1

Mathematical background

This chapter reviews some of the basic mathematical ideas and notations that are used throughout the book. Section 1.1 on set theory and Section 1.2 on functions are rather concise; readers unfamiliar with this type of material are advised to consult a more detailed text on mathematical analysis. Measures and mass distributions play an important part in the theory of fractals and a treatment adequate for our needs is given in Section 1.3. By asking the reader to take on trust the existence of certain measures, we can avoid many of the technical difficulties usually associated with measure theory. Some notes on probability theory are given in Section 1.4; this is needed in Chapters 15 and 16.

1.1 Basic set theory

In this section, we recall some basic notions from set theory and point set topology.

We generally work in *n-dimensional Euclidean space*, $\mathbb{R}^n$, where $\mathbb{R}^1 = \mathbb{R}$ is just the set of real numbers or the 'real line', and $\mathbb{R}^2$ is the (Euclidean) plane. Points in $\mathbb{R}^n$ will generally be denoted by lower case letters x, y, and so on, and we will occasionally use the coordinate form $x = (x_1, \ldots, x_n), y = (y_1, \ldots, y_n)$. Addition and scalar multiplication are defined in the usual manner, so that $x + y = (x_1 + y_1, \ldots, x_n + y_n)$ and $\lambda x = (\lambda x_1, \ldots, \lambda x_n)$, where λ is a real scalar. We use the usual *Euclidean distance* or *metric* on $\mathbb{R}^n$ so if x and y are points of $\mathbb{R}^n$, the distance between them is $|x - y| = \left(\sum_{i=1}^{n} |x_i - y_i|^2\right)^{1/2}$. In particular, the triangle inequality $|x + y| \leqslant |x| + |y|$, the reverse triangle inequality $\big||x| - |y|\big| \leqslant |x - y|$ and the metric triangle inequality $|x - y| \leqslant |x - z| + |z - y|$ hold for all $x, y, z \in \mathbb{R}^n$.

Sets, which will generally be subsets of $\mathbb{R}^n$, are denoted by capital letters E, F, U, and so on. In the usual way, $x \in E$ means that the point x belongs to the set E, and $E \subset F$ means that E is a subset of the set F. We write $\{x : \text{condition}\}$ for the set

Fractal Geometry: Mathematical Foundations and Applications, Third Edition. Kenneth Falconer.

Companion website: http://www.wiley.com/go/fractal

of x for which 'condition' is true. Certain frequently occurring sets have a special notation. The empty set, which contains no elements, is written as $\emptyset$. The integers are denoted by $\mathbb{Z}$, and the rational numbers by $\mathbb{Q}$. We use a superscript $^+$ to denote the positive elements of a set; thus, $\mathbb{R}^+$ are the positive real numbers, and $\mathbb{Z}^+$ are the positive integers. Sometimes we refer to the complex numbers $\mathbb{C}$, which for many purposes may be identified with the plane $\mathbb{R}^2$, with $x_1 + \mathrm{i}x_2$ corresponding to the point (x_1, x_2).

The *closed ball* of centre x and radius r is defined by $B(x, r) = \{y : |y - x| \leqslant r\}$. Similarly, the *open ball* is $B^{\mathrm{o}}(x, r) = \{y : |y - x| < r\}$. Thus, the closed ball contains its bounding sphere, but the open ball does not. Of course, in $\mathbb{R}^2$, a ball is a disc and in $\mathbb{R}^1$ a ball is just an interval. If $a < b$, we write $[a, b]$ for the *closed interval* $\{x : a \leqslant x \leqslant b\}$ and (a, b) for the *open interval* $\{x : a < x < b\}$. Similarly, $[a, b)$ denotes the half-open interval $\{x : a \leqslant x < b\}$, and so on.

The *coordinate cube* of side $2r$ and centre $x = (x_1, \ldots, x_n)$ is the set $\{y = (y_1, \ldots, y_n) : |y_i - x_i| \leqslant r \text{ for all } i = 1, \ldots, n\}$. (A cube in $\mathbb{R}^2$ is just a square and in $\mathbb{R}^1$ is an interval.)

From time to time we refer to the *δ-neighbourhood* or *δ-parallel body*, A_δ, of a set A, that is, the set of points within distance δ of A; thus, $A_\delta = \{x : |x - y| \leqslant \delta \text{ for some } y \text{ in } A\}$ (see Figure 1.1).

We write $A \cup B$ for the *union* of the sets A and B, that is, the set of points belonging to either A or B, or both. Similarly, we write $A \cap B$ for their *intersection*, the points in both A and B. More generally, $\bigcup_\alpha A_\alpha$ denotes the union of an arbitrary collection of sets $\{A_\alpha\}$, that is, those points in at least one of the sets A_α, and $\bigcap_\alpha A_\alpha$ denotes their intersection, consisting of the set of points common to all of the A_α. A collection of sets is *disjoint* if the intersection of any pair is the empty set. The *difference* $A \backslash B$ of A and B consists of the points in A but not B. The set $\mathbb{R}^n \backslash A$ is termed the *complement* of A.

The set of all ordered pairs $\{(a, b) : a \in A \text{ and } b \in B\}$ is called the (*Cartesian*) *product* of A and B and is denoted by $A \times B$. If $A \subset \mathbb{R}^n$ and $B \subset \mathbb{R}^m$, then $A \times B \subset \mathbb{R}^{n+m}$.

