1 Sets and structures

The object of mathematical physics is to describe the physical world in purely mathematical terms. Although it had its origins in the science of ancient Greece, with the work of Archimedes, Euclid and Aristotle, it was not until the discoveries of Galileo and Newton that mathematical physics as we know it today had its true beginnings. Newton's discovery of the calculus and its application to physics was undoubtedly the defining moment. This was built upon by generations of brilliant mathematicians such as Euler, Lagrange, Hamilton and Gauss, who essentially formulated physical law in terms of differential equations. With the advent of new and unintuitive theories such as relativity and quantum mechanics in the twentieth century, the reliance on mathematics moved to increasingly recondite areas such as abstract algebra, topology, functional analysis and differential geometry. Even classical areas such as the mechanics of Lagrange and Hamilton, as well as classical thermodynamics, can be lifted almost directly into the language of modern differential geometry. Today, the emphasis is often more structural than analytical, and it is commonly believed that finding the right mathematical structure is the most important aspect of any physical theory. Analysis, or the consequences of theories, still has a part to play in mathematical physics – indeed, most research is of this nature – but it is possibly less fundamental in the total overview of the subject.

When we consider the significant achievements of mathematical physics, one cannot help but wonder why the workings of the universe are expressable at all by rigid mathematical 'laws'. Furthermore, how is it that purely human constructs, in the form of deep and subtle mathematical structures refined over centuries of thought, have any relevance at all? The nineteenth century view of a clockwork universe regulated deterministically by differential equations seems now to have been banished for ever, both through the fundamental appearance of probabilities in quantum mechanics and the indeterminism associated with chaotic systems. These two aspects of physical law, the deterministic and indeterministic, seem to interplay in some astonishing ways, the impact of which has yet to be fully appreciated. It is this interplay, however, that almost certainly gives our world its richness and variety. Some of these questions and challenges may be fundamentally unanswerable, but the fact remains that mathematics seems to be the correct path to understanding the physical world.

The aim of this book is to present the basic mathematical structures used in our subject, and to express some of the most important theories of physics in their appropriate mathematical setting. It is a book designed chiefly for students of physics who have the need for a more rigorous mathematical education. A basic knowledge of calculus and linear algebra, including matrix theory, is assumed throughout, but little else. While different students will of course come to this book with different levels of mathematical sophistication, the reader should be able to determine exactly what they can skip and where they must take pause. Mathematicians, for example, may be interested only in the later chapters, where various theories of physics are expressed in mathematical terms. These theories will not, however, be developed at great length, and their consequences will only be dealt with by way of a few examples.

The most fundamental notion in mathematics is that of a *set*, or 'collection of objects'. The subject of this chapter is *set theory* – the branch of mathematics devoted to the study of sets as abstract objects in their own right. It turns out that every mathematical structure consists of a collection of sets together with some *defining relations*. Furthermore, as we shall see in Section 1.3, such relations are themselves defined in terms of sets. It is thus a commonly adopted viewpoint that all of mathematics reduces essentially to statements in set theory, and this is the motivation for starting with a chapter on such a basic topic.

The idea of sets as collections of objects has a non-rigorous, or 'naive' quality, although it is the form in which most students are introduced to the subject [1–4]. Early in the twentieth century, it was discovered by Bertrand Russell that there are inherent self-contradictions and paradoxes in overly simple versions of set theory. Although of concern to logicians and those mathematicians demanding a totally rigorous basis to their subject, these paradoxes usually involve inordinately large self-referential sets – not the sort of constructs likely to occur in physical contexts. Thus, while special models of set theory have been designed to avoid contradictions, they generally have somewhat artificial attributes and naive set theory should suffice for our purposes. The reader's attention should be drawn, however, to the remarks at the end of Section 1.5 concerning the possible relevance of fundamental problems of set theory to physics. These problems, while not of overwhelming concern, may at least provide some food for thought.

While a basic familiarity with set theory will be assumed throughout this book, it nevertheless seems worthwhile to go over the fundamentals, if only for the sake of completeness and to establish a few conventions. Many physicists do not have a good grounding in set theory, and should find this chapter a useful exercise in developing the kind of rigorous thinking needed for mathematical physics. For mathematicians this is all bread and butter, and if you feel the material of this chapter is well-worn ground, please feel free to pass on quickly.

1.1 Sets and logic

There are essentially two ways in which we can think of a **set** S. Firstly, it can be regarded as a collection of mathematical objects $a, b, \dots$, called **constants**, written

$$S = \{a,\ b, \dots\}.$$

The constants $a,\ b, \dots$ may themselves be sets and, indeed, some formulations of set theory *require* them to be sets. Physicists in general prefer to avoid this formal nicety, and find it much more natural to allow for 'atomic' objects, as it is hard to think of quantities such as *temperature* or *velocity* as being 'sets'. However, to think of sets as consisting of lists of objects is only suitable for finite or at most countably infinite sets. If we try putting the real numbers into a list we encounter the Cantor diagonalization problem – see Theorems 1.4 and 1.5 of Section 1.5.

The second approach to set theory is much more general in character. Let $P(x)$ be a *logical proposition* involving a **variable** x. Any such proposition symbolically defines a set

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

which can be thought of as symbolically representing the collection of all x for which the proposition $P(x)$ is true. We will not attempt a full definition of the concept of logical proposition here – this is the business of formal logic and is only of peripheral interest to theoretical physicists – but some comments are in order. Essentially, logical propositions are statements made up from an alphabet of symbols, some of which are termed **constants** and some of which are called **variables**, together with logical connectives such as **not**, **and**, **or** and **implies**, to be manipulated according to rules of standard logic. Instead of 'P implies Q' we frequently use the words '**if** P **then** Q' or the symbolic representation $P \Rightarrow Q$. The statement 'P **if and only if** Q', or 'P **iff** Q', symbolically written $P \Leftrightarrow Q$, is a shorthand for

$$(P \Rightarrow Q) \text{ and } (Q \Rightarrow P),$$

and signifies logical equivalence of the propositions P and Q. The two **quantifiers** $\forall$ and $\exists$, said **for all** and **there exists**, respectively, make their appearance in the following way: if $P(x)$ is a proposition involving a variable x, then

$$\forall x(P(x)) \quad \text{and} \quad \exists x(P(x))$$

are propositions.

Mathematical theories such as *set theory*, *group theory*, etc. traditionally involve the introduction of some new symbols with which to generate further logical propositions. The theory must be complemented by a collection of logical propositions called **axioms** for the theory – statements that are taken to be automatically **true** in the theory. All other true statements should in principle follow by the rules of logic.

Set theory involves the introduction of the new phrase **is a set** and new symbols $\{\dots \mid \dots\}$ and $\in$ defined by:

(Set1) If S is any constant or variable then 'S is a set' is a logical proposition.
(Set2) If $P(x)$ is a logical proposition involving a variable x then $\{x \mid P(x)\}$ is a set.
(Set3) If S is a set and a is any constant or variable then $a \in S$ is a logical proposition, for which we say a **belongs to** S or a **is a member of** S, or simply a **is in** S. The negative of this proposition is denoted $a \notin S$ – said a **is not in** S.

These statements say nothing about whether the various propositions are true or false – they merely assert what are 'grammatically correct' propositions in set theory. They merely tell us how the new symbols and phrases are to be used in a grammatically correct fashion. The main axiom of set theory is: if $P(x)$ is any logical proposition depending on a variable x, then for any constant or variable a

$$a \in \{x \mid P(x)\} \Leftrightarrow P(a).$$

Every mathematical theory uses the equality symbol $=$ to express the identity of mathematical objects in the theory. In some cases the concept of mathematical identity needs a separate definition. For example **equality of sets** $A = B$ is defined through the *axiom of extensionality*:

Two sets A and B are equal if and only if they contain the same members. Expressed symbolically,

$$A = B \Leftrightarrow \forall a(a \in A \Leftrightarrow a \in B).$$

A **finite set** $A = \{a_1, a_2, \dots, a_n\}$ is equivalent to

$$A = \{x \mid (x = a_1) \text{ or } (x = a_2) \text{ or } \dots \text{ or } (x = a_n)\}.$$

A set consisting of just one element a is called a **singleton** and should be written as $\{a\}$ to distinguish it from the element a which belongs to it: $\{a\} = \{x \mid x = a\}$.

