now if g is a function which reverses the effect of f, then we cannot decide whether $g(4) = 2$ or $g(4) = -2$. The problem arises because g is not 1-1.

Let's make the preceding discussion precise:

Definition 0.2.36 Let $f : A \to B$. We say that f is *invertible* if and only if there is a function $g : B \to A$ such that

$$g(f(a)) = a \quad \text{for all } a \in A, \qquad f(g(b)) = b \quad \text{for all } b \in B \tag{$*$}$$

The function g, *if it exists*, is called the *inverse* of f, and denoted $g = f^{-1}$. Then $(*)$ amounts to saying

$$f^{-1} \circ f = \mathrm{id}_A \qquad \text{and} \qquad f \circ f^{-1} = \mathrm{id}_B$$

□

Note that if f^{-1} exists, then

$$f^{-1}(b) = a \quad \text{if and only if} \quad f(a) = b$$

Proposition 0.2.37 *A function $f : A \longrightarrow B$ is invertible if and only if it is a bijection.*

Proof: Suppose that f is invertible, i.e. that f^{-1} exists. Then f^{-1} is a function from B to A. We first show that f is surjective: Let $b \in B$. Since the domain is B, $f^{-1}(b)$ must be defined, i.e. there must be some $a \in A$ such that $f^{-1}(b) = a$. But then $f(a) = b$. Hence every $b \in B$ has a preimage.
Next we show that f is injective. For suppose that $f(a_1) = f(a_2) = b$. Then $f^{-1}(b) = a_1$ and $f^{-1}(b) = a_2$. Since f^{-1} is a function, we must have $a_1 = a_2$ (check the definition of function), and hence f is injective.
This proves that if f is invertible, then f is a bijection.

Now we prove the converse. If f is a bijection, then it is onto B. Hence for every $b \in B$ there is some $a \in A$ such that $f(a) = b$. Moreover, since f is one-to-one, that a has to be unique. So we may define $f^{-1}(b)$ to be the unique a such that $f(a) = b$. This makes f^{-1} into a well–defined function $f^{-1} : B \to A$.

⊣

Examples 0.2.38 (a) The function $f(x) = x^3$ is a bijection on the reals, and its inverse is $g(x) = \sqrt[3]{x}$.

(b) The function $f(x) = x^2$ does not have an inverse, since it is not a bijection. However, if we *restrict* f to the non–negative reals, then $f|\mathbb{R}^+$ is a bijection. Its inverse is the square root function.

(c) The function $f : \mathbb{R} \longrightarrow (0, +\infty)$ defined by $f(x) = e^x$ is bijective. Its inverse is the natural logarithm $\ln x$.

(d) The function $\sin x$ is neither injective, nor surjective; however, if we restrict $\sin x$ and regard it as a function $[-\frac{\pi}{2}, \frac{\pi}{2}] \longrightarrow [-1, 1]$, then it is a bijection, and its inverse is $\arcsin x$.

(e) If A is an $n \times n$–matrix, regarded as a function on $\mathbb{R}^n$, then A has an inverse function if and only if A has an inverse matrix. Since composition is just matrix multiplication, the *inverse function* of A is just the *inverse matrix* A^{-1}.

□

Remarks 0.2.39 Note that, in general,

$$f^{-1}(x) \quad \neq \quad \frac{1}{f(x)}$$

e.g. $\sqrt[3]{x} \neq \frac{1}{x^3}$.

The number $x^{-1} = \frac{1}{x}$ is the inverse of x under the operation of *multiplication*, in that

$$x \cdot x^{-1} = 1 \qquad x^{-1} \cdot x = 1$$

noting that 1 is the identity for multiplication.
The function f^{-1} is the inverse of f under the operation of *composition*, in that

$$f \circ f^{-1} = \mathrm{id} \qquad f^{-1} \circ f = \mathrm{id}$$

noting that id is the identity for composition.
The same notation for inverse, i.e. $^{-1}$, refers to *different operations*, so there's no reason to believe that there is any relationship between them.

□

The notion of invertibility can be refined:

Definition 0.2.40 Let $f : A \to B$ and $g : B \to A$.

(a) g is called a *left inverse* of f if $g \circ f = \mathrm{id}_A$.

(b) g is called a *right inverse* of f if $f \circ g = \mathrm{id}_B$.

□

Note that if f is invertible, then f^{-1} is both a left and a right inverse of f, and vice versa.

Exercises 0.2.41 (1) Prove that a function f has a left inverse if and only if it is injective.

(2) Prove that a function f has a right inverse if and only if it is surjective.

(3) Prove that if a function f has a left inverse g *and* a right inverse h, then f is invertible, and $g = h$.

(4) Consider $f : \{a, b, c\} \to \{1, 2\}$ defined by $f(a) = f(b) = 1, f(c) = 2$. Find two distinct right inverses of f.

(5) Consider the inclusion $\iota : \mathbb{Z} \to \mathbb{Q}$. Construct two distinct left inverses of ι.

□

We have already noted the confusion that may possibly arise by the two uses of the symbol $^{-1}$. We have but few symbols at our disposal, and many of them must therefore serve more than one function. Thus you must *always be aware of the context* in which a particular symbol is used.

You have to do this when using ordinary language: You *know* in what sense the newspaper headline

"School kids make great snacks at fund raiser"

is meant, even though the other sense offers greater amusement value.

I say this because we are about to add to the possible confusion. With every function $f : A \to B$ (not necessarily invertible), we can associate two new functions between the power sets of A and B

$$f[\cdot] : \mathcal{P}(A) \to \mathcal{P}(B) : A' \mapsto \{b \in B : \text{ There is } a' \in A' \text{ such that } f(a') = b\} \quad \text{where } A' \subseteq A$$
$$f^{-1}[\cdot] : \mathcal{P}(B) \to \mathcal{P}(A) : B' \mapsto \{a \in A : f(a) \in B'\} \quad \text{where } B' \subseteq B$$

Thus $f[\cdot]$ assigns to each subset A' of A a subset $f[A'] \subseteq B$. Similarly, $f^{-1}[\cdot]$ transforms each subset B' of B into a subset $f^{-1}[B'] \subseteq A$.
We will, for the moment, use square brackets to distinguish the various functions, but will drop this convention later. Which function is meant will be clear from context. We shall also call $f[A']$ the *direct image* of A' along f, and $f^{-1}[B']$ the *inverse image* of B' along f. Note that

$$f[A'] = \text{set of all images of } a \in A'$$

whereas

$$f^{-1}[B'] = \text{set of all preimages of } b \in B'$$

Remarks 0.2.42 Sometimes the notation $f^{\rightarrow}$ is used for direct image, and $f^{\leftarrow}$ for inverse image.

□

Inverse images play a very important role in mathematics. It is therefore useful to remember the following:

$$a \in f^{-1}[B'] \quad \text{if and only if} \quad f(a) \in B'$$

Similarly,

$$b \in f[A'] \quad \text{if and only if there is } a' \in A' \text{ such that } f(a') = b$$

Examples 0.2.43 (a) Suppose that $f : \mathbb{R} \to \mathbb{R} : x \mapsto x^2$. Then

$$f[-1,2] = [0,4] \quad , f[\mathbb{Z}] = \{0,1,4,9,\dots\}, \quad f[\{4\}] = \{16\}$$

Also

$$f^{-1}[0,1] = [-1,1], \quad f^{-1}[\{4\}] = \{2,-2,\}, \quad f^{-1}[\{-4\}] = \emptyset$$

In each case, a *set* is transformed into a *set*.

(b) Suppose that $A = \{a_1, a_2, a_3\}, B = \{b_1, b_2, b_3\}$, and that $f : A \to B$ is defined by $f(a_1) = f(a_3) = b_1$, and $f(a_2) = b_3$. Then

$$f[\{a_1\}] = f[\{a_3\}] = f[\{a_1, a_3\}] = \{b_1\}, \quad f[\{a_2\}] = b_3, \quad f[A] = \{b_1, b_3\}, \quad f[\emptyset] = \emptyset$$

and

$$f^{-1}[\{b_3\}] = \{a_2\}, \quad f^{-1}[\{b_2\}] = f^{-1}[\emptyset] = \emptyset, \quad f^{-1}(B) = f^{-1}[\{b_1, b_3\}] = A$$

□

Exercises 0.2.44 1. Let $f : A \to B$ be a function, and let $A' \subseteq A$, $B' \subseteq B$.

(a) Show that $A' \subseteq f^{-1}[f[A']]$

(b) Show that $B' \supseteq f[f^{-1}[B']]$

(c) Show that $A' = f^{-1}[f[A']]$ if and only if f is injective.

(d) Show that $B' = f[f^{-1}[B']]$ if and only if f is surjective.

[Hints: Reason along the following lines:
(b) If $b \in f[f^{-1}[B']]$ then $b = f(a)$ for some $a \in f^{-1}[B']$. But then $f(a) \in B'$, and so $b \in B'$.
(c) If $a \in f^{-1}[f[A']]$ then $f(a) \in f[A']$. Thus there is $a' \in A'$ such that $f(a) = f(a')$. But since f is injective, $a = a'$, and so $a \in A'$.]