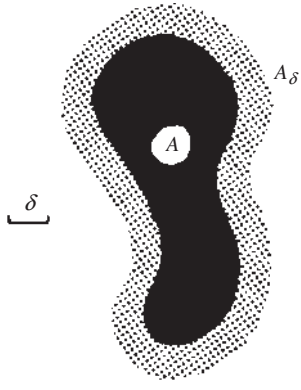

Figure 1.1 A set A and its δ-neighbourhood A_δ.

If A and B are subsets of $\mathbb{R}^n$ and λ is a real number, we define the *vector sum* of the sets as $A + B = \{x + y : x \in A \text{ and } y \in B\}$ and we define the *scalar multiple* $\lambda A = \{\lambda x : x \in A\}$.

An infinite set A is *countable* if its elements can be listed in the form $x_1, x_2, \ldots$ with every element of A appearing at a specific place in the list; otherwise, the set is *uncountable*. The sets $\mathbb{Z}$ and $\mathbb{Q}$ are countable but $\mathbb{R}$ is uncountable. Note that a countable union of countable sets is countable.

If A is any non-empty set of real numbers, then its *supremum* $\sup A$ is the least number m such that $x \leqslant m$ for every x in A or is ∞ if no such number exists. Similarly, the *infimum* $\inf A$ is the greatest number m such that $m \leqslant x$ for all x in A or is $-\infty$. Intuitively, the supremum and infimum are thought of as the maximum and minimum of the set, although it is important to realise that $\sup A$ and $\inf A$ need not be members of the set A itself. For example, $\sup(0, 1) = 1$, but $1 \notin (0, 1)$. We write $\sup_{x \in B}(\ )$ for the supremum of the quantity in brackets, which may depend on x, as x ranges over the set B.

We define the *diameter* $|A|$ of a non-empty subset of $\mathbb{R}^n$ as the greatest distance apart of pairs of points in A. Thus, $|A| = \sup\{|x - y| : x, y \in A\}$. In $\mathbb{R}^n$, a ball of radius r has diameter $2r$, and a cube of side length δ has diameter $\delta\sqrt{n}$. A set A is *bounded* if it has finite diameter or, equivalently, if A is contained in some (sufficiently large) ball.

Convergence of sequences is defined in the usual way. A sequence $\{x_k\}$ in $\mathbb{R}^n$ *converges* to a point x of $\mathbb{R}^n$ as $k \to \infty$ if, given $\varepsilon > 0$, there exists a number K such that $|x_k - x| < \varepsilon$ whenever $k > K$, that is, if $|x_k - x|$ converges to 0. The number x is called the *limit* of the sequence, and we write $x_k \to x$ or $\lim_{k \to \infty} x_k = x$.

The ideas of 'open' and 'closed' that have been mentioned in connection with balls apply to much more general sets. Intuitively, a set is closed if it contains its boundary and open if it contains none of its boundary points. More precisely, a subset A of $\mathbb{R}^n$ is *open* if, for all points x in A, there is some ball $B(x, r)$, centred at x and is of positive radius that is contained in A. A set is *closed* if whenever $\{x_k\}$ is a sequence of points of A converging to a point x of $\mathbb{R}^n$, then x is in A (see Figure 1.2). The empty set $\emptyset$ and $\mathbb{R}^n$ are regarded as both open and closed.

It may be shown that a set is open if and only if its complement is closed. The union of any collection of open sets is open, as is the intersection of any finite

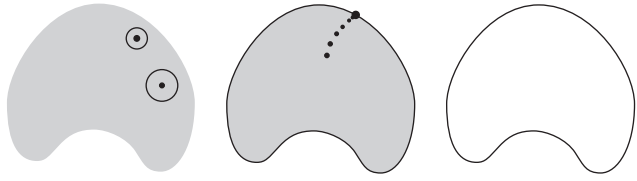

Figure 1.2 (a) An open set – there is a ball contained in the set centred at each point of the set. (b) A closed set – the limit of any convergent sequence of points from the set lies in the set. (c) The boundary of the set in (a) or (b).

number of open sets. The intersection of any collection of closed sets is closed, as is the union of any finite number of closed sets (see Exercise 1.6).

A set A is called a *neighbourhood* of a point x if there is some (small) ball $B(x, r)$ centred at x and contained in A.

The intersection of all the closed sets containing a set A is called the *closure* of A, written $\overline{A}$. The union of all the open sets contained in A is the *interior* $\text{int}A$ of A. The closure of A is thought of as the smallest closed set containing A, and the interior as the largest open set contained in A. The *boundary* ∂A of A is given by $\partial A = \overline{A} \backslash \text{int}A$, thus $x \in \partial A$ if and only if the ball $B(x, r)$ intersects both A and its complement for all $r > 0$.

A set B is a *dense* in A if $A \subset \overline{B}$, that is, if there are points of B arbitrarily close to each point of A.

A set A is *compact* if any collection of open sets that covers A (i.e. with union containing A) has a finite subcollection which also covers A. Technically, compactness is an extremely useful property that enables infinite sets of conditions to be reduced to finitely many. However, as far as most of this book is concerned, it is enough to take the definition of a compact subset of $\mathbb{R}^n$ as one that is both closed and bounded.

The intersection of any collection of compact sets is compact. It may be shown that if $A_1 \supset A_2 \supset \cdots$ is a decreasing sequence of compact sets, then the intersection $\bigcap_{i=1}^{\infty} A_i$ is non-empty (see Exercise 1.7). Moreover, if $\bigcap_{i=1}^{\infty} A_i$ is contained in V for some open set V, then the finite intersection $\bigcap_{i=1}^{k} A_i$ is contained in V for some k.