As remarked above, sets can be members of other sets. A set whose elements are all sets themselves will often be called a **collection** or **family** of sets. Such collections are often denoted by script letters such as $\mathcal{A}$, $\mathcal{U}$, etc. Frequently a family of sets $\mathcal{U}$ has its members **indexed** by another set I, called the **indexing set**, and is written

$$\mathcal{U} = \{U_i \mid i \in I\}.$$

For a finite family we usually take the indexing set to be the first n *natural numbers*, $I = \{1, 2, \dots, n\}$. Strictly speaking, this set must also be given an axiomatic definition such as *Peano's axioms*. We refer the interested reader to texts such as [4] for a discussion of these matters.

Although the finer details of logic have been omitted here, essentially all concepts of set theory can be constructed from these basics. The implication is that all of mathematics can be built out of an alphabet for constants and variables, parentheses $(\dots)$, logical connectives and quantifiers together with the rules of propositional logic, and the symbols $\{\dots \mid \dots\}$ and $\in$. Since mathematical physics is an attempt to express physics in purely mathematical language, we have the somewhat astonishing implication that all of physics should also be reducible to these simple terms. Eugene Wigner has expressed wonderment at this idea in a famous paper entitled *The unreasonable effectiveness of mathematics in the natural sciences* [5].

The presentation of set theory given here should suffice for all practical purposes, but it is not without logical difficulties. The most famous is *Russell's paradox*: consider the set of all sets which are not members of themselves. According to the above rules this set can be written $R = \{A \mid A \notin A\}$. Is R a member of itself? This question does not appear to have an answer. For, if $R \in R$ then by definition $R \notin R$, which is a contradiction. On the other hand, if $R \notin R$ then it satisfies the criterion required for membership of R; that is, $R \in R$.

To avoid such vicious arguments, logicians have been forced to reformulate the axioms of set theory in a very careful way. The most frequently used system is the axiomatic scheme of *Zermelo and Fraenkel* – see, for example, [2] or the Appendix of [6]. We will adopt the 'naive' position and simply assume that the sets dealt with in this book do not exhibit the self-contradictions of Russell's monster.

1.2 Subsets, unions and intersections of sets

A set T is said to be a **subset** of S, or T is **contained in** S, if every member of T belongs to S. Symbolically, this is written $T \subseteq S$,

$$T \subseteq S \quad \text{iff} \quad a \in T \Rightarrow a \in S.$$

We may also say S is a **superset** of T and write $S \supset T$. Of particular importance is the **empty set** $\emptyset$, to which no object belongs,

$$\forall a \, (a \notin \emptyset).$$

The empty set is assumed to be a subset of any set whatsoever,

$$\forall S(\emptyset \subseteq S).$$

This is the default position, consistent with the fact that $a \in \emptyset \Rightarrow a \in S$, since there are no a such that $a \in \emptyset$ and the left-hand side of the implication is never true. We have here an example of the logical dictum that 'a false statement implies the truth of any statement'.

A common criterion for showing the equality of two sets, $T = S$, is to show that $T \subseteq S$ and $S \subseteq T$. The proof follows from the axiom of extensionality:

$$\begin{aligned} T = S &\iff (a \in T \Leftrightarrow a \in S) \\ &\iff (a \in T \Rightarrow a \in S) \text{ and } (a \in S \Rightarrow a \in T) \\ &\iff (T \subseteq S) \text{ and } (S \subseteq T). \end{aligned}$$

Exercise: Show that the empty set is unique; i.e., if $\emptyset'$ is an empty set then $\emptyset' = \emptyset$.

The collection of all subsets of a set S forms a set in its own right, called the **power set** of S, denoted 2^S.

Example 1.1 If S is a finite set consisting of n elements, then 2^S consists of one empty set $\emptyset$ having no elements, n singleton sets having just one member, $\binom{n}{2}$ sets having two elements, etc. Hence the total number of sets belonging to 2^S is, by the binomial theorem,

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

This motivates the symbolic representation of the power set.

Unions and intersections

The **union** of two sets S and T, denoted $S \cup T$, is defined as

$$S \cup T = \{x \mid x \in S \text{ or } x \in T\}.$$

The **intersection** of two sets S and T, denoted $S \cap T$, is defined as

$$S \cap T = \{x \mid x \in S \text{ and } x \in T\}.$$

Two sets S and T are called **disjoint** if no element belongs simultaneously to both sets, $S \cap T = \emptyset$. The **difference** of two sets S and T is defined as

$$S - T = \{x \mid x \in S \text{ and } x \notin T\}.$$

Exercise: If S and T are disjoint, show that $S - T = S$.

The union of an arbitrary (possibly infinite) family of sets $\mathcal{A}$ is defined as the set of all elements x that belong to some member of the family,

$$\bigcup \mathcal{A} = \{x \mid \exists S \text{ such that } (S \in \mathcal{A}) \text{ and } (x \in S)\}.$$

Similarly we define the **intersection** of $\mathcal{S}$ to be the set of all elements that belong to *every* set of the collection,

$$\bigcap \mathcal{A} = \{x \mid x \in S \text{ for all } S \in \mathcal{A}\}.$$

When $\mathcal{A}$ consists of a family of sets S_i indexed by a set I, the union and intersection are frequently written

$$\bigcup_{i \in I} \{S_i\} \quad \text{and} \quad \bigcap_{i \in I} \{S_i\}.$$

Problems

Problem 1.1 Show the distributive laws

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

Problem 1.2 If $\mathcal{B} = \{B_i \mid i \in I\}$ is any family of sets, show that

$$A \cap \bigcup \mathcal{B} = \bigcup \{A \cap B_i \mid i \in I\}, \quad A \cup \bigcap \mathcal{B} = \bigcap \{A \cup B_i \mid i \in I\}.$$

Problem 1.3 Let B be any set. Show that $(A \cap B) \cup C = A \cap (B \cup C)$ if and only if $C \subseteq A$.

Problem 1.4 Show that

$$A - (B \cup C) = (A - B) \cap (A - C), \quad A - (B \cap C) = (A - B) \cup (A - C).$$

Problem 1.5 If $\mathcal{B} = \{B_i \mid i \in I\}$ is any family of sets, show that

$$A - \bigcup \mathcal{B} = \bigcup \{A - B_i \mid i \in I\}.$$

Problem 1.6 If E and F are any sets, prove the identities

$$2^E \cap 2^F = 2^{E \cap F}, \quad 2^E \cup 2^F \subseteq 2^{E \cup F}.$$

Problem 1.7 Show that if $\mathcal{C}$ is any family of sets then

$$\bigcap_{X \in \mathcal{C}} 2^X = 2^{\bigcap \mathcal{C}}, \quad \bigcup_{X \in \mathcal{C}} 2^X \subseteq 2^{\bigcup \mathcal{C}}.$$

1.3 Cartesian products and relations

Ordered pairs and cartesian products

As it stands, there is no concept of order in a set consisting of two elements, since $\{a, b\} = \{b, a\}$. Frequently we wish to refer to an **ordered pair** (a, b). Essentially this is a set of two elements $\{a, b\}$ where we specify the *order* in which the two elements are to be written. A purely set-theoretical way of expressing this idea is to adjoin the element a that is to be regarded as the 'first' member. An ordered pair (a, b) can thus be thought of as a set consisting of $\{a, b\}$ together with the element a singled out as being the *first*,

$$(a, b) = \{\{a, b\}, a\}. \tag{1.1}$$

While this looks a little artificial at first, it does demonstrate how the concept of 'order' can be defined in purely set-theoretical terms. Thankfully, we only give this definition for illustrative purposes – there is essentially no need to refer again to the formal representation (1.1).

Exercise: From the definition (1.1) show that $(a, b) = (a', b')$ iff $a = a'$ and $b = b'$.

Similarly, an **ordered n-tuple** $(a_1, a_2, \dots, a_n)$ is a set in which the order of the elements must be specified. This can be defined inductively as

$$(a_1, a_2, \dots, a_n) = (a_1, (a_2, a_3, \dots, a_n)).$$

Exercise: Write out the ordered triple (a, b, c) as a set.

The **(cartesian) product** of two sets, $S \times T$, is the set of all ordered pairs (s, t) where s belongs to S and t belongs to T,

$$S \times T = \{(s, t) \mid s \in S \text{ and } t \in T\}.$$

The product of n sets is defined as

$$S_1 \times S_2 \times \cdots \times S_n = \{(s_1, s_2, \dots, s_n) \mid s_1 \in S_1, s_2 \in S_2, \dots, s_n \in S_n\}.$$

If the n sets are equal, $S_1 = S_2 = \cdots = S_n = S$, then their product is denoted S^n.

Exercise: Show that $S \times T = \emptyset$ iff $S = \emptyset$ or $T = \emptyset$.