2. Inverse images preserve the set operations: Let $f : A \to B$, and suppose that G, H are subsets of B. Then

   (a) If $G \subseteq H$, then $f^{-1}[G] \subseteq f^{-1}[H]$;

   (b) $f^{-1}[G \cap H] = f^{-1}[G] \cap f^{-1}[H]$;

   (c) $f^{-1}[G \cup H] = f^{-1}[G] \cup f^{-1}[H]$;

   (d) $f^{-1}[G - H] = f^{-1}[G] - f^{-1}[H]$;

3. Direct images are not quite so well behaved: Let $f : A \to B$, and suppose that $G, H \subseteq A$.

   (a) Suppose that $G \subseteq H$. Show that $f[G] \subseteq f[H]$;

   (b) Show that $f[G \cup H] = f[G] \cup f[H]$;

   (c) Show that $f[G \cap H] \subseteq f[G] \cap f[H]$;

   (d) Give an example to show that we may not have $f[G \cap H] = f[G] \cap f[H]$;

   (e) Show that $f[G] - f[H] \subseteq f[G - H] \subseteq f[G]$;

   (f) Give an example to show, in (e), that both $\subseteq$'s may fail to be $=$'s.

□

We end this section with some notation: Suppose that A, B are finite sets, and that A has n elements, and B m elements. How many functions are there from A to B?
For each $a \in A$ we have m choices for the value $f(a) \in B$. Thus there are m^n functions from A to B. For that reason

Definition 0.2.45 Let A, B be sets. Then we define

$$B^A = \text{set of all functions from } A \text{ to } B$$

Some authors use ${}^AB$ instead of B^A.

□

Note that each function $f : A \to B$ is a subset of $A \times B$. Hence B^A is a set of subsets of $A \times B$, i.e. $B^A \in \mathcal{P}(\mathcal{P}(A \times B))$.

0.2.3 Relations

We want to capture mathematically the idea that two objects are somehow related. For example, suppose that we have two sets

$$M = \{\text{Archie, Reggie, Forsythe}\} \qquad W = \{\text{Betty, Veronica, Ethel}\}$$

and suppose that A is married to B, and that R is married to V, but that F and E remain unmarried. The relation of being married is described by the set

$$\mathbf{R} = \{(\text{A,B}), (\text{R,V})\}$$

Note that $\mathbf{R}$ is a subset of the cartesian product $M \times W$. We will sometimes write $x\mathbf{R}y$ instead of $(x, y) \in \mathbf{R}$. Thus in this case, $x\mathbf{R}y$ if and only if x is married to y.

As for functions, the general definition of a relation is quite abstract:

Definition 0.2.46 A *relation* from a set A to a set B is just a subset of $A \times B$. If $A = B$, we just say that $\mathbf{R}$ is a relation on A.

□

Thus if $A = \mathbb{N}$ and $B = \mathbb{N} \cup \{0\}$, then

$$\mathbf{L} = \{(1, 3), (2, 1), (3, 4), (4, 1), (5, 5), (6, 9) \dots\} \subseteq A \times B$$

is a relation from A to B. Here what the relation actually *is* may not be obvious. Could you have guessed that $7\mathbf{L}2$ and $8\mathbf{L}6$? In fact, $n\mathbf{L}m$ if and only if m is the n^{th} number in the decimal expansion of $\pi = 3.14159265\dots$. Since there may often be a relation without you being able to see it, we have adopted a completely general definition of *relation*, which does not assume any visible relationship between the objects.

Relations are ubiquitous in mathematics, and you know many already:

Examples 0.2.47 (a) Consider the relation $\leq$ on $\mathbb{R}$:

$$\leq = \{(x, y) \in \mathbb{R}^2 : x \leq y\}$$

(b) If A is a set, there is a similar relation $\subseteq$ on $\mathcal{P}(A)$:

$$\subseteq = \{(X, Y) \in \mathcal{P}(A) \times \mathcal{P}(A) : X \subseteq Y\}$$

(c) The divisor relation on $\mathbb{Z}$: We say that $n|m$ if and only if m is a multiple of n. It is described by the set

$$\{(n, m) \in \mathbb{Z}^2 : \text{There exists } a \in \mathbb{Z} \text{ such that } m = an\}$$

(d) Congruency modulo n: Two integers are *congruent modulo* n if they leave the same remainder when divided by n. For example $3, 8, 13, 18\dots$ all leave a remainder of 3 when divided by 5, and thus they are congruent modulo 5. Symbolically, we say that

$$a \equiv b \mod n \text{ if } n|(b - a)$$

(e) Perpendicularity is a relation between vectors in $\mathbb{R}^3$. We have

$$(x_1, x_2, x_3) \perp (y_1, y_2, y_3) \Longleftrightarrow x_1y_1 + x_2y_2 + x_3y_3 = 0$$

(f) *Equality* is a relation. Actually, it is several relations, all denoted by the same symbol =. Thus we have equality of numbers, of vectors, of sets, etc. If A is a set, then the relation of equality on A is called the *identity* relation. It is just

$$\Delta_A = \{(a, a) : a \in A\}$$

□

Note that function is a special kind of relation: We defined a function $f : A \times B$ to be a subset of $A \times B$ with the additional property that for every a there is exactly one b such that $a\, f\, b$. Instead of writing $a\, f\, b$, however, we write $f(a) = b$.

If $\mathbf{R}$ is a relation from A to B, then its *inverse* $\mathbf{R}^{-1}$ is a relation from B to A. It is defined by:

$$b\, \mathbf{R}^{-1} a \iff a \mathbf{R} b$$

i.e. $(b, a) \in \mathbf{R}^{-1}$ if and only if $(a, b) \in \mathbf{R}$.

If R is a relation from A to B and S is a relation from B to C, we can define a relation $S \circ R$ from A to C as follows:

$$(a, c) \in S \circ R \quad \Leftrightarrow \exists b \in B[(a, b) \in R \wedge (b, c) \in S]$$

Thus

$$a(S \circ R)c \text{ iff there is } b \in B \text{ such that } a\, R\, b\, S\, c$$

Note the change in order!!
$S \circ R$ is called the *composition* of S with R.

Exercise 0.2.48 If the relations R, S are functions, then their composition as relations is the same as their composition as functions.
Similarly, if the relation R is a bijective function, then the inverse R^{-1} of R as a relation is the same as its inverse as a function.

□

Examples 0.2.49 (a) If X is the set of all people, and if $\mathbf{P}$ is the relation "parent of", then $\mathbf{P}^{-1}$ is the relation "child of".

(b) Similarly, if $\mathbf{P}$ is the relation "parent of", then $\mathbf{P} \circ \mathbf{P}$ is the relation "grandparent of": For $a(\mathbf{P} \circ \mathbf{P})c$ if and only if there is b such that $a\mathbf{P}b$ and $b\mathbf{P}c$.

(c) Moreover, $\mathbf{S} = \mathbf{P} \circ \mathbf{P}^{-1} - \Delta_X$ is the relation "sibling of": For $a(\mathbf{P} \circ \mathbf{P}^{-1})c$ if and only if there is a b such that $a\, \mathbf{P}^{-1}\, b\, \mathbf{P}\, c$, i.e. iff a is the child of b and b is the parent of c, i.e. iff a, c have a common parent. Thus $(a, c) \in \mathbf{S}$ implies that a, c have a common parent. Since $(a, c) \in \mathbf{S}$ implies $(a, c) \notin \Delta_X$, we see that $a \neq c$, and thus that a, c are brother and/or sister.

(d) $\leq^{-1} = \geq$, since $b \geq a$ if and only if $a \leq b$.

(e) n divides m if and only if m is a multiple of n. Thus the "multiple of" relation is the inverse of the "divisor of" relation.

(f) Perpendicularity between vectors is its own inverse, i.e. $\perp^{-1} = \perp$: $\vec{x} \perp \vec{y}$ iff $\vec{y} \perp \vec{x}$.

□

Exercises 0.2.50 1. Let W be the set of all women, and let S, M be relations from W to W described as follows: aSb iff a us a sister of b; aMb iff a is a mother of b. Describe

(a) $M \circ S$;

(b) $(M \circ S)^{-1}$;

(c) S^{-1} and M^{-1};

(d) $S^{-1} \circ M^{-1}$

2. Suppose that R is a relation from A to B and that S is a relation form B to C. Show that

$$(S \circ R)^{-1} = R^{-1} \circ S^{-1}$$

□

There are two important classes of relations in mathematics, namely *equivalence relations* and *partial orderings*. Equivalence relations have many of the same properties of $=$, and partial orderings have similar properties to $\leq$ and $\subseteq$.

Definition 0.2.51 Suppose that $\mathbf{R}$ is a relation from on a set A. R is said to be

(i) *reflexive* if $a\mathbf{R}a$ for all $a \in A$.

(ii) *symmetric* if $a\mathbf{R}b$ implies $b\,\mathbf{R}a$ for all $a, b \in A$.

(iii) *antisymmetric* if $a\mathbf{R}b$ *and* $b\,\mathbf{R}a$ together imply $a = b$, for all $a, b \in A$.

(iv) *transitive* if $a\mathbf{R}b$ and $b\mathbf{R}c$ together imply $a\mathbf{R}c$, for all $a, b, c \in A$.

An *equivalence relation* is a reflexive, symmetric, transitive relation.
A *partial ordering* is a reflexive, antisymmetric, transitive relation.