A subset A of $\mathbb{R}^n$ is *connected* if there do not exist open sets U and V such that $U \cup V$ contains A with $A \cap U$ and $A \cap V$ disjoint and non-empty. Intuitively, we think of a set A as connected if it consists of just one 'piece'. The largest connected subset of A containing a point x is called the *connected component* of x. The set A is *totally disconnected* if the connected component of each point consists of just that point. This will certainly be so if for every pair of points x and y in A we can find disjoint open sets U and V such that $x \in U, y \in V$ and $A \subset U \cup V$.

There is one further class of set that must be mentioned, although its precise definition is indirect and should not concern the reader unduly. The class of *Borel sets* is the smallest collection of subsets of $\mathbb{R}^n$ with the following properties:

1. Every open set and every closed set is a Borel set.
2. The union of every finite or countable collection of Borel sets is a Borel set, and the intersection of every finite or countable collection of Borel sets is a Borel set.

Throughout this book, virtually all of the subsets of $\mathbb{R}^n$ that will be of any interest to us will be Borel sets. Any set that can be constructed using a sequence of countable unions or intersections starting with the open sets or closed sets will certainly be Borel. The reader will not go far wrong with the material of the sort described in this book by assuming that all the sets encountered are Borel sets.

1.2 Functions and limits

Let X and Y be any sets. A *mapping, function* or *transformation* f from X to Y is a rule or formula that associates a point $f(x)$ of Y with each point x of X. We write $f : X \to Y$ to denote this situation; X is called the *domain* of f and Y is called the *codomain*. If A is any subset of X, we write $f(A)$ for the *image* of A, given by $\{f(x) : x \in A\}$. If B is a subset of Y, we write $f^{-1}(B)$ for the *inverse image* or *pre-image* of B, that is, the set $\{x \in X : f(x) \in B\}$; note that in this context, the inverse image of a single point can contain many points.

A function $f : X \to Y$ is called an *injection* or a *one-to-one* function if $f(x) \neq f(y)$ whenever $x \neq y$, that is, different elements of X are mapped to different elements of Y. The function is called a *surjection* or an *onto* function if, for every y in Y, there is an element x in X with $f(x) = y$, that is, every element of Y is the image of some point in X. A function that is both an injection and a surjection is called a *bijection* or *one-to-one correspondence* between X and Y. If $f : X \to Y$ is a bijection, then we may define the *inverse function* $f^{-1} : Y \to X$ by taking $f^{-1}(y)$ as the unique element of X such that $f(x) = y$. In this situation, $f^{-1}(f(x)) = x$ for all x in X and $f(f^{-1}(y)) = y$ for all y in Y.

The *composition* of the functions $f : X \to Y$ and $g : Y \to Z$ is the function $g \circ f : X \to Z$ given by $(g \circ f)(x) = g(f(x))$. This definition extends to the composition of any finite number of functions in the obvious way.

Certain functions from $\mathbb{R}^n$ to $\mathbb{R}^n$ have a particular geometric significance; often, in this context, they are referred to as *transformations* and are denoted by capital letters. Their effects are shown in Figure 1.3. The transformation $S : \mathbb{R}^n \to \mathbb{R}^n$ is called a *congruence* or *isometry* if it preserves distances, that is if $|S(x) - S(y)| = |x - y|$ for x, y in $\mathbb{R}^n$. Congruences also preserve angles and transform sets into geometrically congruent ones. Special cases include *translations*, which are of the form $S(x) = x + a$ and have the effect of shifting points parallel to the vector a, *rotations* which have a centre a such that $|S(x) - a| = |x - a|$ for all x (for convenience, we also regard the identity transformation given by $I(x) = x$ as a rotation) and *reflections*, which maps points to their mirror images in some $(n-1)$-dimensional plane. A congruence that may be achieved by a combination of a rotation and a translation, that is, does not involve reflection, is called a *rigid motion* or *direct congruence*. A transformation $S : \mathbb{R}^n \to \mathbb{R}^n$ is a *similarity* of *ratio* or *scale* $c > 0$ if $|S(x) - S(y)| = c|x - y|$ for all x, y in $\mathbb{R}^n$. A similarity transforms sets into geometrically similar ones with all lengths multiplied by the factor c.

A transformation $T : \mathbb{R}^n \to \mathbb{R}^n$ is *linear* if $T(x + y) = T(x) + T(y)$ and $T(\lambda x) = \lambda T(x)$ for all $x, y \in \mathbb{R}^n$ and $\lambda \in \mathbb{R}$; linear transformations may be represented by matrices in the usual way. Such a linear transformation is *non-singular* if $T(x) = 0$ if and only if $x = 0$. If $S : \mathbb{R}^n \to \mathbb{R}^n$ is of the form $S(x) = T(x) + a$, where T is a non-singular linear transformation and a is a vector in $\mathbb{R}^n$, then S is called an *affine transformation* or an *affinity*. An affinity may be thought of as a shearing transformation; its contracting or expanding effect need not be the same in every direction.