Relations

Any subset of S^n is called an **n-ary relation** on a set S. For example,

unary relation ≡ 1-ary relation = subset of S
binary relation ≡ 2-ary relation = subset of $S^2 = S \times S$
ternary relation ≡ 3-ary relation = subset of $S^3 = S \times S \times S$, etc.

We will focus attention on binary relations as these are by far the most important. If $R \subseteq S \times S$ is a binary relation on S, it is common to use the notation aRb in place of $(a, b) \in R$.

Some commonly used terms describing relations are the following:

R is said to be a **reflexive** relation if aRa for all $a \in S$.
R is called **symmetric** if $aRb \Rightarrow bRa$ for all $a, b \in S$.
R is **transitive** if $(aRb \text{ and } bRc) \Rightarrow aRc$ for all $a, b, c \in S$.

Example 1.2 Let $\mathbb{R}$ be the set of all real numbers. The usual ordering of real numbers is a relation on $\mathbb{R}$, denoted $x \leq y$, which is both reflexive and transitive but not symmetric. The relation of strict ordering $x < y$ is transitive, but is neither reflexive nor symmetric. Similar statements apply for the ordering on subsets of $\mathbb{R}$, such as the integers or rational numbers. The notation $x \leq y$ is invariably used for this relation in place of the rather odd-looking $(x, y) \in \leq$ where $\leq \subseteq \mathbb{R}^2$.

Equivalence relations

A relation that is reflexive, symmetric and transitive is called an **equivalence relation**. For example, equality $a = b$ is always an equivalence relation. If R is an equivalence relation on a set S and a is an arbitrary element of S, then we define the **equivalence class corresponding to** a to be the subset

$$[a]_R = \{b \in S \,|\, aRb\}.$$

The equivalence class is frequently denoted simply by $[a]$ if the equivalence relation R is understood. By the reflexive property $a \in [a]$ – that is, equivalence classes 'cover' the set S in the sense that every element belongs to at least one class. Furthermore, if aRb then $[a] = [b]$. For, let $c \in [a]$ so that aRc. By symmetry, we have bRa, and the transitive property implies that bRc. Hence $c \in [b]$, showing that $[a] \subseteq [b]$. Similarly $[b] \subseteq [a]$, from which it follows that $[a] = [b]$.

Furthermore, if $[a]$ and $[b]$ are any pair of equivalence classes having non-empty intersection, $[a] \cap [b] \neq \emptyset$, then $[a] = [b]$. For, if $c \in [a] \cap [b]$ then aRc and cRb. By transitivity, aRb, or equivalently $[a] = [b]$. Thus any pair of equivalence classes are either disjoint, $[a] \cap [b] = \emptyset$, or else they are equal, $[a] = [b]$. The equivalence relation R is therefore said to **partition** the set S into disjoint equivalence classes.

It is sometimes useful to think of elements of S belonging to the same equivalence class as being 'identified' with each other through the equivalence relation R. The set whose elements are the equivalence classes defined by the equivalence relation R is called the **factor space**, denoted S/R,

$$S/R = \{[a]_R \,|\, a \in S\} \equiv \{x \,|\, x = [a]_R,\ a \in S\}.$$

Example 1.3 Let p be a positive integer. On the set of all integers $\mathbb{Z}$, define the equivalence relation R by mRn if and only if there exists $k \in \mathbb{Z}$ such that $m - n = kp$, denoted

$$m \equiv n \ (\text{mod } p).$$

This relation is easily seen to be an equivalence relation. For example, to show it is transitive, simply observe that if $m - n = kp$ and $n - j = lp$ then $m - j = (k + l)p$. The equivalence class $[m]$ consists of the set of integers of the form $m + kp$, $(k = 0, \pm 1, \pm 2, \dots)$. It follows that there are precisely p such equivalence classes, $[0],\ [1],\dots,[p-1]$, called the **residue classes modulo** p. Their union spans all of $\mathbb{Z}$.

Example 1.4 Let $\mathbb{R}^2 = \mathbb{R}\times\mathbb{R}$ be the cartesian plane and define an equivalence relation $\equiv$ on $\mathbb{R}^2$ by

$$(x, y) \equiv (x', y') \quad \text{iff} \quad \exists n, m \in \mathbb{Z} \text{ such that } x' = x + m,\ y' = y + n.$$

Each equivalence class $[(x, y)]$ has one representative such that $0 \le x < 1,\ 0 \le y < 1$. The factor space

$$T^2 = \mathbb{R}^2/\equiv\ = \{[(x, y)] \mid 0 \le x,\ y < 1\}$$

is called the **2-torus**. The geometrical motivation for this name will become apparent in Chapter 10.

Order relations and posets

The characteristic features of an 'order relation' have been discussed in Example 1.2, specifically for the case of the real numbers. More generally, a relation R on a set S is said to be a **partial order** on S if it is reflexive and transitive, and in place of the symmetric property it satisfies the 'antisymmetric' property

$$aRb \text{ and } bRa \implies a = b.$$

The ordering $\le$ on real numbers has the further special property of being a **total order**, by which it is meant that for every pair of real numbers x and y, we have either $x \le y$ or $y \le x$.

Example 1.5 The power set 2^S of a set S is partially ordered by the relation of set inclusion $\subseteq$,

$$\begin{gathered} U \subseteq U \text{ for all } U \in S, \\ U \subseteq V \text{ and } V \subseteq W \implies U \subseteq W, \\ U \subseteq V \text{ and } V \subseteq U \implies U = V. \end{gathered}$$

Unlike the ordering of real numbers, this ordering is *not* in general a total order.

A set S together with a partial order $\le$ is called a **partially ordered set** or more briefly a **poset**. This is an example of a *structured set*. The words 'together with' used here are a rather casual type of mathspeak commonly used to describe a set with an imposed structure. Technically more correct is the definition of a poset as an ordered pair,

$$\text{poset } S \equiv (S, \le)$$

where $\le\ \subseteq S \times S$ satisfies the axioms of a partial order. The concept of a poset could be totally reduced to its set-theoretical elements by writing ordered pairs (s, t) as sets of the form $\{\{s,\ t\},\ s\}$, etc., but this uninstructive task would only serve to demonstrate how simple mathematical concepts can be made totally obscure by overzealous use of abstract definitions.

Problems

Problem 1.8 Show the following identities:

$$(A \cup B) \times P = (A \times P) \cup (B \times P),$$
$$(A \cap B) \times (P \cap Q) = (A \times P) \cap (B \times P),$$
$$(A - B) \times P = (A \times P) - (B \times P).$$

Problem 1.9 If $\mathcal{A} = \{A_i \mid i \in I\}$ and $\mathcal{B} = \{B_j \mid j \in J\}$ are any two families of sets then

$$\bigcup \mathcal{A} \times \bigcup \mathcal{B} = \bigcup_{i \in I, j \in J} A_i \times B_j,$$
$$\bigcap \mathcal{A} \times \bigcap \mathcal{B} = \bigcap_{i \in I, j \in J} A_i \times B_j.$$

Problem 1.10 Show that both the following two relations:

$$\begin{aligned} (a, b) \le (x, y) \quad &\text{iff} \quad a < x \text{ or } (a = x \text{ and } b \le y) \\ (a, b) \preceq (x, y) \quad &\text{iff} \quad a \le x \text{ and } b \le y \end{aligned}$$

are partial orders on $\mathbb{R} \times \mathbb{R}$. For any pair of partial orders $\le$ and $\preceq$ defined on an arbitrary set A, let us say that $\le$ is *stronger* than $\preceq$ if $a \le b \;\rightarrow\; a \preceq b$. Is $\le$ stronger than, weaker than or incomparable with $\preceq$?

1.4 Mappings

Let X and Y be any two sets. A **mapping** φ from X to Y, often written $\varphi : X \to Y$, is a subset of $X \times Y$ such that for every $x \in X$ there is a *unique* $y \in Y$ for which $(x, y) \in \varphi$. By **unique** we mean

$$(x, y) \in \varphi \text{ and } (x, y') \in \varphi \implies y = y'.$$

Mappings are also called **functions** or **maps**. It is most common to write $y = \varphi(x)$ for $(x, y) \in \varphi$. Whenever $y = \varphi(x)$ it is said that x **is mapped to** y, written $\varphi : x \mapsto y$.

In elementary mathematics it is common to refer to the subset $\varphi \subseteq X \times Y$ as representing the *graph* of the function φ. Our definition essentially identifies a function with its graph. The set X is called the **domain** of the mapping φ, and the subset $\varphi(X) \subseteq Y$ defined by

$$\varphi(X) = \{y \in Y \mid y = \varphi(x),\ x \in X\}$$

is called its **range**.