□

Exercises 0.2.52 1. $=$ is an equivalence relation on any set.

2. $\leq$ is a partial ordering on the set of reals.

3. $\subseteq$ is a partial ordering on $\mathcal{P}(A)$.

4. Congruency modulo n is an equivalence relation on $\mathbb{Z}$. (Recall that $a \equiv b \mod n$ if and only if a, b leave the same remainder when divided by n, if and only if $a - b$ is divisible by n.)

5. The divisor relation $n|m$ on $\mathbb{Z}$ is reflexive and transitive, but not symmetric, nor antisymmetric.

6. Define a relation $\mathbf{L}_1$ on $\mathbb{R}$ by:

$$\vec{x}\,\mathbf{L}_1\vec{y} \text{ if } |x| \leq |y|$$

Is $\mathbf{L}_1$ is a partial ordering?

7. Is $\perp$ an equivalence relation or a partial ordering (on $\mathbb{R}^3$)?

8. Let R be a relation on a set A.

(a) $\Delta_A \subseteq R$ iff R is reflexive.

(b) $R = R^{-1}$ iff R is symmetric.

(c) $R \cap R^{-1} = \Delta_A$ if and only if R is antisymmetric.

(d) $R \circ R \subseteq R$ if and only if R is transitive.

□

Let's take a look at equivalence relations from another angle: They are very closely related to *partitions*.

Definition 0.2.53 Let A be a set. A family $\mathcal{A} = \{A_i : i \in I\}$ is called a **partition** of A provided that

(i) The A_i are **mutually disjoint**, i.e. if $i \neq j$, then $A_i \cap A_j = \emptyset$ for all $i, j \in I$.

(ii) $\bigcup_I A_i = A$

□

Thus $\{A_i : i \in I\}$ is a partition of A provided that every element of A belongs to exactly one A_i. If $\{A_i : i \in I\}$ is a partition of A, then we can define an equivalence relation $\approx$ on A by:

$$a \approx b \iff a, b \text{ belong to the same } A_i$$

Exercise 0.2.54 Prove that $\approx$ is an equivalence relation.

□

On the other hand, if $\approx$ is an equivalence relation on A, then $\approx$ behave roughly like $=$. When we lump together all elements that are the same under $\approx$, we get an *equivalence class*.

Definition 0.2.55 Let $\approx$ be an equivalence relation on A. For each $a \in A$, define the *equivalence class* $E(a)$ of a as follows:

$$E(a) = \{b \in A : a \approx b\}$$

□

Note that $E(a) = E(b)$ if and only if $a \approx b$. If $a \not\approx b$, then $E(a) \cap E(b) = \emptyset$. Thus the sets $E(a)$ are either equal or disjoint. Hence the set $\{E(a) : a \in A\}$ is a partition of A.

Exercise 0.2.56 Verify the above statements.

□

Examples 0.2.57 (a) If $\approx$ is the identity relation on A, then the equivalence classes are singletons: $E(a) = \{a\}$.

(b) Suppose that $\approx$ is congruency modulo 3. Then the equivalence classes are $A_1 = \{\dots, -6, -3, 0, 3, 6, \dots\}$, $A_2 = \{\dots, -5, -2, 1, 4, 7, \dots\}$ and $A_3 = \{\dots, -4, -1, 2, 6, 8, \dots\}$. The elements of A_1 leave remainder 0 when divided by 3, those of A_2 leave remainder 1, and those of A_3 leave remainder 2. Note that A_1, A_2, A_3 are mutually disjoint, and that $A_1 \cup A_2 \cup A_3 = \mathbb{Z}$.

(c) Let $\approx$ be the "equal length" relation $\mathbf{L}_2$ on $\mathbb{R}^3$. The equivalence classes are spheres centred at the origin.

□

0.2.4 Countable and Uncountable Sets

In this section, we investigate the idea of the *size* or *cardinality* of a set. For finite sets, we can determine the size of a set by counting its elements. Thus for example, the set $\{a, b, c\}$ has cardinality 3 (it has 3 elements). We are going to extend this idea of counting to obtain the size to infinite sets, and we will show that infinity comes in many sizes.

Let's explore the idea of *counting*: For the moment, let $\mathbf{n} = \{1, 2, \ldots, n\}$ be the set of the first n natural numbers. To say that $A = \{a, b, c\}$ has 3 elements is equivalent to saying that there is a one-to-one correspondence between the sets A and $\mathbf{3}$. Indeed, this is the heart of the idea of counting: When we count the elements of A, we are setting up a bijection between A and $\mathbf{3}$. We go "a first, b second, c third". This is equivalent to a map $f : A \cong \mathbf{3}$ defined by $f(a) = 1, f(b) = 2, f(c) = 3$. Thus the idea of counting the elements of a finite set X involves finding a bijection between X and some $\mathbf{n}$. If there is a bijection from X to $\mathbf{n}$, then X has n elements.

Now for some reason, mathematicians often like to start counting at zero. In the mathematical literature, the sets $\mathbf{n}$ are therefore often defined as

$$\mathbf{n} = \{0, 1, 2, \ldots, n-1\}$$

This is the convention that we shall adopt henceforth.

It is obvious that two finite sets A and Δ have the same size if and only if there is a one-to-one correspondence $f : A \cong \Delta$. We don't even have to count A and Δ to know that they have the same number of elements. If $A = \{a, b, c, d\}$ and $\Delta = \{\alpha, \beta, \gamma, \delta\}$, then the existence of the bijection $f : A \cong \Delta$ given by

$$f(a) = \beta, f(b) = \delta, f(c) = \alpha, f(d) = \gamma$$

is sufficient to show that A and Δ have the same number of elements. It doesn't tell us that this number is 4.

Thus two sets have the same size if and only if there is a bijection between them; we can bypass the idea of number. This is important, because we cannot actually *count* infinite sets. But we can establish bijective correspondences between infinite sets. We shall adopt this idea as our basic idea of size.

Definition 0.2.58 *We define an equivalence relation $\approx$ between sets as follows: If A, B are sets, we say that $A \approx B$ if and only if there is a bijection from A to B. If $A \approx B$, we say that A and B have the same* cardinality*. We may also indicate this by saying $|A| = |B|$.*

□

Note that having the same cardinality is an *equivalence relation* between sets, i.e. that

(i) $|A| = |A|$ (Reflexivity)

(ii) If $|A| = |B|$, then $|B| = |A|$ (Symmetry)

(iii) If $|A| = |B|$ and $|B| = |C|$, then $|A| = |C|$ (Transitivity)

Exercise 0.2.59 Prove this assertion. (Note that the assertion is *not obvious*: When we say that $|A| = |B|$, we are not actually claiming that there are two equal numbers. What we *are* saying is that there is a bijection from A to B. To prove (i), for example, you have to find a bijection from A to A.)

□

Examples 0.2.60 (a) Two finite sets have the same cardinality if and only if they have the same number of elements.

(b) For finite sets, if A is a *proper subset* of B, then $|A| < |B|$. This breaks down completely for infinite sets. Consider, for example, the sets $\mathbb{N}$ and $\mathbb{Z}$. It is certainly true that $\mathbb{N} \subset \mathbb{Z}$. However, the map $\mathbb{N} \xrightarrow{f} \mathbb{Z}$ defined by

$$f(n) = \begin{cases} \dfrac{n}{2} & \text{if } n \text{ is even} \\ -\dfrac{n-1}{2} & \text{if } n \text{ is odd} \end{cases}$$

is a bijection: $f(1) = 0, f(2) = 1, f(3) = -1, f(4) = 2, f(5) = -2, f(6) = 3 \ldots$. (Note that we are zig–zagging from the positive integers to the negative integers.) Thus $\mathbb{N}$ and $\mathbb{Z}$ have the same cardinality, even though $\mathbb{N}$ seems to contain fewer elements than $\mathbb{Z}$.

(c) We also have $|\mathbb{Q}| = |\mathbb{N}|$. This can be seen as follows. Put the set of strictly positive rational numbers $\mathbb{Q}^+$ in an array

$$\begin{array}{ccccccc} 1/1 & 2/1 & 3/1 & 4/1 & 5/1 & \ldots \\ 1/2 & 2/2 & 3/2 & 4/2 & 5/2 & \ldots \\ 1/3 & 2/3 & 3/3 & 4/3 & 5/3 & \ldots \\ 1/4 & 2/4 & 3/4 & 4/4 & 5/4 & \ldots \\ 1/5 & 2/5 & 3/5 & 4/5 & 5/5 & \ldots \\ \vdots & \vdots & \vdots & \vdots & \vdots & \end{array}$$

We can then trace a zig–zag path that moves through all the rational numbers as follows. Start at the top line and move diagonally down to the left until you reach the leftmost line. Repeat. We thus obtain a sequence

$$\frac{1}{1}, \frac{2}{1}, \frac{1}{2}, \frac{3}{1}, \frac{2}{2}, \frac{1}{3}, \frac{4}{1}, \frac{3}{2}, \frac{2}{3}, \frac{1}{4}, \frac{5}{1} \ldots$$

All of the strictly positive rational numbers occur in this sequence, and they all occur infinitely many times. For example, $\frac{1}{1}, \frac{2}{2}, \frac{3}{3} \ldots$ lie along the diagonal, and they are all equal. To obtain a bijection from $\mathbb{N}$ to $\mathbb{Q}^+$, we follow the above sequence of rationals,

but we omit any number that has already occurred to ensure that the function is one-to-one, i.e. we *prune* away the repeated values. We therefore define the function $\mathbb{N} \xrightarrow{f} \mathbb{Q}^+$ by

$$f(1) = \frac{1}{1},\ f(2) = \frac{2}{1},\ f(3) = \frac{1}{2},\ f(4) = \frac{3}{1},\ f(5) = \frac{1}{3},\ f(6) = \frac{4}{1}, \ldots$$

Note that $f(5) \neq \frac{2}{2}$, which is after $f(4) = \frac{3}{1}$ in the sequence, because $\frac{2}{2} = \frac{1}{1}$ has already occurred as $f(1)$. Then f is a bijection from $\mathbb{N}$ to $\mathbb{Q}^+$. Now even though we haven't found a *formula* for f, it is nevertheless a perfectly good function, and all its values can be calculated. Can you see that $f(16) = \frac{2}{5}$?