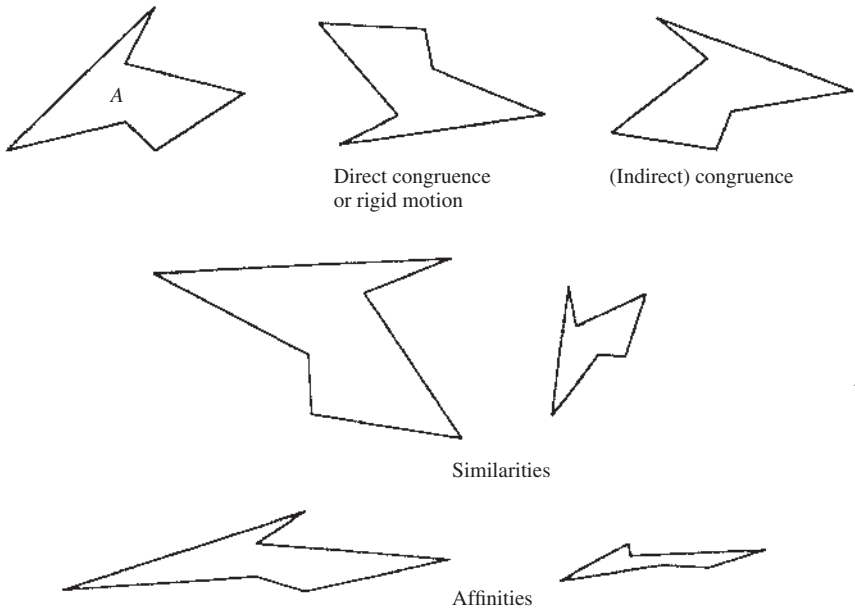

Figure 1.3 The effect of various transformations on a set A.

However, if T is orthonormal, then S is a congruence, and if T is a scalar multiple of an orthonormal transformation, then T is a similarity.

It is worth pointing out that such classes of transformation form groups under composition of mappings. For example, the composition of two translations is a translation, the identity transformation is trivially a translation, and the inverse of a translation is a translation. Finally, the associative law $S \circ (T \circ U) = (S \circ T) \circ U$ holds for all translations S, T, U. Similar group properties hold for the congruences, the rigid motions, the similarities and the affinities.

A function $f : X \to Y$ is called a *Hölder function of exponent* α if

$$|f(x) - f(y)| \leqslant c|x - y|^{\alpha} \quad (x, y \in X)$$

for some constant $c \geqslant 0$. The function f is called *Lipschitz* if α may be taken to be equal to 1, that is if

$$|f(x) - f(y)| \leqslant c|x - y| \quad (x, y \in X)$$

and *bi-Lipschitz* if

$$c_1|x - y| \leqslant |f(x) - f(y)| \leqslant c_2|x - y| \quad (x, y \in X)$$

for $0 < c_1 \leqslant c_2 < \infty$, in which case both f and $f^{-1} : f(X) \to X$ are Lipschitz functions. Lipschitz and Hölder functions play an important role in fractal geometry.

We next remind readers of the basic ideas of limits and continuity of functions. Let X and Y be subsets of $\mathbb{R}^n$ and $\mathbb{R}^m$, respectively, let $f : X \to Y$ be a function, and let a be a point of $\overline{X}$. We say that $f(x)$ has *limit* y (or *tends to* y, or *converges to* y) *as* x *tends to* a, if, given $\varepsilon > 0$, there exists $\delta > 0$ such that $|f(x) - y| < \varepsilon$ for all $x \in X$ with $|x - a| < \delta$. We denote this by writing $f(x) \to y$ as $x \to a$ or by $\lim_{x \to a} f(x) = y$. For a function $f : X \to \mathbb{R}$, we say that $f(x)$ *tends to infinity* (written $f(x) \to \infty$) *as* $x \to a$ if, given M, there exists $\delta > 0$ such that $f(x) > M$ whenever $|x - a| < \delta$. The definition of $f(x) \to -\infty$ is similar.

Suppose, now, that $f : \mathbb{R}^+ \to \mathbb{R}$. We shall frequently be interested in the values of such functions for small positive values of x. Note that if $f(x)$ is increasing as x decreases, then $\lim_{x \to 0} f(x)$ exists either as a finite limit or as ∞, and if $f(x)$ is decreasing as x decreases, then $\lim_{x \to 0} f(x)$ exists and is finite or $-\infty$. Of course, $f(x)$ can fluctuate wildly for small x and $\lim_{x \to 0} f(x)$ need not exist at all. We use lower and upper limits to describe such fluctuations. We define the *lower limit* as

$$\underline{\lim_{x \to 0}} f(x) \equiv \lim_{r \to 0} (\inf \{f(x) : 0 < x < r\}).$$

As $\inf \{f(x) : 0 < x < r\}$ is either $-\infty$ for all positive r or else increases as r decreases, $\underline{\lim}_{x \to 0} f(x)$ always exists. Similarly, the *upper limit* is defined as

$$\overline{\lim_{x \to 0}} f(x) \equiv \lim_{r \to 0} (\sup \{f(x) : 0 < x < r\}).$$

The lower and upper limits exist (as real numbers or $-\infty$ or ∞) for every function f and are indicative of the variation of f for x close to 0 (see Figure 1.4). Clearly, $\underline{\lim}_{x \to 0} f(x) \leqslant \overline{\lim}_{x \to 0} f(x)$; if the lower and upper limits are equal, then $\lim_{x \to 0} f(x)$ exists and equals this common value. Note that if $f(x) \leqslant g(x)$ for $x > 0$,