Let U be any subset of Y. The **inverse image of** U is defined to be the set of all points of X that are mapped by φ into U, denoted

$$\varphi^{-1}(U) = \{x \in X \mid \varphi(x) \in U\}.$$

This concept makes sense even when the *inverse map* φ^{-1} does not exist. The notation $\varphi^{-1}(U)$ is to be regarded as one entire symbol for the inverse image set, and should not be broken into component parts.

Example 1.6 Let $\sin\colon \mathbb{R} \to \mathbb{R}$ be the standard sine function on the real numbers $\mathbb{R}$. The inverse image of 0 is $\sin^{-1}(0) = \{0, \pm\pi, \pm 2\pi, \pm 3\pi, \dots\}$, while the inverse image of 2 is the empty set, $\sin^{-1}(2) = \emptyset$.

An n**-ary function** from X to Y is a function $\varphi : X^n \to Y$. In this case we write $y = \varphi(x_1, x_2, \dots, x_n)$ for $((x_1, x_2, \dots, x_n), y) \in \varphi$ and say that φ has n **arguments** in the set S, although strictly speaking it has just one argument from the product set $X^n = X \times \dots \times X$.

It is possible to generalize this concept even further and consider maps whose domain is a product of n possibly different sets,

$$\varphi : X_1 \times X_2 \times \dots \times X_n \to Y.$$

Important maps of this type are the **projection maps**

$$\mathrm{pr}_i : X_1 \times X_2 \times \cdots X_n \to X_i$$

defined by

$$\mathrm{pr}_i : (x_1, x_2, \dots, x_n) \mapsto x_i.$$

If $\varphi : X \to Y$ and $\psi : Y \to Z$, the **composition** map $\psi \circ \varphi : X \to Z$ is defined by

$$\psi \circ \varphi\,(x) = \psi(\varphi(x)).$$

Composition of maps satisfies the **associative law**

$$\alpha \circ (\psi \circ \varphi) = (\alpha \circ \psi) \circ \varphi$$

where $\alpha : Z \to W$, since for any $x \in X$

$$\alpha \circ (\psi \circ \varphi)(x) = \alpha(\psi(\varphi(x))) = (\alpha \circ \psi)(\varphi(x)) = (\alpha \circ \psi) \circ \varphi(x).$$

Hence, there is no ambiguity in writing $\alpha \circ \psi \circ \varphi$ for the composition of three maps.

Surjective, injective and bijective maps

A mapping $\varphi : X \to Y$ is said to be **surjective** or **a surjection** if its range is all of T. More simply, we say φ is a mapping of X **onto** Y if $\varphi(X) = Y$. It is said to be **one-to-one** or **injective**, or **an injection**, if for every $y \in Y$ there is a *unique* $x \in X$ such that $y = \varphi(x)$; that is,

$$\varphi(x) = \varphi(x') \implies x = x'.$$

A map φ that is injective and surjective, or equivalently one-to-one and onto, is called **bijective** or **a bijection**. In this and only this case can one define the **inverse map** $\varphi^{-1} : Y \to X$ having the property

$$\varphi^{-1}(\varphi(x)) = x, \quad \forall x \in X.$$

Two sets X and Y are said to be in **one-to-one correspondence** with each other if there exists a bijection $\varphi : X \to Y$.

Exercise: Show that if $\varphi : X \to Y$ is a bijection, then so is φ^{-1}, and that $\varphi(\varphi^{-1}(x)) = x, \quad \forall x \in X$.

A bijective map $\varphi : X \to X$ from X onto itself is called a **transformation** of X. The most trivial transformation of all is the **identity map** id_X defined by

$$\mathrm{id}_X(x) = x, \quad \forall x \in X.$$

Note that this map can also be described as having a 'diagonal graph',

$$\mathrm{id}_X = \{(x, x) \,|\, x \in X\} \subseteq X \times X.$$

Exercise: Show that for any map $\varphi : X \to Y$, $\mathrm{id}_Y \circ \varphi = \varphi \circ \mathrm{id}_X = \varphi$.

When $\varphi : X \to Y$ is a bijection with inverse φ^{-1}, then we can write

$$\varphi^{-1} \circ \varphi = \mathrm{id}_X, \qquad \varphi \circ \varphi^{-1} = \mathrm{id}_Y.$$

If both φ and ψ are bijections then so is $\psi \circ \varphi$, and its inverse is given by

$$(\psi \circ \varphi)^{-1} = \varphi^{-1} \circ \psi^{-1}$$

since

$$\varphi^{-1} \circ \psi^{-1} \circ \psi \circ \varphi = \varphi^{-1} \circ \mathrm{id}_Y \circ \varphi = \varphi^{-1} \circ \varphi = \mathrm{id}_X.$$

If U is any subset of X and $\varphi : X \to Y$ is any map having domain X, then we define the **restriction** of φ to U as the map $\varphi\big|_U : U \to Y$ by $\varphi\big|_U(x) = \varphi(x)$ for all $x \in U$. The restriction of the identity map

$$i_U = \mathrm{id}_X\Big|_U : U \to X$$

is referred to as the **inclusion map** for the subset U. The restriction of an arbitrary map φ to U is then its composition with the inclusion map,

$$\varphi\big|_U = \varphi \circ i_U.$$

Example 1.7 If U is a subset of X, define a function $\chi_U : X \to \{0, 1\}$, called the **characteristic function** of U, by

$$\chi_U(x) = \begin{cases} 0 & \text{if } x \notin U, \\ 1 & \text{if } x \in U. \end{cases}$$

Any function $\varphi : X \to \{0, 1\}$ is evidently the characteristic function of the subset $U \subseteq X$ consisting of those points that are mapped to the value 1,

$$\varphi = \chi_U \quad \text{where} \quad U = \varphi^{-1}(\{1\}).$$

Thus the power set 2^X and the set of all maps $\varphi : X \to \{0, 1\}$ are in one-to-one correspondence.

Example 1.8 Let R be an equivalence relation on a set X. Define the **canonical map** $\varphi : X \to X/R$ from X onto the factor space by

$$\varphi(x) = [x]_R, \quad \forall x \in X.$$

It is easy to verify that this map is onto.

More generally, any map $\varphi : X \to Y$ defines an equivalence relation R on X by aRb iff $\varphi(a) = \varphi(b)$. The equivalence classes defined by R are precisely the inverse images of the singleton subsets of Y,

$$X/R = \{\varphi^{-1}(\{y\}) \mid y \in T\},$$

and the map $\psi : Y \to X/R$ defined by $\psi(y) = \varphi^{-1}(\{y\})$ is one-to-one, for if $\psi(y) = \psi(y')$ then $y = y'$ – pick any element $x \in \psi(y) = \psi(y')$ and we must have $\varphi(x) = y = y'$.

1.5 Infinite sets

A set S is said to be **finite** if there is a natural number n such that S is in one-to-one correspondence with the set $N = \{1, 2, 3, \dots, n\}$ consisting of the first n natural numbers. We call n the **cardinality** of the set S, written $n = \mathrm{Card}(S)$.

Example 1.9 For any two sets S and T the set of all maps $\varphi : S \to T$ will be denoted by T^S. Justification for this notation is provided by the fact that if S and T are both finite and $s = \mathrm{Card}(S)$, $t = \mathrm{Card}(T)$ then $\mathrm{Card}(T^S) = t^s$. In Example 1.7 it was shown that for any set S, the power set 2^S is in one-to-one correspondence with the set of characteristic functions on $\{1, 2\}^S$. As shown in Example 1.1, for a finite set S both sets have cardinality 2^s.

A set is said to be **infinite** if it is not finite. The concept of infinity is intuitively quite difficult to grasp, but the mathematician Georg Cantor (1845–1918) showed that infinite sets could be dealt with in a completely rigorous manner. He even succeeded in defining different 'orders of infinity' having a *transfinite arithmetic* that extended the ordinary arithmetic of the natural numbers.

Countable sets

The lowest order of infinity is that belonging to the natural numbers. Any set S that is in one-to-one correspondence with the set of natural numbers $\mathbb{N} = \{1, 2, 3, \dots\}$ is said to be **countably infinite**, or simply **countable**. The elements of S can then be displayed as a **sequence**, $s_1, s_2, s_3, \dots$ on setting $s_i = f^{-1}(i)$.