In the same way, we can set up a bijection g from $\mathbb{N}$ to the negative rationals. Just put $g(n) = -f(n)$. Finally, we can define a bijection $h : \mathbb{N} \longrightarrow \mathbb{Q}$ using f, g and another zig–zag: We define

$$h(1) = 0, h(2) = f(1),\ h(3) = g(1),\ h(4) = f(2),$$
$$h(5) = g(2), h(6) = f(3),\ h(7) = g(3), \ldots$$

Again, we have no formula for h, but it is certainly a well–defined function, and all its values can be calculated. Check that $h(23) = -\frac{1}{5}$.

(d) If A is any set, finite or infinite, then $\mathcal{P}(A) \approx \mathbf{2}^A$.(Recall that $\mathbf{2}^A$ is the set of all functions from A to $\mathbf{2} = \{0, 1\}$). This can be seen as follows: If $B \subseteq A$, define the *indicator function* (or *characteristic function*) $I_B : A \longrightarrow B$ by

$$I_B(a) = \begin{cases} 1 \text{ if } a \in B \\ 0 \text{ else} \end{cases}$$

Clearly $I_B = I_C$ if and only if $B = C$, and so the map $\mathcal{I} : \mathcal{P}(A) \longrightarrow \mathbf{2}^A$ defined by $\mathcal{I}(B) = I_B$ is an injection. Now suppose that $\chi \in \mathbf{2}^A$, i.e. $A \xrightarrow{\chi} \{0, 1\}$. Define a subset $B \subseteq A$ by

$$a \in B \Longleftrightarrow \chi(a) = 1$$

It is clear that $\mathcal{I}(B) = I_B = \chi$, and thus that $\mathcal{I}$ is surjective as well. This proves that $|\mathcal{P}(A)| = |\mathbf{2}^A|$.

□

Definition 0.2.61 A set A is said to be *countable* if it is either finite or can be put into a one-to-one correspondence with the natural numbers, i.e. if $|A| = \mathbf{n}$ for some $n \in \mathbb{N}$, or $|A| = |\mathbb{N}|$.

□

Remarks 0.2.62 (a) Basically a set A is countable if its elements can be indexed by the natural numbers, i.e. if it *can* be written as $A = \{a_n : n \in \mathbb{N}\}$. For if A is countable and not finite, then there is a bijection $\mathbb{N} \xrightarrow{f} A$, and we can take $a_n = f(n)$. Conversely, if $A = \{a_n : n \in \mathbb{N}\}$ is infinite, we can define a bijection from $\mathbb{N}$ to A by letting $f(n) = a_n$ (although here some *pruning* is necessary if the a_n aren't all distinct; see Example 0.2.60(c)).

(b) In Example 0.2.60, we proved that the sets $\mathbb{Z}$ and $\mathbb{Q}$ are countable sets.

(c) The "zig–zag" technique, used above to prove that the rational numbers are countable, is often very useful.

□

A very basic question that arises is the following: Are all infinite sets countable? The answer is **"No!"**

Example 0.2.63 We show that the unit interval $I = [0, 1]$ is *uncountable*, i.e. that we cannot find an enumeration

$$I = \{x_n : n \in \mathbb{N}\}$$

The proof is by *contradiction*: Suppose that we *can* find such an enumeration $I = \{x_1, x_2, x_3, x_4, \dots\}$, i.e that every real number in $[0, 1]$ is equal to x_n for some n. Now every number x_n has a decimal expansion of the form

$$x_n = 0.x_{n1}x_{n2}x_{n3}x_{n4}x_{n5}\dots$$

where x_{nm} is the m^{th} number in the decimal expansion of x_n. Of course some real numbers have two distinct decimal expansions, a terminating one and a non–terminating one. For example, $1.0000\cdots = 0.9999\dots$. We will choose the non–terminating decimal expansions for our x_n.

We now create a new real number x from the x_n by a process called *diagonalization*. We choose $a_n \in \{1, 2, \dots, 9\}$ such that the following hold:

$$a_1 \neq x_{11}, a_2 \neq x_{22}, a_3 \neq x_{33}, \dots, a_n \neq x_{nn}, \dots$$

To avoid a situation where we obtain a number x with a terminating decimal expansion, we haven't permitted $a_n = 0$; this is just a technicality. We can now define x: Put

$$x = 0.a_1a_2a_3a_4\dots$$

Here comes the heart of the argument: Clearly $x \in I = [0, 1]$. Now if I can be written as a list $\{x_1, x_2, x_3, \dots\}$, then there must be some n such that $x = x_n$. But the first decimal place of x differs from the first decimal place of x, since $a_1 \neq x_{11}$; hence $x \neq x_1$. Similarly, the second decimal place of x differs from the second decimal place of x_2, since $a_2 \neq x_{22}$;

hence $x \neq x_2$. We can continue in this way to show that $x \neq x_n$ for any $n \in \mathbb{N}$, i.e. x is not on the list $\{x_1, x_2, x_3, \dots\}$.

This proves the result! Given any list $x_1, x_2, x_3, \dots$ of real numbers in $[0,1]$, we now have a technique for producing a new real number x that is not on the list. It thus follows that no such list can contain all the real numbers in $[0,1]$, i.e. there is no bijection from $\mathbb{N}$ to $[0,1]$.

□

Remarks 0.2.64 Cantor, who discovered the above argument for the uncountability of the reals, wrote to a friend

"I see it, but I don't believe it."

□

Hence there are uncountable sets. Clearly $\mathbb{R}$ is also uncountable, because otherwise we could find an enumeration $\{r_1, r_2, r_3, \dots\}$ of $\mathbb{R}$. By omitting any reals which are not in $[0,1]$, we could prune this into an enumeration of $[0,1]$.

The fact that $\mathbb{R}$ is uncountable causes much trouble in analysis. We shall see some more examples of uncountable sets later on.

Definition 0.2.65 *If A, B are sets, we say that the cardinality of A is less than or equal to the cardinality of B, and write*

$$|A| \leq |B|$$

if there is an injection from A into B. We write $|A| < |B|$ if $|A| \leq |B|$, but $|A| \neq |B|$, i.e. if there is an injection from A to B, but no bijection.

□

The idea is that $|A| < |B|$ if and only if A has "fewer" elements than $|B|$. Clearly the following holds:

Proposition 0.2.66 *(a) If $A \subseteq B$, then $|A| \leq |B|$.*

(b) If $|A| \leq |B|$ and $|B| \leq |C|$, then $|A| \leq |C|$.

(c) If $|A| \leq |B|$, then $|\mathcal{P}(A)| \leq |\mathcal{P}(B)|$.

(d) If $|A| \leq |B|$, then $|C^A| \leq |C^B|$

□

Exercise 0.2.67 Prove the above proposition.

□

In fact, the $\leq$–relation is a *partial ordering* between sets: Reflexivity is obvious, and transitivity was left to the exercise above. The main thing that needs to be shown is *antisymmetry*:

Theorem 0.2.68 *(Schröder–Bernstein Theorem) Suppose that $|A| \leq |B|$ and that $|B| \leq |A|$. Then $|A| = |B|$.*

⊣

We will omit the proof. It can be found in any text–book on set theory. Again, I must stress that this result is **not obvious**, because $|A|, |B|$ aren't really numbers. What we *have* to do is show that if there exists an injection from A to B, and an injection from B to A, then there exists a bijection from A to B.

Proposition 0.2.69 *(a) If A is countable, and if B is a subset of A, then B is countable.*

(b) If A, B are countable, then $A \times B$ is countable.

(c) If A, B are countable, the $A \cup B$ is countable.

(d) If $\mathcal{A} = \{A_n : n \in \mathbb{N}\}$ is a family of countable sets, then $\bigcup_n A_n$ is countable.