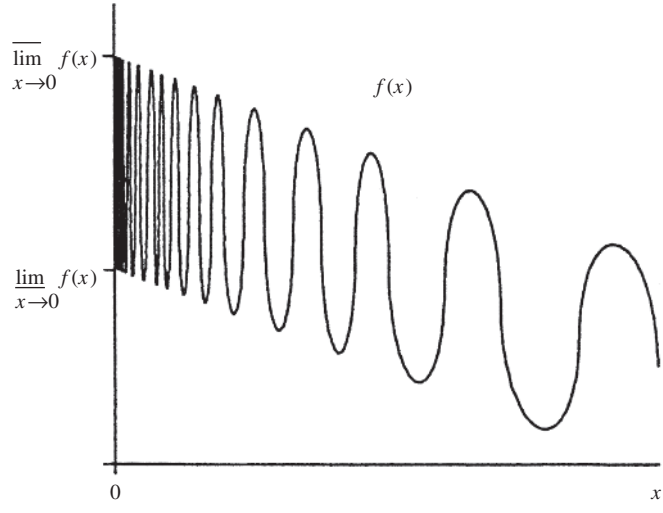

Figure 1.4 The upper and lower limits of a function.

then $\underline{\lim}_{x\to 0} f(x) \leqslant \underline{\lim}_{x\to 0} g(x)$ and $\overline{\lim}_{x\to 0} f(x) \leqslant \overline{\lim}_{x\to 0} g(x)$. In the same way, it is possible to define lower and upper limits as $x \to a$ for functions $f : X \to \mathbb{R}$ where X is a subset of $\mathbb{R}^n$ with a in $\overline{X}$.

We sometimes need to compare two functions $f, g : \mathbb{R}^+ \to \mathbb{R}$ for small values. We write $f(x) \sim g(x)$ to mean that $f(x)/g(x) \to 1$ as $x \to 0$. We will often have that $f(x) \sim x^s$; in other words, f obeys an approximate power law of exponent s when x is small. We use the notation $f(x) \simeq g(x)$ more loosely, to mean that $f(x)$ and $g(x)$ are approximately equal in some sense, to be specified in the particular circumstances.

Recall that a function $f : X \to Y$ is *continuous* at a point a of X if $f(x) \to f(a)$ as $x \to a$ and is *continuous on* X if it is continuous at all points of X. In particular, Lipschitz and Hölder mappings are continuous. If $f : X \to Y$ is a continuous bijection with continuous inverse $f^{-1} : Y \to X$, then f is called a *homeomorphism*, and X and Y are termed *homeomorphic* sets. Congruences, similarities and affine transformations on $\mathbb{R}^n$ are examples of homeomorphisms.

The function $f : \mathbb{R} \to \mathbb{R}$ is *differentiable* at x with the number $f'(x)$ as *derivative* if

$$\lim_{h\to 0} \frac{f(x+h) - f(x)}{h} = f'(x).$$

A function f is termed *continuously differentiable* if $f'(x)$ is continuous in x. Very significant is the *mean value theorem* that states that, given $x < y$ and a real-valued function f that is differentiable over an interval containing x and y, there exists w with $x < w < y$ such that

$$\frac{f(y) - f(x)}{y - x} = f'(w)$$

(intuitively, any chord of the graph of f is parallel to the slope of f at some intermediate point). A consequence of the mean value theorem is that if $|f'(x)|$ is bounded over an interval, then f is Lipschitz over that interval.

More generally, if $f : \mathbb{R}^n \to \mathbb{R}^n$, we say that f is *differentiable* at x and has *derivative* given by the linear mapping $f'(x) : \mathbb{R}^n \to \mathbb{R}^n$ if

$$\lim_{h\to 0} \frac{|f(x+h) - f(x) - f'(x)h|}{|h|} = 0.$$

Occasionally, we shall be interested in the convergence of a sequence of functions $f_k : X \to Y$ where X and Y are subsets of Euclidean spaces. We say that functions f_k converge *pointwise* to a function $f : X \to Y$ if $f_k(x) \to f(x)$ as $k \to \infty$ for each x in X. We say that the convergence is *uniform* if $\sup_{x\in X} |f_k(x) - f(x)| \to 0$ as $k \to \infty$. Uniform convergence is a rather stronger property than pointwise convergence; the rate at which the limit is approached is uniform across X. If the functions f_k are continuous and converge uniformly to f, then f is continuous.

Finally, we remark that logarithms will always be to base e. Recall that, for $a, b > 0$, we have that $\log ab = \log a + \log b$ and that $\log a^c = c \log a$ for real numbers c. The identity $a^c = b^{c \log a / \log b}$ will often be used. The logarithm is the inverse of the exponential function, so that $e^{\log x} = x$, for $x > 0$, and $\log e^y = y$ for $y \in \mathbb{R}$.

1.3 Measures and mass distributions

Anyone studying the mathematics of fractals will not get far before encountering measures in some form or other. Many people are put off by the seemingly technical nature of measure theory – often unnecessarily so, because for most fractal applications only a few basic ideas are needed. Moreover, these ideas are often already familiar in the guise of the mass or charge distributions encountered in basic physics.

We need only be concerned with measures on subsets of $\mathbb{R}^n$. Basically, a measure is just a way of ascribing a numerical 'size' to sets, such that if a set is decomposed into a finite or countable number of pieces in a reasonable way, then the size of the whole is the sum of the sizes of the pieces.

We call μ a *measure* on $\mathbb{R}^n$ if μ assigns a non-negative number, possibly ∞, to each subset of $\mathbb{R}^n$ such that

(a) $\mu(\emptyset) = 0$; (1.1)

(b) $\mu(A) \leqslant \mu(B)$ if $A \subset B$; (1.2)

(c) if $A_1, A_2, \ldots$ is a countable (or finite) sequence of sets, then

$$\mu\left(\bigcup_{i=1}^{\infty} A_i\right) \leqslant \sum_{i=1}^{\infty} \mu(A_i) \tag{1.3}$$

with equality in (1.3), that is

$$\mu\left(\bigcup_{i=1}^{\infty} A_i\right) = \sum_{i=1}^{\infty} \mu(A_i), \tag{1.4}$$

if the A_i are disjoint Borel sets.