Example 1.10 The set of all integers $\mathbb{Z} = \{0, \pm 1, \pm 2, \dots\}$ is countable, for the map $f : \mathbb{Z} \to \mathbb{N}$ defined by $f(0) = 1$ and $f(n) = 2n$, $f(-n) = 2n + 1$ for all $n > 0$ is clearly a bijection,

$$f(0) = 1,\ f(1) = 2,\ f(-1) = 3,\ f(2) = 4,\ f(-2) = 5, \dots$$

Theorem 1.1 *Every subset of a countable set is either finite or countable.*

Proof: Let S be a countable set and $f : S \to \mathbb{N}$ a bijection, such that $f(s_1) = 1$, $f(s_2) = 2, \dots$ Suppose S' is an infinite subset of S. Let s'_1 be the first member of the sequence $s_1, s_2, \dots$ that belongs to S'. Set s'_2 to be the next member, etc. The map $f' : S' \to \mathbb{N}$

$$\begin{tikzcd}
(s_1, t_1) \arrow[r] & (s_1, t_2) \arrow[dl] & (s_1, t_3) \arrow[d] & \bullet & \bullet \\
(s_2, t_1) \arrow[d] & (s_2, t_2) \arrow[ur] & (s_2, t_3) \arrow[dl] & \bullet & \\
(s_3, t_1) \arrow[ur] & (s_3, t_2) \arrow[dl] & \bullet & \bullet & \\
\bullet & \bullet & \bullet & &
\end{tikzcd}$$

Figure 1.1 Product of two countable sets is countable

defined by

$$f'(s'_1) = 1, \;\; f'(s'_2) = 2, \ldots$$

is a bijection from S' to $\mathbb{N}$. ■

Theorem 1.2 *The cartesian product of any pair of countable sets is countable.*

Proof: Let S and T be countable sets. Arrange the ordered pairs (s_i, t_j) that make up the elements of $S \times T$ in an infinite rectangular array and then trace a path through the array as depicted in Fig. 1.1, converting it to a sequence that includes every ordered pair. ■

Corollary 1.3 *The rational numbers $\mathbb{Q}$ form a countable set.*

Proof: A rational number is a fraction n/m where m is a natural number (positive integer) and n is an integer having no common factor with m. The rationals are therefore in one-to-one correspondence with a subset of the product set $\mathbb{Z} \times \mathbb{N}$. By Example 1.10 and Theorem 1.2, $\mathbb{Z} \times \mathbb{N}$ is a countable set. Hence the rational numbers $\mathbb{Q}$ are countable. ■

In the set of real numbers ordered by the usual $\leq$, the rationals have the property that for any pair of real numbers x and y such that $x < y$, there exists a rational number q such that $x < q < y$. Any subset, such as the rationals $\mathbb{Q}$, having this property is called a **dense set** in $\mathbb{R}$. The real numbers thus have a *countable* dense subset; yet, as we will now show, the entire set of real numbers turns out to be uncountable.

Uncountable sets

A set is said to be **uncountable** if it is neither finite nor countable; that is, it cannot be set in one-to-one correspondence with any subset of the natural numbers.

Theorem 1.4 *The power set 2^S of any countable set S is uncountable.*

Proof: We use **Cantor's diagonal argument** to demonstrate this theorem. Let the elements of S be arranged in a sequence $S = \{s_1, s_2, \dots\}$. Every subset $U \subseteq S$ defines a unique sequence of 0's and 1's

$$x = \{\epsilon_1, \epsilon_2, \epsilon_3, \dots\}$$

where

$$\epsilon_i = \begin{cases} 0 & \text{if } s_i \notin U, \\ 1 & \text{if } s_i \in U. \end{cases}$$

The sequence x is essentially the characteristic function of the subset U, discussed in Example 1.7. If 2^S is countable then its elements, the subsets of S, can be arranged in sequential form, $U_1, U_2, U_3, \dots$, and so can their set-defining sequences,

$$\begin{aligned} x_1 &= \epsilon_{11},\ \epsilon_{12},\ \epsilon_{13}, \dots \\ x_2 &= \epsilon_{21},\ \epsilon_{22},\ \epsilon_{23}, \dots \\ x_3 &= \epsilon_{31},\ \epsilon_{32},\ \epsilon_{33}, \dots \\ &\text{etc.} \end{aligned}$$

Let x' be the sequence of 0's and 1's defined by

$$x' = \epsilon'_1,\ \epsilon'_2,\ \epsilon'_3, \dots$$

where

$$\epsilon'_i = \begin{cases} 0 & \text{if } \epsilon_{ii} = 1, \\ 1 & \text{if } \epsilon_{ii} = 0. \end{cases}$$

The sequence x' cannot be equal to any of the sequences x_i above since, by definition, it differs from x_i in the ith place, $\epsilon'_i \neq \epsilon_{ii}$. Hence the set of all subsets of S cannot be arranged in a sequence, since their characteristic sequences cannot be so arranged. The power set 2^S cannot, therefore, be countable. ■

Theorem 1.5 *The set of all real numbers $\mathbb{R}$ is uncountable.*

Proof: Each real number in the interval [0, 1] can be expressed as a binary decimal

$$0.\epsilon_1\epsilon_2\epsilon_3 \dots \quad \text{where} \quad \text{each } \epsilon_i = 0 \text{ or } 1 \ (i = 1, 2, 3, \dots).$$

The set [0, 1] is therefore uncountable since it is in one-to-one correspondence with the power set $2^{\mathbb{N}}$. Since this set is a subset of $\mathbb{R}$, the theorem follows at once from Theorem 1.1. ■

Example 1.11 We have seen that the rational numbers form a countable dense subset of the set of real numbers. A set is called *nowhere dense* if it is not dense in any open interval (a, b). Surprisingly, there exists a nowhere dense subset of $\mathbb{R}$ called the **Cantor set**, which is uncountable – the surprise lies in the fact that one would intuitively expect such a set to

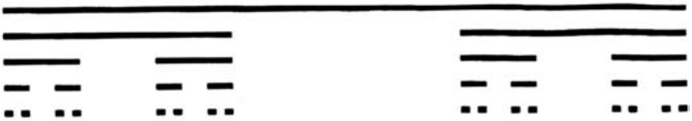

Figure 1.2 The Cantor set (after the four subdivisions)

be even sparser than the rationals. To define the Cantor set, express the real numbers in the interval [0, 1] as ternary decimals, to the base 3,

$$x = 0.\epsilon_1\epsilon_2\epsilon_3 \dots \quad \text{where} \quad \epsilon_i = 0, 1 \text{ or } 2, \quad \forall i.$$

Consider those real numbers whose ternary expansion contains only 0's and 2's. These are clearly in one-to-one correspondence with the real numbers expressed as binary expansions by replacing every 2 with 1.

Geometrically one can picture this set in the following way. From the closed real interval [0, 1] remove the middle third $(1/3, 2/3)$, then remove the middle thirds of the two pieces left over, then of the four pieces left after doing that, and continue this process *ad infinitum*. The resulting set can be visualized in Fig. 1.2.

This set may appear to be little more than a mathematical curiosity, but sets displaying a similar structure to the Cantor set can arise quite naturally in non-linear maps relevant to physics.

The continuum hypothesis and axiom of choice

All infinite subsets of $\mathbb{R}$ described above are either countable or in one-to-one correspondence with the real numbers themselves, of cardinality $2^{\mathbb{N}}$. Cantor conjectured that this was true of all infinite subsets of the real numbers. This famous **continuum hypothesis** proved to be one of the most challenging problems ever postulated in mathematics. In 1938 the famous logician Kurt Gödel (1906–1978) showed that it would never be possible to prove the converse of the continuum hypothesis – that is, no mathematical inconsistency could arise by assuming Cantor's hypothesis to be true. While not proving the continuum hypothesis, this meant that it could never be proved using the time-honoured method of *reductio ad absurdum*. The most definitive result concerning the continuum hypothesis was achieved by Cohen [7], who demonstrated that it was a genuinely independent axiom, neither provable, nor demonstrably false.

In many mathematical arguments, it is assumed that from any family of sets it is always possible to create a set consisting of a representative element from each set. To justify this seemingly obvious procedure it is necessary to postulate the following proposition:

Axiom of choice Given a family of sets $\mathcal{S} = \{S_i \mid i \in I\}$ labelled by an indexing set I, there exists a *choice function* $f : I \to \bigcup \mathcal{S}$ such that $f(i) \in S_i$ for all $i \in I$.