Proof: (a) If $\{a_n : n \in \mathbb{N}\}$ is an enumeration of A, we can obtain an enumeration of B by pruning the elements of A which are not in B. This can be accomplished inductively as follows. Let $b_1 = a_n$, where n is the least positive integer such that $a_n \in B$. Suppose now that b_m has been defined and that $b_m = a_i$. Then let $b_{m+1} = a_j$, where j is the least positive integer $> i$ such that $a_j \in B$. Clearly $\{b_m : m \in \mathbb{N}\}$ is an enumeration of B.
(b) One can easily prove that $\mathbb{N} \times \mathbb{N}$ is countable by copying Example 0.2.60(c). Just form an array

$$\begin{array}{cccc} (1,1) & (2,1) & (3,1) & (4,1) & \dots \\ 1,2) & (2,2) & (3,2) & (4,2) & \dots \\ (1,3) & (2,3) & (3,3) & (4,3) & \dots \\ \vdots & \vdots & \vdots & \vdots & \end{array}$$

and zig–zag your way across this array. Let $A \xrightarrow{f} \mathbb{N}$ and $B \xrightarrow{g} \mathbb{N}$ be bijections. Then the map $h : A \times B \longrightarrow \mathbb{N} \times \mathbb{N}$ defined by $h(a, b) = (f(a), g(b))$ is clearly a bijection. Hence $|A \times B| = |\mathbb{N} \times \mathbb{N}| = |\mathbb{N}|$ as required.
(c) follows from (d).
(d) Again we use a zig–zag: Let $\{a_{n1}, a_{n2}, a_{n3}, \dots\}$ be a listing of the elements of A_n. Form an array

$$\begin{array}{cccc} a_{11} & a_{12} & a_{13} & \dots \\ a_{21} & a_{22} & a_{23} & \dots \\ a_{31} & a_{32} & a_{33} & \dots \\ \vdots & \vdots & \vdots & \end{array}$$

and take a path which goes through each element once, pruning duplications.

⊣

This proposition shows that you can't make uncountable sets using finite products and countable unions. You can, however, make uncountable sets using infinite products and the powerset operation.

Proposition 0.2.70 *Let A be a set. Then $|A| < |\mathcal{P}(A)| = \mathbf{2}^A$.*

Proof: We know that for any set A, $\mathcal{P}(A) \approx \mathbf{2}^A$, by Example 0.2.60(d). So it suffices to show that $|A| < |\mathcal{P}(A)|$. Now it is obvious that there is an injection from A into $\mathcal{P}(A)$: The map $a \mapsto \{a\}$ will do the trick. Hence certainly $|A| \leq |\mathcal{P}(A)|$. Suppose now that there is a bijection $A \xrightarrow{f} \mathcal{P}(A)$, and define $A_a \subseteq A$ by $A_a = f(a)$. Since a bijection is surjective, we must have $\mathcal{P}(A) = \{A_a : a \in A\}$. We shall now show that this is impossible.

Note that $A_a \subseteq A$ and that $a \in A$. Thus it may happen that $a \in A_a$, or it may not. Define B to be the set of all a for which it does not happen, i.e. let

$$a \in B \Longleftrightarrow a \notin A_a$$

The $B \subseteq A$. Since the listing $\{A_a : a \in A\}$ is supposed to be a *complete* list of all the elements of $\mathcal{P}(A)$, there must be some $b \in A$ such that $B = A_b$. However, if $b \in B$, then $b \notin A_b$, and if $b \notin B$, then $b \in A_b$. Hence B cannot equal A_b, since b belongs to one set, but not the other. The assumption that $\{A_a : a \in A\}$ is a complete list of all the subsets of A therefore leads to a contradiction.

⊣

The following proposition is very useful:

Proposition 0.2.71 *Suppose that A, B are infinite sets, and that $|A| \leq |B|$. Then:*

(a) $|A \cup B| = |B|$

(b) $|A \times B| = |B|$

(c) $|A^B| = |\mathbf{2}^B|$.

⊣

We omit the proof, which can be found in almost any textbook on set theory.

Exercises 0.2.72 (1) Prove that if A is uncountable and B is countable, then $A - B$ is uncountable.

(2) Prove that $\mathbb{R} \approx [0, 1]$. (Hint: Note that all non-empty finite intervals have the same cardinality as $[0, 1]$. First prove that all closed intervals have the same cardinality. If I is any finite interval, whether open, closed, or half–open, we can find closed intervals I_1, I_2 such that $I_1 \subseteq I \subseteq I_2$. The Schröder–Bernstein Theorem then implies that they all have the same cardinality. Now define a map $\mathbb{Z} \times [0, 1) \xrightarrow{f} \mathbb{R}$ as follows: If $n \in \mathbb{Z}$ and if $x \in [0, 1)$, then define

$$f(n, x) = n + x$$

This is clearly a bijection. Now $|\mathbb{Z}| \leq |[0, 1]| = |[0, 1)|$, and therefore $\mathbb{R} \approx \mathbb{Z} \times [0, 1) \approx [0, 1]$.)

□

Example 0.2.73 $\mathbb{R} \approx \mathbf{2}^{\mathbb{N}}$. Here's a clever way of seeing this: Every real number has a *dyadic* or *binary* expansion, as opposed to a decimal expansion. The dyadic expansion uses only the numbers 0 and 1. For example, if we have a dyadic number 101.011, this is

$$\underbrace{101.011}_{\text{dyadic}} = 1 \cdot 2^2 + 0 \cdot 2^1 + 1 \cdot 2^0 + 0 \cdot 2^{-1} + 1 \cdot 2^{-2} + 1 \cdot 2^{-3} = \underbrace{5.375}_{\text{decimal}}$$

Hence every real number can be turned into a sequence of zeroes and ones, and vice versa. Now such a sequence is essentially just a map from $\mathbb{N}$ to $\mathbf{2}$. For example, the sequence $1011001\ldots$ can be thought of as the function $f : \mathbb{N} \longrightarrow \mathbf{2}$ which has $f(1) = 1, f(2) = 0, f(3) = 1, f(4) = 1, f(5) = 0, f(6) = 0, f(7) = 1 \ldots$. There are only two problems: (i) Where to put the decimal point, and (ii) Some real numbers have two distinct dyadic expansions. For example, $\frac{1}{2} = 0.1000\cdots = 0.01111\ldots$. However, if a real number has two dyadic expansions, it is easy to see that the one must eventually end in all 0's, and the other must end in all 1's. We call the former expansion terminating, and the latter expansion non–terminating.

We now overcome the two problems as follows: Since $\mathbb{R} \approx [0, 1)$, it suffices to show that $\mathbf{2}^{\mathbb{N}} \approx [0, 1)$. Now given any $x \in [0, 1)$, its non–terminating dyadic expansion $x = 0.x_1x_2x_3\ldots$ will give us a sequence $x_1, x_2, x_3 \ldots$ of zeroes and ones. This clearly gives us an injective map $F : [0, 1) \longrightarrow \mathbf{2}^{\mathbb{N}}$. It is, however, not a surjective map. But only the sequences that eventually end in all zeroes have been missed out, and there are only countably many such. To be precise, if $\mathcal{X} = \text{range } F$, then $\mathcal{Y} = \mathbf{2}^{\mathbb{N}} - \mathcal{X}$ is countable. Hence $|\mathbf{2}^{\mathbb{N}}| = |\mathcal{X} \cup \mathcal{Y}| = |\mathcal{X}| = |(0, 1]| = |\mathbb{R}|$.

□

Example 0.2.74 (The Cantor set) The *Cantor set* is a subset of $[0,1]$ which is constructed as follows: Let $C_0 = [0,1]$. It is a single interval of length 1. Now let C_1 be C_0 with its *middle third* removed, i.e. $C_1 = [0, \frac{1}{3}] \cup [\frac{2}{3}, 1]$. Thus C_1 consists of two disjoint intervals, each of length $\frac{1}{3}$. Now remove the middle thirds of these two intervals to form C_2, i.e. $C_2 = [0, \frac{1}{9}] \cup [\frac{2}{9}, \frac{1}{3}] \cup [\frac{2}{3}, \frac{7}{9}] \cup [\frac{8}{9}, 1]$. Then C_2 is a disjoint union of 4 intervals, each of length $\frac{1}{9}$. Continue in this way, removing the middle thirds of each of the intervals comprising C_n to form C_{n+1}. It follows that C_n consists of 2^n intervals, each of length $(\frac{1}{3})^n$, and thus that $\lambda(C_n) = (\frac{2}{3})^n$. Finally, let $C = \bigcap_{n=0}^{\infty} C_n$. C is the Cantor set.

How much of $[0,1]$ did we remove when we created C? First we removed an interval of length $\frac{1}{3}$, then we removed 2 intervals, each of length $\frac{1}{9}$. After that, we removed 4 intervals, each of length $\frac{1}{27}$, etc. Thus we have removed disjoint sets with a total length

$$\frac{2^0}{3^1} + \frac{2^1}{3^2} + \frac{2^2}{3^3} + \cdots = \frac{1}{3} \sum_{k=0}^{\infty} \left(\frac{2}{3}\right)^k$$

which is a geometric series with sum $\frac{1}{3} \cdot \frac{1}{1 - \frac{2}{3}} = 1$. It seems, therefore, that we have removed the entire length of the unit interval $[0,1]$. There is no length left.