We call $\mu(A)$ the *measure* of the set A and think of $\mu(A)$ as the size of A measured in some way. Condition (a) says that the empty set has zero measure, condition (b) says 'the larger the set, the larger the measure' and condition (c) says that if a set is a union of a countable number of pieces (which may overlap), then the sum of the measure of the pieces is at least equal to the measure of the whole. If a set is decomposed into a countable number of disjoint Borel sets, then the total measure of the pieces equals the measure of the whole.

Technical note. For the measures that we shall encounter, (1.4) generally holds for a much wider class of sets than just the Borel sets, in particular for all images of Borel sets under continuous functions. However, for reasons that need not concern us here, we cannot in general require that (1.4) holds for every countable collection of disjoint sets A_i. The reader who is familiar with measure theory will realise that our definition of a measure on $\mathbb{R}^n$ is the definition of what would normally be termed 'an outer measure on $\mathbb{R}^n$ for which the Borel sets are measurable'. However, to save frequent referral to 'measurable sets', it is convenient to have $\mu(A)$ defined for every

set A, and because we are usually interested in measures of Borel sets, it is enough to have (1.4) holding for Borel sets rather than for a larger class. If μ is defined and satisfies (1.1)–(1.4) for the Borel sets, the definition of μ may be extended to an outer measure on all sets in such a way that (1.1)–(1.3) hold, so our definition is consistent with the usual one.

If $A \supset B$, then A may be expressed as a disjoint union $A = B \cup (A \backslash B)$, so it is immediate from (1.4) that, if A and B are Borel sets with $\mu(B)$ finite,

$$\mu(A \backslash B) = \mu(A) - \mu(B). \tag{1.5}$$

Similarly, if $A_1 \subset A_2 \subset \cdots$ is an increasing sequence of Borel sets, then

$$\mu\left(\bigcup_{i=1}^{\infty} A_i\right) = \lim_{i \to \infty} \mu(A_i). \tag{1.6}$$

To see this, note that $\bigcup_{i=1}^{\infty} A_i = A_1 \cup (A_2 \backslash A_1) \cup (A_3 \backslash A_2) \cup \ldots$, with this union disjoint, so that

$$\begin{aligned}
\mu\left(\bigcup_{i=1}^{\infty} A_i\right) &= \mu(A_1) + \sum_{i=1}^{\infty} (\mu(A_{i+1}) - \mu(A_i)) \\
&= \mu(A_1) + \lim_{k \to \infty} \sum_{i=1}^{k} (\mu(A_{i+1}) - \mu(A_i)) \\
&= \lim_{k \to \infty} \mu(A_k).
\end{aligned}$$

A simple extension of this is that if, for $\delta > 0, A_\delta$ are Borel sets that are increasing as δ decreases, that is, $A_{\delta'} \subset A_\delta$ for $0 < \delta < \delta'$, then

$$\mu\left(\bigcup_{\delta > 0} A_\delta\right) = \lim_{\delta \to 0} \mu(A_\delta). \tag{1.7}$$

We think of the support of a measure as the set on which the measure is concentrated. Formally, the *support* of μ, written spt μ, is the smallest closed set X such that $\mu(\mathbb{R}^n \backslash X) = 0$. Thus, x is in the support if and only if $\mu(B(x, r)) > 0$ for all positive radii r. We say that μ is a measure *on* a set A if A contains the support of μ.

A measure on a bounded subset of $\mathbb{R}^n$ for which $0 < \mu(\mathbb{R}^n) < \infty$ will be called a *mass distribution*, and we think of $\mu(A)$ as the mass of the set A. We often think of this intuitively: we take a finite mass and spread it in some way across a set X to get a mass distribution on X; the conditions for a measure will then be satisfied.

We give some examples of measures and mass distributions. In general, we omit the proofs that measures with the stated properties exist. Much of technical measure theory concerns the existence of such measures, but, as far as applications go, their existence is intuitively reasonable, and can be taken on trust.

Example 1.1 The counting measure

For each subset A of $\mathbb{R}^n$, let $\mu(A)$ be the number of points in A if A is finite and ∞ otherwise. Then μ is a measure on $\mathbb{R}^n$.

Example 1.2 Point mass

Let a be a point in $\mathbb{R}^n$ and define $\mu(A)$ to be 1 if A contains a and 0 otherwise. Then μ is a mass distribution, thought of as a unit point mass concentrated at a.