While correct for finite and countably infinite families of sets, the status of this axiom is much less clear for uncountable families. Cohen in fact showed that the axiom of choice was an independent axiom and was independent of the continuum hypothesis. It thus appears that there are a variety of alternative set theories with differing axiom schemes, and the real numbers have different properties in these alternative theories. Even though the real numbers are at the heart of most physical theories, no truly challenging problem for mathematical physics has arisen from these results. While the axiom of choice is certainly useful, its availability is probably not critical in physical applications. When used, it is often invoked in a slightly different form:

Theorem 1.6 (Zorn's lemma) *Let $\{P, \leq\}$ be a partially ordered set (poset) with the property that every totally ordered subset is bounded above. Then P has a maximal element.*

Some words of explanation are in order here. Recall that a subset Q is *totally ordered* if for every pair of elements $x, y \in Q$ either $x \leq y$ or $y \leq x$. A subset Q is said to be **bounded above** if there exists an element $x \in P$ such that $y \leq x$ for all $y \in Q$. A **maximal element** of P is an element z such that there is no $y \neq z$ such that $z \leq y$. The proof that Zorn's lemma is equivalent to the axiom of choice is technical though not difficult; the interested reader is referred to Halmos [4] or Kelley [6].

Problems

Problem 1.11 There is a technical flaw in the proof of Theorem 1.5, since a decimal number ending in an endless sequence of 1's is identified with a decimal number ending with a sequence of 0's, for example,

$$.011011111\ldots = .0111000000\ldots$$

Remove this hitch in the proof.

Problem 1.12 Prove the assertion that the Cantor set is nowhere dense.

Problem 1.13 Prove that the set of all real functions $f : \mathbb{R} \to \mathbb{R}$ has a higher cardinality than that of the real numbers by using a Cantor diagonal argument to show it cannot be put in one-to-one correspondence with $\mathbb{R}$.

Problem 1.14 If $f : [0, 1] \to \mathbb{R}$ is a non-decreasing function such that $f(0) = 0,\ f(1) = 1$, show that the places at which f is not continuous form a countable subset of $[0, 1]$.

1.6 Structures

Physical theories have two aspects, the *static* and the *dynamic*. The former refers to the general background in which the theory is set. For example, special relativity takes place in Minkowski space while quantum mechanics is set in Hilbert space. These mathematical structures are, to use J. A. Wheeler's term, the 'arena' in which a physical system evolves; they are of two basic kinds, algebraic and geometric.

In very broad terms, an *algebraic structure* is a set of binary relations imposed on a set, and 'algebra' consists of those results that can be achieved by formal manipulations using the rules of the given relations. By contrast, a *geometric structure* is postulated as a set of relations on the power set of a set. The objects in a geometric structure can in some sense be 'visualized' as opposed to being formally manipulated. Although mathematicians frequently divide themselves into 'algebraists' and 'geometers', these two kinds of structure interrelate in all kinds of interesting ways, and the distinction is generally difficult to maintain.

Algebraic structures

A **(binary) law of composition** on a set S is a binary map

$$\varphi :\ S \times S \to S.$$

For any pair of elements $a, b \in S$ there thus exists a new element $\varphi(a, b) \in S$ called their **product**. The product is often simply denoted by ab, while at other times symbols such as $a \cdot b,\ a \circ b,\ a + b,\ a \times b,\ a \wedge b,\ [a, b]$, etc. may be used, depending on the context.

Most algebraic structures consist of a set S together with one or more laws of composition defined on S. Sometimes more than one set is involved and the law of composition may take a form such as $\varphi : S \times T \to S$. A typical example is the case of a vector space, where there are two sets involved consisting of vectors and scalars respectively, and the law of composition is *scalar multiplication* (see Chapter 3). In principle we could allow laws of composition that are n-ary maps ($n > 2$), but such laws can always be thought of as families of binary maps. For example, a ternary map $\phi : S^3 \to S$ is equivalent to an indexed family of binary maps $\{\phi_a \mid a \in S\}$ where $\phi_a : S^2 \to S$ is defined by $\phi_a(b, c) = \phi(a, b, c)$.

A law of composition is said to be **commutative** if $ab = ba$. This is *always* assumed to be true for a composition denoted by the symbol $+$; that is, $a + b = b + a$. The law of composition is **associative** if $a(bc) = (ab)c$. This is true, for example, of matrix multiplication or functional composition $f \circ (g \circ h) = (f \circ g) \circ h$, but is not true of vector product $\mathbf{a} \times \mathbf{b}$ in ordinary three-dimensional vector calculus,

$$\mathbf{a} \times (\mathbf{b} \times \mathbf{c}) = (\mathbf{a}.\mathbf{c})\mathbf{b} - (\mathbf{a}.\mathbf{b})\mathbf{c} \neq (\mathbf{a} \times \mathbf{b}) \times \mathbf{c}.$$

Example 1.12 A **semigroup** is a set S with an associative law of composition defined on it. It is said to have an **identity element** if there exists an element $e \in S$ such that

$$ea = ae = a, \quad \forall a \in S.$$

Semigroups are one of the simplest possible examples of an algebraic structure. The theory of semigroups is not particularly rich, and there is little written on their general theory, but particular examples have proved interesting.

(1) The positive integers $\mathbb{N}$ form a commutative semigroup under the operation of addition. If the number 0 is adjoined to this set it becomes a semigroup with identity $e = 0$, denoted $\hat{\mathbb{N}}$.

(2) A map $f : S \to S$ of a set S into itself is frequently called a **discrete dynamical system**. The successive iterates of the function f, namely $F = \{f,\ f^2, \dots,\ f^n = f \circ (f^{n-1}), \dots\}$, form a commutative semigroup with functional iteration as the law of composition. If we include the identity map and set $f^0 = \mathrm{id}_S$, the semigroup is called the **evolution semigroup generated by the function** f, denoted E_f.

The map $\phi : \hat{\mathbb{N}} \to E_f$ defined by $\phi(n) = f^n$ preserves semigroup products,

$$\phi(n + m) = f^{n+m}.$$

Such a product-preserving map between two semigroups is called a **homomorphism**. If the homomorphism is a one-to-one map it is called a **semigroup isomorphism**. Two semigroups that have an isomorphism between them are called **isomorphic**; to all intents and purposes they have the same semigroup structure. The map ϕ defined above need not be an isomorphism. For example on the set $S = \mathbb{R} - \{2\}$, the real numbers excluding the number 2, define the function $f : S \to S$ by

$$f(x) = \frac{2x - 3}{x - 2}.$$

Simple algebra reveals that $f(f(x)) = x$, so that $f^2 = \mathrm{id}_S$. In this case E_f is isomorphic with the residue class of integers modulo 2, defined in Example 1.3.

(3) All of mathematics can be expressed as a semigroup. For example, set theory is made up of finite strings of symbols such as $\{\dots \mid \dots\}$, and, not, $\in, \forall$, etc. and a countable collection of symbols for variables and constants, which may be denoted $x_1, x_2, \dots$ Given two strings σ_1 and σ_2 made up of these symbols, it is possible to construct a new string $\sigma_1\sigma_2$, formed by concatenating the strings. The set of all possible such strings is a semigroup, where 'product' is defined as string concatenation. Of course only some strings are logically meaningful, and are said to be *well-formed*. The rules for a well-formed string are straightforward to list, as are the rules for 'universally valid statements' and the rules of inference. Gödel's famous *incompleteness theorem* states that if we include statements of ordinary arithmetic in the semigroup then there are propositions P such that neither P nor its negation, not P, can be reached from the axioms by any sequence of logically allowable operations. In a sense, the truth of such statements is unknowable. Whether this remarkable theorem has any bearing on theoretical physics has still to be determined.

Geometric structures

In its broadest terms, a geometric structure defines certain classes of subsets of S as in some sense 'acceptable', together with rules concerning their intersections and unions. Alternatively, we can think of a geometric structure $\mathcal{G}$ on a set S as consisting of one or more subsets of 2^S, satisfying certain properties. In this section we briefly discuss two examples: *Euclidean geometry* and *topology*.

Example 1.13 **Euclidean geometry** concerns points (singletons), straight lines, triangles, circles, etc., all of which are subsets of the plane. There is a 'visual' quality of these concepts, even though they are idealizations of the 'physical' concepts of points and lines that must have size or thickness to be visible. The original formulation of plane geometry as set out in Book 1 of *Euclid's Elements* would hardly pass muster by today's criteria as a rigorous axiomatic system. For example, there is considerable confusion between definitions and undefined terms. Historically, however, it is the first systematic approach to an area of mathematics that turns out to be both axiomatic and interesting.