Nevertheless, C is not empty; in fact, C is uncountable. Here is one way to see this: Every real number $a \in [0,1]$ can be written as an infinite sum $\sum_{i=1}^{\infty} \frac{a_i}{3^i}$, where $a_i = 0, 1$ or 2. Thus the *ternary expansion* (as opposed to *decimal expansion*) of a is $0.a_1a_2a_3\ldots$. For example, $\frac{1}{3} = 0.1000\ldots$, $\frac{5}{9} = \frac{1}{3} + \frac{2}{9} = 0.1200\ldots$, etc. A little thought will reveal that the Cantor set is formed by removing all numbers which have a 1 occurring in their ternary expansion. Thus C_1 is formed by removing all numbers which have a 1 in the first decimal place, C_2 is formed by removing all numbers in C_1 which have a 1 in the second decimal place, and so on. Thus the Cantor set is just the set of all numbers a in $[0,1]$ which can be written as a sum $\sum_{i=1}^{\infty} \frac{a_i}{3^i}$, where $a_i = 0$ or 2, but not 1. There is a bijection $\Phi : 2^{\mathbb{N}} \longrightarrow C$ defined as follows: If $f \in 2^{\mathbb{N}}$, then $\Phi(f)$ is the number with decimal expansion $0.a_1a_2a_3\ldots$, where $a_n = 0$ if $f(n) = 0$, and $a_n = 2$ if $f(n) = 1$. Hence $|C| = |2^{\mathbb{N}}| = |\mathbb{R}|$.

□

0.3 Prelude to an Axiomatic Development of the Real Number System

0.3.1 Why we need Axioms

Consider the following questions:

Question 1: Many years ago, you were taught the following algorithm for multiplying

two numbers:

$$\begin{array}{r} 23 \\ \underline{\ 17} \\ 161 \\ \underline{230} \\ 391 \end{array}$$

Why does this algorithm work?

Question 2: Why is $-1 \times -1 = 1$? Alternatively, why is the product of two negative numbers a positive number?

If you think that these are silly questions, think again. The answers to these questions are *not* obvious. You are merely so used to the answers that the questions never occur to you.

An explanation for why the multiplication algorithm works might go along the following lines:

$$\begin{aligned} 23 \times 17 &= 23 \cdot (7 + 10) \\ &= 23 \cdot 7 + 23 \cdot 10 \\ &= (20 + 3) \cdot 7 + (20 + 3) \cdot 10 \\ &= [20 \cdot 7 + 3 \cdot 7] + [20 \cdot 10 + 3 \cdot 10] \\ &= [140 + 21] + [200 + 30] \qquad = 161 + 230 \\ &= 391 \end{aligned}$$

To do this calculation, we performed the following operations:

(i) We used the fact that $a \cdot (b + c) = a \cdot b + a \cdot c$ several times.

(ii) We retrieve certain results, like $3 \cdot 7 = 21$, from memory. Such results were learnt by rote, in the form of multiplication tables. Thus all the values of $a \times b$ for $1 \leq a, b \leq 10$ are stored in a mental look–up table.
The values in the look–up table were determined *empirically*, i.e. by observation. To see that $7 \times 8 = 56$, take 8 small bags, each containing 7 stones, and empty them into a big bag. If you now count the number of stones in the big bag, you will get 56. That's just a fact that's been observed over and over again, in many different places and at many different times.

(iii) We use the fact that multiplying a number by 10 is accomplished by adding a zero to the end of that number. Thus $20 \cdot 10 = 200$.

(iv) To calculate the value of a term such as $20 \cdot 7$ (which is not in the mental look–up table), we have to argue that $20 \cdot 7 = 7 \cdot 20 = 7 \cdot (2 \cdot 10) = (7 \cdot 2) \cdot 10 = 14 \cdot 10 = 140$. Thus, in addition to the look–up table and the multiply–by–ten rule, we also used the following facts about multiplication: $a \cdot b = b \cdot a$, and $a \cdot (b \cdot c) = (a \cdot b) \cdot c$.

(v) We used another algorithm (also learnt long ago) for adding numbers, such as $161 + 230$. Try and justify that algorithm yourself.

As you can see, in order to explain why the multiplication algorithm works, you need to invoke quite a few simpler results about addition and multiplication. Question 1 is not as obvious as it looks! As for Question 2, you should be able to explain why $-1 \times -1 = 1$ by the end of this chapter.

Now note the following (empirically verifiable) facts: Human beings have a certain intuition (or idea) about non–physical objects called numbers. These numbers can be combined in various ways to form new numbers, e.g. they can be added and multiplied. Moreover, there are some simple rules which govern the combination of numbers, e.g.

(i) The product of two numbers does not depend on where or when the multiplication is performed.

(ii) $a + b = b + a \qquad ab = ba$

(iii) $a + (b + c) = (a + b) + c \qquad a(bc) = (ab)c$

(iv) $a(b + c) = ab + ac$

et cetera. Our aim is now to find a set of rules, or *axioms*, which completely captures our intuition about the arithmetic of the reals. In other words, we seek a set of rules which (1) is in accord with our intuition about arithmetic, and (2) is sufficiently rich that any informal, intuitive arithmetic argument can be made *formal*, i.e. we can reach the same conclusion by applying no intuition at all, but just the axioms.

Why do we need axioms? For several reasons.

- Axioms tend to be simple, and most people will accept them as in agreement with their intuition. Thus the axioms are a common starting point for all people. People who disagree on the axioms are probably talking about different things.

- The agreed–upon rules can be applied over and over again, to arbitrary levels of complexity. Any two people who agree on the (simple) axioms will also agree on the (complicated) conclusions that may be reached by formal application of those axioms.

  On the other hand, intuition becomes less and less reliable as we increase the level of complexity, and thus conclusions obtained solely by a intuition are more suspect. For example, you and I may agree that Euclid's 5 axioms for geometry are in accordance with our intuition of *space*. These axioms are simple, and difficult to disbelieve. *You* may have a powerful intuition, however: You intuit that the square of the (length of) the hypotenuse of a right–angled triangle is equal to the sum of the squares of the other two sides. But *my* intuition is far less developed than yours: I just don't see it, and so I don't believe you. Should you provide a step–by–step argument, starting from our common ground (the 5 axioms), using only commonly agreed rules, I will be forced to admit that your intuition is correct. In this way, I can verify the truth of your assertion myself, and don't just have to take your word for it.

- If we use the axiomatic method, we are constantly aware of our assumptions. It therefore becomes much simpler to discern similarities and differences between various mathematical objects and operations. This will make the arguments *portable* (in the Computer Science sense — arguments (computer code) can easily be moved from one problem to (platform) to another).
  For example, the rules $a \cdot (b+c) = a \cdot b + a \cdot c$ and $a \cdot (b \cdot c) = (a \cdot b) \cdot c$ apply not only to multiplication of numbers, but also to multiplication of matrices. Thus any proposition that can be proved about numbers, using just those rules, will also be true for matrices. However, the rule $a \cdot b = b \cdot a$ does not hold for matrices.

- Finally, axioms allow us to circumvent metaphysical speculation about the nature and existence of mathematical objects. What, for example *is* a real number? Is it an irreducible, or is it made up of simpler things? This question was first given a satisfactory answer in 1872. Indeed, it was given *two* different but satisfactory answers in that year, by Dedekind and Cantor. In each case, the real numbers are "constructed" from some previously constructed, simpler, objects, e.g. the rational numbers.

  And the rational numbers are in turn constructed from the positive integers, which are constructed from the empty set — We'll go into a bit more depth later in this chapter, but for a proper explanation, you will need to read a book on set theory.

  Thus there is no single answer to the question: "What is a real number?". But the exact nature of the reals is unimportant for mathematical purposes. What is important is how they *behave*, i.e. how they can be recombined, using various operations, to form new numbers. The axioms are essentially just a description of such behaviour, and though the three constructions disagree about the essential nature of the reals, they *do* agree on how they behave.

(i) George Cantor held that a real number *is* a set of sequences of rational numbers. Thus, for example, $\sqrt{2}$ is just the set of all rational sequences that converge to $\sqrt{2}$. (This definition may seem circular, but the apparent circularity can be removed.)

(ii) For Richard Dedekind, a real number *is* an ordered pair of sets of rational numbers. Thus,

$$\sqrt{2} = (\{q \in \mathbb{Q} : q^2 \leq 2\}, \{q \in \mathbb{Q} : Q^2 > 2\})$$

(iii) John Horton Conway regards a real number as a *game* played by two individuals. I won't elaborate.

0.3.2 A Brief Note on the Philosophy of Mathematics*

In the previous section, we got quite philosophical. Before we continue, it's a good idea to have a look at the main mathematical schools of thought. Our main concern is with the schools of *Platonism* and *Formalism*. For completeness, I present brief and over–simplified caricatures of these and other leading schools below. If you want an honest exposition, you'd better consult a book on the philosophy of mathematics.

- **Platonism:** The belief that mathematical objects, though not part of the physical universe, nevertheless have an existence which is independent of the human mind. We speak of making mathematical *discoveries*, which suggests that we are somehow able to observe mathematical objects.

  Some people also speak about mathematical *creations*. Did da Vinci have the freedom to create a Mona Lisa with a grimace, rather than a smile? Probably. He preferred a smile. But did Pythagoras have the freedom to create a right–angled triangle for which his theorem — that the square of the hypotenuse is the sum of the squares on the right–angle sides — fails?

- **Logicism:** An attempt to reduce mathematics to logic. Frege and Russell are the main protagonists. The four–volume *Principia Mathematica*, by Russell and Whitehead, is the most well–known exposition of this school.