Example 1.3 Lebesgue measure on $\mathbb{R}$

Lebesgue measure $\mathcal{L}^1$ extends the idea of 'length' to a large collection of subsets of $\mathbb{R}$ that includes the Borel sets. For open and closed intervals, we take $\mathcal{L}^1(a,b) = \mathcal{L}^1[a,b] = b-a$. If $A = \bigcup_i [a_i, b_i]$ is a finite or countable union of disjoint intervals, we let $\mathcal{L}^1(A) = \sum(b_i - a_i)$ be the length of A, thought of as the sum of the length of the intervals. This leads us to the definition of the *Lebesgue measure* $\mathcal{L}^1(A)$ of an arbitrary set A. We define

$$\mathcal{L}^1(A) = \inf\left\{ \sum_{i=1}^{\infty}(b_i - a_i) \,:\, A \subset \bigcup_{i=1}^{\infty}[a_i, b_i] \right\},$$

that is, we look at all coverings of A by countable collections of intervals and take the smallest total interval length possible. It is not hard to see that (1.1)–(1.3) hold; it is rather harder to show that (1.4) holds for disjoint Borel sets A_i, and we avoid this question here. (In fact, (1.4) holds for a much larger class of sets than the Borel sets, 'the Lebesgue measurable sets', but not for all subsets of $\mathbb{R}$.) Lebesgue measure on $\mathbb{R}$ is generally thought of as 'length', and we often write length(A) for $\mathcal{L}^1(A)$ when we wish to emphasise this intuitive meaning.

Example 1.4 Lebesgue measure on $\mathbb{R}^n$

We call a set of the form $A = \{(x_1, \ldots, x_n) \in \mathbb{R}^n \,:\, a_i \leqslant x_i \leqslant b_i\}$ a *coordinate parallelepiped* in $\mathbb{R}^n$, its n-dimensional volume of A is given by

$$\mathrm{vol}^n(A) = (b_1 - a_1)(b_2 - a_2)\cdots(b_n - a_n).$$

(Of course, if $n = 1$, a coordinate parallelepiped is just an interval with vol^1 as length, as in Example 1.3; if $n = 2$, it is a rectangle with vol^2 as area, and if $n = 3$, it is a cuboid with vol^3 the usual 3-dimensional volume.) Then *n-dimensional Lebesgue measure* $\mathcal{L}^n$ may be thought of as the extension of n-dimensional volume to a large class of sets. Just as in Example 1.3, we obtain a measure on $\mathbb{R}^n$ by defining

$$\mathcal{L}^n(A) = \inf\left\{ \sum_{i=1}^{\infty}\mathrm{vol}^n(A_i) \,:\, A \subset \bigcup_{i=1}^{\infty} A_i \right\}$$

where the infimum is taken over all coverings of A by coordinate parallelepipeds A_i. We get that $\mathcal{L}^n(A) = \mathrm{vol}^n(A)$ if A is a coordinate parallelepiped or, indeed, any set

for which the volume can be determined by the usual rules of mensuration. Again, to aid intuition, we sometimes write area(A) in place of $\mathcal{L}^2(A)$, vol(A) for $\mathcal{L}^3(A)$ and vol$^n(A)$ for $\mathcal{L}^n(A)$.

Sometimes, we need to define 'k-dimensional' volume on a k-dimensional plane X in $\mathbb{R}^n$; this may be done by identifying X with $\mathbb{R}^k$ and using $\mathcal{L}^k$ on subsets of X in the obvious way.

Example 1.5 Uniform mass distribution on a line segment

Let L be a line segment of unit length in the plane. For $A \subset \mathbb{R}^2$ define $\mu(A) = \mathcal{L}^1(L \cap A)$, that is, the 'length' of intersection of A with L. Then μ is a mass distribution with support L, because $\mu(A) = 0$ if $A \cap L = \emptyset$. We may think of μ as unit mass spread evenly along the line segment L.

Example 1.6 Restriction of a measure

Let μ be a measure on $\mathbb{R}^n$ and E a Borel subset of $\mathbb{R}^n$. We may define a measure ν on $\mathbb{R}^n$, called the *restriction of μ to E*, by $\nu(A) = \mu(E \cap A)$ for every set A. Then ν is a measure on $\mathbb{R}^n$ with support contained in $\overline{E}$.

As far as this book is concerned, the most important measures we shall meet are the s-dimensional Hausdorff measures $\mathcal{H}^s$ on subsets of $\mathbb{R}^n$, where $0 \leqslant s \leqslant n$. These measures, which are introduced in Section 3.1, are a generalisation of Lebesgue measures to dimensions that are not necessarily integral.

The following method is often used to construct a mass distribution on a subset of $\mathbb{R}^n$. It involves repeated subdivision of a mass between parts of a bounded Borel set E. Let $\mathcal{E}_0$ consist of the single set E. For $k = 1, 2, \ldots$, we let $\mathcal{E}_k$ be a collection of disjoint Borel subsets of E such that each set U in $\mathcal{E}_k$ is contained in one of the sets of $\mathcal{E}_{k-1}$ and contains a finite number of the sets in $\mathcal{E}_{k+1}$. We assume that the maximum diameter of the sets in $\mathcal{E}_k$ tends to 0 as $k \to \infty$. We define a mass distribution on E by repeated subdivision (see Figure 1.5). We let $\mu(E)$ satisfy $0 < \mu(E) < \infty$, and we split this mass between the sets $U_1, \ldots, U_m$ in $\mathcal{E}_1$ by defining $\mu(U_i)$ in such a way that $\sum_{i=1}^m \mu(U_i) = \mu(E)$. Similarly, we assign masses to the sets of $\mathcal{E}_2$ so that if $U_1, \ldots, U_m$ are the sets of $\mathcal{E}_2$ contained in a set U of $\mathcal{E}_1$, then $\sum_{i=1}^m \mu(U_i) = \mu(U)$. In general, we assign masses so that

$$\sum_i \mu(U_i) = \mu(U) \tag{1.8}$$

for each set U of $\mathcal{E}_k$, where the $\{U_i\}$ are the disjoint sets in $\mathcal{E}_{k+1}$ contained in U. For each k, we let E_k be the union of the sets in $\mathcal{E}_k$, and we define $\mu(A) = 0$ for all A with $A \cap E_k = \emptyset$.