The undefined terms are *point*, *line segment*, *line*, *angle*, *circle* and relations such as *incidence on*, *endpoint*, *length* and *congruence*. Euclid's five postulates are:

1. Every pair of points are on a unique line segment for which they are end points.
2. Every line segment can be extended to a unique line.
3. For every point A and positive number r there exists a unique circle having A as its centre and radius r, such that the line connecting every other point on the circle to A has length r.
4. All right angles are equal to one another.
5. *Playfair's axiom*: given any line ℓ and a point A not on ℓ, there exists a unique line through A that does not intersect ℓ – said to be *parallel* to ℓ.

The undefined terms can be defined as subsets of some basic set known as the Euclidean plane. Points are singletons, line segments and lines are subsets subject to Axioms 1 and 2, while the relation *incidence on* is interpreted as the relation of set-membership $\in$. An angle would be defined as a set $\{A, \ell_1, \ell_2\}$ consisting of a point and two lines on which it is incident. Postulates 1–3 and 5 seem fairly straightforward, but what are we to make of Postulate 4? Such inadequacies were tidied up by Hilbert in 1921.

The least 'obvious' of Euclid's axioms is Postulate 5, which is not manifestly independent of the other axioms. The challenge posed by this axiom was met in the nineteenth century by the mathematicians Bolyai (1802–1860), Lobachevsky (1793–1856), Gauss (1777–1855) and Riemann (1826–1866). With their work arose the concept of *non-Euclidean geometry*, which was eventually to be of crucial importance in Einstein's theory of gravitation known as *general relativity*; see Chapter 18. Although often regarded as a product of pure thought, Euclidean geometry was in fact an attempt to classify logically the geometrical relations in the world around us. It can be regarded as one of the earliest exercises in mathematical physics. Einstein's general theory of relativity carried on this ancient tradition of unifying geometry and physics, a tradition that lives on today in other forms such as gauge theories and string theory.

The discovery of analytic geometry by René Descartes (1596–1650) converted Euclidean geometry into algebraic language. The cartesian method is simply to define the Euclidean plane as $\mathbb{R}^2 = \mathbb{R} \times \mathbb{R}$ with a distance function $d : \mathbb{R}^2 \times \mathbb{R}^2 \to \mathbb{R}$ given by the Pythagorean formula

$$d((x, y), (u, v)) = \sqrt{(x-u)^2 + (y-v)^2}. \tag{1.2}$$

This theorem is central to the analytic version of Euclidean geometry – it underpins the whole Euclidean edifice. The generalization of Euclidean geometry to a space of arbitrary dimensions $\mathbb{R}^n$ is immediate, by setting

$$d(\mathbf{x}, \mathbf{y}) = \sqrt{\sum_{i=1}^{n} (x_i - y_i)^2} \quad \text{where} \quad \mathbf{x} = (x_1,\ x_2, \dots, x_n), \text{ etc.}$$

The ramifications of Pythagoras' theorem have revolutionized twentieth century physics in many ways. For example, Minkowski discovered that Einstein's special theory of relativity could be represented by a four-dimensional *pseudo-Euclidean* geometry where time is interpreted as the fourth dimension and a minus sign is introduced into Pythagoras' law. When gravitation is present, Einstein proposed that Minkowski's geometry must be 'curved', the pseudo-Euclidean structure holding only *locally* at each point. A *complex* vector space having a natural generalization of the Pythagorean structure is known as a *Hilbert space* and forms the basis of quantum mechanics (see Chapters 13 and 14). It is remarkable to think that the two pillars of twentieth century physics, relativity and quantum theory, both have their basis in mathematical structures based on a theorem formulated by an eccentric mathematician over two and a half thousand years ago.

Example 1.14 In Chapter 10 we will meet the concept of a **topology** on a set S, defined as a subset $\mathcal{O}$ of 2^S whose elements (subsets of S) are called *open sets*. To qualify as a topology, the open sets must satisfy the following properties:

1. The empty set and the whole space are open sets, $\emptyset \in \mathcal{O}$ and $S \in \mathcal{O}$.
2. If $U \in \mathcal{O}$ and $V \in \mathcal{O}$ then $U \cap V \in \mathcal{O}$.
3. If $\mathcal{U}$ is any subset of $\mathcal{O}$ then $\bigcup \mathcal{U} \in \mathcal{O}$.

The second axiom says that the intersection of any pair of open sets, and therefore of any finite collection of open sets, is open. The third axiom says that an arbitrary, possibly infinite, union of open sets is open. According to our criterion, a topology is clearly a geometrical structure on S.

The basic view presented here is that the key feature distinguishing an algebraic structure from a geometric structure on a set S is

$$\text{algebraic structure} = \text{a map } S \times S \to S = \text{a subset of } S^3,$$

while

$$\text{geometric structure} = \text{a subset of } 2^S.$$

This may look to be a clean distinction, but it is only intended as a guide, for in reality many structures exhibit both algebraic and geometric aspects. For example, Euclidean geometry as originally expressed in terms of relations between subsets of the plane such as points, lines and circles is the geometric or 'visual' approach. On the other hand, cartesian geometry is the algebraic or analytic approach to plane geometry, in which points are represented as elements of $\mathbb{R}^2$. In the latter approach we have two basic maps: the *difference map* $- : \mathbb{R}^2 \times \mathbb{R}^2 \to \mathbb{R}^2$ defined as $(x, y) - (u, v) = (x - u, y - v)$, and the *distance map* $d : \mathbb{R}^2 \times \mathbb{R}^2 \to \mathbb{R}$ defined by Eq. (1.2). The emphasis on maps places this method much more definitely in the algebraic camp, but the two representations of Euclidean geometry are essentially interchangeable and may indeed be used simultaneously to best understand a problem in plane geometry.

Dynamical systems

The evolution of a system with respect to its algebraic/geometric background invokes what is commonly known as 'laws of physics'. In most cases, particularly when describing a continuous evolution, these laws are expressed in the form of differential equations. Providing they have a well-posed initial value problem, such equations generally give rise to a unique evolution for the system, wherein lies the predictive power of physics. However, exact solutions of differential equations are only available in some very specific cases, and it is frequently necessary to resort to numerical methods designed for digital computers with the time parameter appearing in discrete packets. Discrete time models can also serve as a useful technique for formulating 'toy models' exhibiting features similar to those of a continuum theory, which may be too difficult to prove analytically.

There is an even more fundamental reason for considering discretely evolving systems. We have good reason to believe that on time scales less than the *Planck time*, given by

$$T_{\text{Planck}} = \sqrt{\frac{G\hbar}{c^5}},$$

the continuum fabric of space-time is probably invalid and a quantum theory of gravity becomes operative. It is highly likely that differential equations have little or no physical relevance at or below the Planck scale.

As already discussed in Example 1.12, a *discrete dynamical system* is a set S together with a map $f : S \to S$. The map $f : S \to S$ is called a **discrete dynamical structure** on S. The complexities generated by such a simple structure on a single set S can be enormous. A well-known example is the *logistic map* $f : [0, 1] \to [0, 1]$ defined by

$$f(x) = Cx(1 - x) \quad \text{where} \quad 0 < C \leq 4,$$

and used to model population growth with limited resources or predator–prey systems in ecology. Successive iterates give rise to the phenomena of chaos and *strange attractors* – limiting sets having a Cantor-like structure. The details of this and other maps such as the **Hénon map** [8], $f : \mathbb{R}^2 \to \mathbb{R}^2$ defined by

$$f(x, y) = (y + 1 - ax^2, bx)$$

can be found in several books on non-linear phenomena, such as [9].

Discrete dynamical structures are often described on the set of states on a given set S, where a *state* on S is a function $\phi : S \to \{0, 1\}$. As each state is the characteristic function of some subset of S (see Example 1.7), the set of states on S can be identified with 2^S. A discrete dynamical structure on the set of all states on S is called a **cellular automaton** on S.

Any discrete dynamical system (S, f) induces a cellular automaton $(2^S, f^* : 2^S \to 2^S)$, by setting $f^* : \phi \mapsto \phi \circ f$ for any state $\phi : S \to \{0, 1\}$. This can be pictured in the following way. Every state ϕ on S attaches a 1 or 0 to every point p on S. Assign to p the new value $\phi(f(p))$, which is the value 0 or 1 assigned by the original state ϕ to the *mapped point* $f(p)$. This process is sometimes called a *pullback* – it carries state values 'backwards' rather than forwards. We will frequently meet this idea that a mapping operates on functions, states in this case, in the opposite direction to the mapping.