- **Constructivism:** Constructivists require that, in order to show that a mathematical object exists, one must explicitly show *how* to construct it. This leads them to reject of the *Law of the Excluded Middle*, which states that for any statement φ, either φ is true, or not-φ is true, i.e. there's no "middle" between φ and not-φ. It also leads to the rejection of proofs by contradiction. There are many varieties of constructivism; the most well–known is *Intuitionism*, whose main proponent was Brouwer.
  For example, suppose that S is a set, and that φ is a property. In *classical logic*, the following statement is true:

  Either there is a member of S that has property φ,
  or every member of S has property not-φ. $\quad (*)$

  For example, consider the *Riemann Hypothesis*, which states that all the (complex) roots of the Riemann ζ–function have a real part equal to $\frac{1}{2}$:

  $$1 + \frac{1}{2^z} + \frac{1}{3^z} + \cdots = 0 \quad \Longrightarrow \quad \mathrm{Re}(z) = \frac{1}{2}$$

  This is currently *the* unsolved problem in mathematics. For the classical logician, the Riemann Hypothesis is either true or false — we just don't know which. Not so for the constructivist: To say that it is either true or false, we must either prove that it is true, or show that it is false. So the constructivist does not accept $(*)$. For this statement to hold, we must either show how to construct a member of S with the property φ, or we must show that each member of S has the property not-φ.
  If S is a finite set, we could, in principle, look at each of the elements of S in turn, to see if it satisfies φ. If S is infinite, however, this is generally not possible. Constructivists are happy to apply the Law of the Excluded Middle to finite sets; its application to infinite sets they regard as a colossal mistake — an unwarranted and unjustifiable extrapolation of methods of reasoning designed for the finite to the infinite. Indeed, some constructivists deny the existence of infinite objects altogether.
  As another example, suppose that we want to prove that every real cubic polynomial $p(x) = x^3 + ax^2 + bx + c$ has a real root. One way to do it is to appeal to the *Intermediate Value Theorem*: We see that $p(x) > 0$ for all sufficiently large positive x, and thus that $p(x)$ lies above the X–axis, for all sufficiently large positive x. Similarly, $p(x) < 0$ for all sufficiently large negative x, so that $p(x)$ lies below the X–axis, for all sufficiently large negative x. Hence, since $p(x)$ is continuous,

there must be a place where $p(x)$ cuts the X–axis, and that place would be a root of $p(x)$. This proof is *non–constructive*: We've shown that there is a root, but we haven't shown how to find it.

- **Formalism:** This school of thought dates back to Hilbert in the late 19^{th} century. At that time, certain paradoxes in *set theory* shook the foundations of mathematics, and mathematicians were suddenly confronted with the possibility that their subject is inconsistent, i.e. self–contradictory.
The most famous of these is *Russell's paradox*. If we admit a naive concept of set — a set is any old collection of objects — then it is possible for a set to belong to itself. For example if

$$A = \text{The set of all objects that can be defined in English using fewer than twenty words}$$

then $A \in A$, because we've just defined A using fewer than twenty words. Now consider a set of sets R, defined as follows:

$$A \in R \quad \text{iff} \quad A \notin A$$

Since R is a set, we may legitimately ask if it belongs to itself. By definition of R, we see that

$$R \in R \quad \text{iff} \quad R \notin R$$

If R belongs to R, then it doesn't; and if R does not belong to R, then it does! This paradox, usually credited to the logicist Russell in 1899??, but already noted by Zermelo in 1896?? caused quite a lot of concern.

Hilbert, the most powerful mathematician of his era, set up a programme aimed at proving the internal consistency of mathematics by so–called *finitist means*. The formalist regards mathematics as a one–player game, rather like Patience (or Free-cell). A proof of a statement ψ, for example, is merely a sequence of statements $\varphi_1, \varphi_2, \ldots, \varphi_n$, ending with $\varphi_n = \psi$. Each φ_k must either be an axiom, or must be obtained from previous φ_j by certain permitted "moves", or rules of deduction.
For example, a commonly used rule of deduction is *Modus Ponens*:

$$\text{From } \varphi \to \psi \text{ and } \varphi \text{ deduce } \psi$$

The mathematician seeking to prove the statement ψ is like the player of Patience, trying out permitted sequences of moves until she hits upon a sequence that works. The idea behind the Hilbert Programme is to *formalize* mathematics:

- Write all mathematics in a *formal language*;
- Reduce all proofs to *formal deductions*;
- Show that no contradictions can be derived within this formal system.

Hilbert had hoped that it would be possible to show that all of mathematics could be thus reduced, and proved consistent. Thus commenced a massive attempt to formalize and axiomatize all of mathematics, and the way that we now do and see mathematics has been heavily influenced by the Hilbert programme.

One of the first branches of mathematics to be formalized was *set theory*, where the paradoxes had been found. The Zermelo–Fraenkel axioms of set theory banish Russell's paradox, but at a costs: It

is no longer possible for a set to belong to itself, and the intuition of a set as "just any old collection of objects" had to be abandoned. It was found possible to squeeze nearly all of mathematics inside the formal system of axiomatic set theory. Unfortunately, Hilbert's student Gödel proved in 1931 that the Hilbert programme was doomed to failure. In a paper entitled *On formally undecidable statements in Principia Mathematica and related systems* he showed that in any formalist reduction of mathematics there would be statements that are *true*, but *unprovable*. He also showed that no such reduction is capable of proving its own consistency. This proved the death knell for the Hilbert programme, though not for the Formalist school. (Gödel himself was a Platonist.) With hindsight, it is remarkable how close the Hilbert programme came to succeeding.

Platonism and Formalism disagree (quite violently) about the *nature* of mathematical objects: For the Platonist, these have an existence independent of the human mind; by the mysterious faculty of intuition we apprehend basic truths (axioms) about mathematical objects, and then use reason to deduce ever more complex truths (theorems). For the Formalist, there are no mathematical objects, just rules for transforming one string of symbols into another.

Nevertheless, both schools of thought *agree* on what constitutes a valid mathematical proof. The average practicing mathematician has been described as "a Platonist on weekdays, and a Formalist on Sundays. That is, when he is doing mathematics he is convinced hat he is dealing with an objective reality whose properties he is attempting to determine. But then, when challenged to give a philosophical account of this reality, he finds it easiest to pretend that he does not believe in it after all."

And that's what our position will be. To begin with, we will be firm Platonists: We will believe in the objective existence of the real number system, and use our intuition to apprehend basic truths. Recognising that our intuition is fallible, however, we won't let it stray to far. Instead, we opt soon to formalise our intuitions into a system of axioms. After that, intuition is only allowed to make suggestions, and only those statements that can be seen to admit a *formal* proof will be admitted to the status of theoremhood.

0.3.3 Logic, Formal Languages, Quantifiers

The aim of this section is to cover the bare minimum about formal theories — just enough to make our construction of the real number system intelligible.

A *formal language* is a collection of $\mathcal{L}$ whose *logical symbols* include

- **Logical Connectives**

$\wedge$	and
$\vee$	or
$\rightarrow$	implies
$\leftrightarrow$	if and only if
$\neg$	not

It is enough to use just two connectives, e.g. $\wedge$ and $\neg$. We can then define the remainder by

$$\varphi \vee \psi \equiv \neg(\neg\varphi \wedge \neg\psi)$$
$$\varphi \to \psi \equiv \neg\varphi \vee \psi$$
$$\varphi \leftrightarrow \psi \equiv (\varphi \to \psi) \wedge (\psi \to \varphi)$$

Just a reminder: $\vee$ is *inclusive–or*: $p \vee q$ is true if and only if at least one of p, q is true, possibly both.

- **Quantifiers**

$\forall$	For all
$\exists$	There exists

We have

$$\forall x\varphi \equiv \neg\exists x(\neg\varphi) \qquad \exists x\varphi \equiv \neg\forall x(\neg\varphi)$$

- **Variables**
  $x, y, z, x_1, x_2, x_3 \ldots$

- **Identity relation**
  A special binary relation symbol denoted $=$.

Logical symbols have the same meaning, regardless of context. $\mathcal{L}$ also has *non–logical* symbols, whose meaning depends on context:

- **Relation symbols**
  For example, if we want to talk about partial orderings, we will want a symbol $\leq$; if we want to talk about sets, we will want symbols $\in$ and $\subseteq$.

- **Function symbols**
  For example, if we want to talk about arithmetic, we will want binary function symbols $+, \times$. We may want unary function symbols $-, ^{-1}$. If we want to talk about sets, we will want binary function symbols $\cap, \cup$, unary function symbols $^c, \mathcal{P}$;

- **Constant symbols**
  These are specially named elements, and are often regarded as *nullary* function symbols. For example, if we want to talk about addition, a *distinguished element* denoted by 0 plays an important role. If we want to talk about sets, the set $\emptyset$ deserves its own name.

A formal language will generally not contain all of the above non–logical symbols, only those needed to talk about the domain of discourse. $\mathcal{L}$ will also have brackets $(,), [,]$, etc.

The symbols of a formal language may be "strung" together to form two types: *terms* and *formulas*.