Not all dynamical structures defined on 2^S, however, can be obtained in the way just described. For example, if S has n elements, then the number of dynamical systems on S is n^n. However, the number of discrete dynamical structures on 2^S is the much larger number $(2^n)^{2^n} = 2^{n2^n}$. Even for small initial sets this number is huge; for example, for $n = 4$ it is $2^{64} \approx 2 \times 10^{19}$, while for slightly larger n it easily surpasses all numbers normally encountered in physics. One of the most intriguing cellular automata is Conway's **game of life**, which exhibits complex behaviour such as the existence of stable structures with the capacity for self-reproducibility, all from three simple rules (see [9, 10]). Graphical versions for personal computers are readily available for experimentation.

1.7 Category theory

Mathematical structures generally fall into 'categories', such as sets, semigroups, groups, vector spaces, topological spaces, differential manifolds, etc. The mathematical theory devoted to this categorizing process can have enormous benefits in the hands of skilled practioners of this abstract art. We will not be making extensive use of category theory, but in this section we provide a flavour of the subject. Those who find the subject too obscure for their taste are urged to move quickly on, as little will be lost in understanding the rest of this book.

A **category** consists of:

(Cat1) A class $\mathcal{O}$ whose elements are called **objects**. Note the use of the word 'class' rather than 'set' here. This is necessary since the objects to be considered are generally themselves sets and the collection of all possible sets with a given type of structure is too vast to be considered as a set without getting into difficulties such as those presented by Russell's paradox discussed in Section 1.1.

(Cat2) For each pair of objects A, B of $\mathcal{O}$ there is a set $\mathrm{Mor}(A, B)$ whose elements are called **morphisms** from A to B, usually denoted $A \xrightarrow{\phi} B$.

(Cat3) For any pair of morphisms $A \xrightarrow{\phi} B$, $B \xrightarrow{\psi} C$ there is a morphism $A \xrightarrow{\psi\circ\phi} C$, called the **composition** of ϕ and ψ such that

1. Composition is associative: for any three morphisms $A \xrightarrow{\phi} B, B \xrightarrow{\psi} C, C \xrightarrow{\rho} D$,

$$(\rho \circ \psi) \circ \phi = \rho \circ (\psi \circ \phi).$$

2. For each object A there is a morphism $A \xrightarrow{\iota_A} A$ called the **identity morphism** on A, such that for any morphism $A \xrightarrow{\phi} B$ we have

$$\phi \circ \iota_A = \phi,$$

and for any morphism $C \xrightarrow{\psi} A$ we have

$$\iota_A \circ \psi = \psi.$$

Example 1.15 The simplest example of a category is the **category of sets**, in which the objects are all possible sets, while morphisms are mappings from a set A to a set B. In this case the set $\mathrm{Mor}(A, B)$ consists of all possible mappings from A to B. Composition of morphisms is simply composition of mappings, while the identity morphism on an object A is the identity map id_A on A. Properties (Cat1) and (Cat2) were shown in Section 1.4.

Exercise: Show that the class of all semigroups, Example 1.12, forms a category, where morphisms are defined as semigroup homomorphisms.

The following are some other important examples of categories of structures to appear in later chapters:

Objects	*Morphisms*	*Refer to*
Groups	Homomorphisms	Chapter 2
Vector spaces	Linear maps	Chapter 3
Algebras	Algebra homomorphisms	Chapter 6
Topological spaces	Continuous maps	Chapter 10
Differential manifolds	Differentiable maps	Chapter 15
Lie groups	Lie group homomorphisms	Chapter 19

Two important types of morphisms are defined as follows. A morphism $A \xrightarrow{\varphi} B$ is called a **monomorphism** if for any object X and morphisms $X \xrightarrow{\alpha} A$ and $X \xrightarrow{\alpha'} A$ we have that

$$\varphi \circ \alpha = \varphi \circ \alpha' \implies \alpha = \alpha'.$$

The morphism φ is called an **epimorphism** if for any object X and morphisms $B \xrightarrow{\beta} Y$ and $B \xrightarrow{\beta'} Y$

$$\beta \circ \varphi = \beta' \circ \varphi \implies \beta = \beta'.$$

These requirements are often depicted in the form of **commutative diagrams**. For example, φ is a monomorphism if the morphism α is uniquely defined by the diagram shown in Fig. 1.3. The word 'commutative' here means that chasing arrows results in composition of morphisms, $\psi = (\varphi \circ \alpha)$.

$$\begin{tikzcd}
& X \arrow[dl, "\alpha"'] \arrow[dr, "\psi"] & \\
A \arrow[rr, "\varphi"'] & & B
\end{tikzcd}$$

Figure 1.3 Monomorphism φ

$$\begin{tikzcd}
A \arrow[rr, "\varphi"] \arrow[dr, "\psi"'] & & B \arrow[dl, "\beta"] \\
& Y &
\end{tikzcd}$$

Figure 1.4 Epimorphism φ

On the other hand, φ is an epimorphism if the morphism β is uniquely defined in the commutative diagram shown on Fig. 1.4.

In the case of the category of sets a morphism $A \xrightarrow{\varphi} B$ is a monomorphism if and only if it is a one-to-one mapping.

Proof: 1. If $\varphi : A \to B$ is one-to-one then for any pair of maps $\alpha : X \to A$ and $\alpha' : X \to A$,

$$\varphi(\alpha(x)) = \varphi(\alpha'(x)) \Longrightarrow \alpha(x) = \alpha'(x)$$

for all $x \in X$. This is simply another way of stating the monomorphism property $\varphi \circ \alpha = \varphi \circ \alpha' \Longrightarrow \alpha = \alpha'$.

2. Conversely, suppose φ is a monomorphism. Since X is an arbitrary set, in the definition of the monomorphism property, we may choose it to be a singleton $X = \{x\}$. For any pair of points $a, a' \in A$ define the maps $\alpha, \alpha' : X \to A$ by setting $\alpha(x) = a$ and $\alpha'(x) = a'$. Then

$$\begin{aligned}
\varphi(a) = \varphi(a') &\Longrightarrow \varphi \circ \alpha(x) = \varphi \circ \alpha'(x) \\
&\Longrightarrow \varphi \circ \alpha = \varphi \circ \alpha' \\
&\Longrightarrow \alpha = \alpha' \\
&\Longrightarrow a = \alpha(x) = \alpha'(x) = a'.
\end{aligned}$$

Hence φ is one-to-one. ■

It is left as a problem to show that in the category of sets a morphism is an epimorphism if and only if it is surjective. A morphism $A \xrightarrow{\varphi} B$ is called an **isomorphism** if there exists a morphism $B \xrightarrow{\varphi'} A$ such that

$$\varphi' \circ \varphi = \iota_A \quad \text{and} \quad \varphi \circ \varphi' = \iota_B.$$

In the category of sets a mapping is an isomorphism if and only if it is bijective; that is, it is both an epimorphism and a monomorphism. There can, however, be a trap for the unwary here. While every isomorphism is readily shown to be both a monomorphism and an epimorphism, the converse is not always true. A classic case is the category of Hausdorff topological spaces in which there exist continuous maps that are epimorphisms and monomorphisms but are not invertible. The interested reader is referred to [11] for further development of this subject.

Problems

Problem 1.15 Show that in the category of sets a morphism is an epimorphism if and only if it is onto (surjective).

Problem 1.16 Show that every isomorphism is both a monomorphism and an epimorphism.

References

[1] T. Apostol. *Mathematical Analysis*. Reading, Mass., Addison-Wesley, 1957.

[2] K. Devlin. *The Joy of Sets*. New York, Springer-Verlag, 1979.

[3] N. B. Haaser and J. A. Sullivan. *Real Analysis*. New York, Van Nostrand Reinhold Company, 1971.

[4] P. R. Halmos. *Naive Set Theory*. New York, Springer-Verlag, 1960.

[5] E. Wigner. The unreasonable effectiveness of mathematics in the natural sciences. *Communications in Pure and Applied Mathematics*, **13**:1–14, 1960.

[6] J. Kelley. *General Topology*. New York, D. Van Nostrand Company, 1955.

[7] P. J. Cohen. *Set Theory and the Continuum Hypothesis*. New York, W. A. Benjamin, 1966.

[8] M. Hénon. A two-dimensional map with a strange attractor. *Communications in Mathematical Physics*, **50**:69–77, 1976.

[9] M. Schroeder. *Fractals, Chaos, Power Laws*. New York, W. H. Freeman and Company, 1991.

[10] W. Poundstone. *The Recursive Universe*. Oxford, Oxford University Press, 1987.

[11] R. Geroch. *Mathematical Physics*. Chicago, The University of Chicago Press, 1985.