- **Terms** are defined as follows:

(i) Every variable and every constant is a term;

(ii) If $t_1, \ldots, t_n$ are terms, and if F is an n–ary function symbol, then $F(t_1, \ldots, t_n)$ is a term;

(iii) A string is a term only if it can be shown to be so by a finite number of applications of (i) and (ii);

- **Formulas** are defined as follows:

(i) If $t_1, \ldots, t_n$ are terms, and if R is an n–ary relation symbol, then $R(t_1, \ldots, t_n)$ is a formula. (This includes the case where R is the logical binary relation symbol $=$).

(ii) If φ, ψ are formulas, then so are $(\varphi \wedge \psi)$, $(\varphi \vee \psi), (\varphi \to \psi), (\varphi \leftrightarrow \psi)$;

(iii) If φ is a formula, then so is $\neg\varphi$;

(iv) If φ is a formula and x is a variable, then $\forall x\varphi$ and $\exists x\varphi$ are formulas;

(v) A string is a formula only if it can be shown to be so by a finite number of applications of (i)-(iv).

We often omit brackets when there is no danger of confusion. Moreover, we may also abbreviate $\forall x \forall y \varphi$ by $\forall x, y \varphi$.

If φ is a formula, we write $\varphi(x, y, z)$ to show that the variables of φ are (amongst) x, y, z.

Example 0.3.1 Partial orderings
Consider the following language $\mathcal{L}$: In addition to the logical symbols, $\mathcal{L}$ has a single binary relation symbol $\leq$. There are no function and constant symbols. Thus the only terms of $\mathcal{L}$ are the variables. Some example of formulas are

$$x \leq y, \quad \forall x(x \leq y \wedge y \leq z) \to \exists z(\neg(z \leq x))$$

The theory of partial orderings has the following axioms

(i) $\forall x(x \leq x)$;

(ii) $\forall x, y(x \leq y \wedge y \leq x \to x = y)$;

(iii) $\forall x, y, z(x \leq y \wedge y \leq z \to x \leq z)$.

This theory has many *interpretations*. One is the two–element chain $C_2 = \{0, 1\}$ with $0 \leq 1$. This is a linear ordering, i.e. it satisfies the axiom $\forall x, y(x \leq y \vee y \leq x)$. Another example is the powerset $\mathcal{P}(A)$ of a set A, where $\leq$ is interpreted as "subset". This ordering is non–linear if A has more than one element.

Thus different structures may satisfy the same axioms.

□

Example 0.3.2 Peano Arithmetic
We give here another example of a formal theory. In addition to the logical symbols, $\mathcal{L}$ has the following non–logical symbols:

- Binary function symbols + and .;
- A unary function symbol S;

- A constant symbol 0.

Some examples of terms are:

$$x,\ S(x),\ 0,\ S(S(S(0))),\ x+y,\ (x+S(y))\cdot z,\ (y\cdot S(z))+w$$

Some examples of formulas are:

$$x+y=0,\ \forall x\exists y(x+S(S(0))=y),\ \forall x(x=S(y)\vee\neg(S(x)=y))$$

Peano arithmetic is a formal theory in the language $\mathcal{L}$. The axioms are:

(i) $\forall x[\neg(S(x)=0)]$;

(ii) $\forall x\forall y(S(x)=S(y)\to x=y)$;

(iii) $\forall x(x+0=x)$;

(iv) $\forall x\forall y(x+S(y)=S(x+y))$;

(v) $\forall x(x\cdot 0=0)$;

(vi) $\forall x\forall y(x\cdot S(y)=x\cdot y+x)$;

(vii) For every formula $\varphi(x_0,\ldots,x_n)$, we have

$$\forall x_1\ldots\forall x_n[(\varphi(0,x_1,\ldots,x_n)\wedge(\forall x_0(\varphi(x_0,x_1,\ldots,x_n)\to\varphi(S(x_0),x_1,\ldots,x_n))\\ \to\forall x_0\varphi(x_0,x_1,\ldots,x_n)]$$

We will now show that in this system we can prove the following identity:

$$\forall x,y[x+y=y+x]$$

- We first show that for all x, $0+x=x$. Let $\varphi(x)$ be the formula

$$0+x=x$$

Then certainly $\varphi(0)$ is true, because $0+0=0$ by (iii). Now suppose that $\varphi(x)$ is true. Then by (iv)

$$0+S(x)=S(0+x)=S(x)$$

and so $\varphi(S(x))$ is also true. We have thus shown that $\varphi(0)$ holds, and that if $\varphi(x)$ holds, then so does $\varphi(S(x))$. Axiom (vii) allows us to deduce that $\forall x\varphi(x)$.

- Next, let $\psi(y,x)$ be the formula

$$x+S(y)=S(x)+y$$

Then $x+S(0)=S(x+0)=S(x)=0+S(x)$, by (iv) and what we've just shown. Hence $\psi(0,x)$ is true, for every x. Next, suppose that $\psi(y,x)$ is true for every x. Then

$$\begin{aligned} x+S(S(y)) &= S(x+S(y)) && \text{by (iv)}\\ &= S(S(x)+y) && \text{because } \psi(y,x)\\ &= S(x)+S(y) && \text{by (iv)} \end{aligned}$$

so that $\psi(S(y),x)$ is true, for every x.
By (vii), it follows that $\psi(y,x)$ is true for all y and all x.

- Finally, let $\xi(y, x)$ be the formula

$$x + y = y + x$$

Then we know that $\xi(0, x)$ is true, for all x, because $\varphi(x)$ is true for all x. Assume now that $\xi(y, x)$ is true for all x. Then

$$\begin{aligned} S(S(y) + x) &= S(y) + S(x) && \text{by (iv)} \\ &= y + S(S(x)) && \text{because } \psi(S(x), y) \\ &= S(S(x)) + y && \text{because } \xi(y, S(S(x))) \\ &= S(x) + S(y) && \text{because } \psi(y, S(x)) \\ &= x + S(S(y)) && \text{because } \psi(S(y), x) \\ &= S(x + S(y)) && \text{by (iv)} \end{aligned}$$

Thus by (ii), $S(y) + x = x + S(y)$. Hence $\xi(y, x) \to \xi(S(y), x)$, so that by (vii) we can conclude that $\xi(y, x)$ is true for all x, y.

Right now, you probably don't know what $S(x)$ actually *means*. Like good formalists,we've proved the commutativity of the binary operation $+$ by playing a game of deduction from the axioms. We invoked the mysterious symbol S in several places, without knowing its meaning.

The meaning, or *natural interpretation*, or *canonical model* of the Peano axioms is as follows: The axioms are "about" the natural numbers $\mathbb{N} = \{0, 1, 2, \dots\}$. The binary function symbol $+, \cdot$ are to be interpreted, respectively, as addition and multiplication. The unary symbol S is to be interpreted as *successor*: $S(x) = x + 1$. (However, there is no constant symbol 1.) The constant symbol 0 is to be interpreted as the number zero. Thus $S(0) = 1$ (1 is the successor of 0), $S(S(0)) = 2$, etc.

The first axiom says that 0 is not the successor of any other number. The second axiom says that if x, y have the same successor, then x, y are equal. You can interpret the other axioms yourself. Just note that (vii) is not a single axiom, but an infinite set of axioms, one for every formula φ. The *axiom schema* (vii) formalizes mathematical induction: If 0 has property φ, and if whenever x_0 has property φ, then also $S(x)$ has property φ, then we can conclude that every number x_0 has property φ.

Just like the axioms for partial orderings, the Peano axioms have many other interpretations as well. These are the so–called *non–standard* models of arithmetic.

□

In this section, I have presented the briefest possible introduction to formal theories, and I've taken numerous short cuts. If you want more extensive (and more accurate) coverage, you will have to consult a text on mathematical logic. We end this section with some brief comments on quantifiers and negation.

Consider the following formulas:

$$\forall x \exists y (y > x)$$

To check the truth of such a statement, it is convenient to regard it as a game between two players, $\forall$ and $\exists$. In this game, $\forall$ opens play and chooses and x. If $\exists$ can find a y such that $y > x$, then $\exists$ wins the game. If she can't, $\forall$ wins. The formula is true if $\exists$ can always win, i.e. if $\exists$ has a *winning strategy*; else, the formula is false.

Whether or not the formula is true or false depends on where it is played. If we play it on the natural numbers $\mathbb{N}$, then $\exists$ has a winning strategy: If $\forall$ chooses x, the $\exists$ can choose $y = x + 1$. Then $y > x$. This works for any x that $\forall$ might choose. Hence $\exists$ has a winning strategy: The formula is true for $\mathbb{N}$.

Suppose, however, that the game is played not in $\mathbb{N}$, but on the two–element chain $C_2 = \{0, 1\}$. Then if $\forall$ chooses $x = 1$, $\exists$ cannot find a $y \in C_2$ with $y > x$. Hence $\forall$ has a winning strategy, and the statement is false for C_2.

Exercise 0.3.3 Give a similar analysis for the statement

$$\exists y \forall x (y > x)$$

□

Finally, a note about negating quantifiers: A negation sign can "creep" past a quantifier, but it *flips* the quantifier in the process:

$$\neg \forall x \varphi \equiv \exists x (\neg \varphi) \qquad \neg \exists x \varphi \equiv \forall x (\neg \varphi)$$

For example,

$$\begin{aligned} \neg [\forall x \exists y (y > x)] &\equiv \exists x \neg [\exists y (y > x)] \\ &\equiv \exists x \forall y (y \not> x) \end{aligned}$$