Let $\mathcal{E}$ denote the collection of sets that belong to $\mathcal{E}_k$ for some k together with the subsets of the $\mathbb{R}^n \backslash E_k$. The above procedure defines the mass $\mu(A)$ of every set A in $\mathcal{E}$, and it should seem reasonable that, by building up sets from the sets in $\mathcal{E}$, it specifies enough about the distribution of the mass μ across $\mathcal{E}$ to determine $\mu(A)$ for any (Borel) set A. This is indeed the case, as the following proposition states.

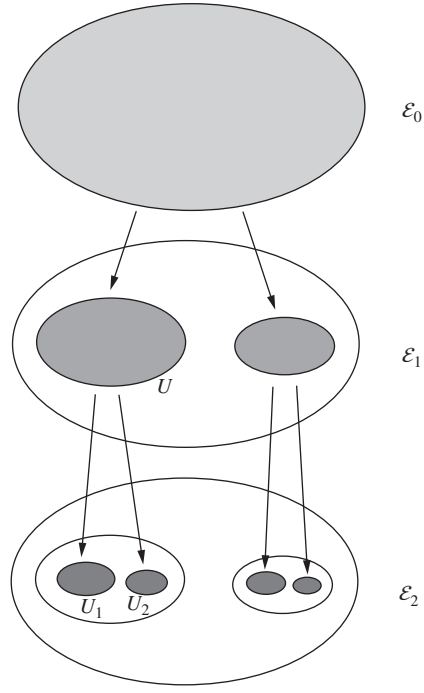

Figure 1.5 Steps in the construction of a mass distribution μ by repeated subdivision. The mass on the sets of $\mathcal{E}_k$ is divided between the sets of $\mathcal{E}_{k+1}$, so for example, $\mu(U) = \mu(U_1) + \mu(U_2)$.

Proposition 1.7

Let μ be defined on a collection of sets $\mathcal{E}$ as above. Then the definition of μ may be extended to all subsets of $\mathbb{R}^n$ so that μ becomes a measure. The value of $\mu(A)$ is uniquely determined if A is a Borel set. The support of μ is contained in $E_\infty = \bigcap_{k=1}^{\infty} \overline{E}_k$.

Note on Proof. If A is any subset of $\mathbb{R}^n$, let

$$\mu(A) = \inf \left\{ \sum_{i=1}^{\infty} \mu(U_i) \,:\, A \cap E_\infty \subset \bigcup_{i=1}^{\infty} U_i \text{ and } U_i \in \mathcal{E} \right\}. \tag{1.9}$$

(Thus, we take the smallest value we can of $\sum_{i=1}^{\infty} \mu(U_i)$ where the sets U_i are in $\mathcal{E}$ and cover $A \cap E_\infty$; we have already defined $\mu(U_i)$ for such U_i.) It is not difficult to see that if A is one of the sets in $\mathcal{E}$, then (1.9) reduces to the mass $\mu(A)$ specified in the construction. The complete proof that μ satisfies all the conditions of a measure and that its values on the sets of $\mathcal{E}$ determine its values on the Borel sets is quite involved, and need not concern us here. As $\mu(\mathbb{R}^n \backslash E_k) = 0$, we have $\mu(A) = 0$ if A is an open set that does not intersect E_k for some k, so the support of μ is in $\overline{E}_k$ for all k. □

Example 1.8 Lebesgue measure by repeated subdivision

Let $\mathcal{E}_k$ denote the collection of 'binary intervals' of length 2^{-k}, that is of the form $[r2^{-k}, (r+1)2^{-k})$ where $0 \leqslant r \leqslant 2^k - 1$. If we take $\mu[r2^{-k}, (r+1)2^{-k}) = 2^{-k}$ in the above construction, we get that μ is Lebesgue measure on $[0, 1]$.

To see this, note that if I is an interval in $\mathcal{E}_k$ of length 2^{-k} and I_1, I_2 are the two subintervals of I in $\mathcal{E}_{k+1}$ of length 2^{-k-1}, we have $\mu(I) = \mu(I_1) + \mu(I_2)$ which is (1.8). By Proposition 1.7, μ extends to a mass distribution on $[0, 1]$. We have $\mu(I) = \text{length}(I)$ for I in $\mathcal{E}$, and it may be shown that this implies that μ coincides with Lebesgue measure on any set.

We say that a property holds for *almost all* x, or *almost everywhere* (with respect to a measure μ) if the set for which the property fails has μ-measure zero. For example, we might say that almost all real numbers are irrational with respect to Lebesgue measure. The rational numbers $\mathbb{Q}$ are countable; they may be listed as $x_1, x_2, \ldots$, say, so that $\mathcal{L}^1(\mathbb{Q}) = \sum_{i=1}^{\infty} \mathcal{L}^1\{x_i\} = 0$.

Although we shall usually be interested in measures in their own right, we shall sometimes need to integrate functions with respect to measures. There are technical difficulties concerning which functions can be integrated. We may get around these difficulties by assuming that for $f : D \to \mathbb{R}$ a function defined on a Borel subset D of $\mathbb{R}^n$, the set $f^{-1}(-\infty, a] = \{x \in D : f(x) \leqslant a\}$ is a Borel set for all real numbers a. A very large class of functions satisfies this condition, including all continuous functions (for which $f^{-1}(-\infty, a]$ is closed and therefore a Borel set). We make the assumption throughout this book that all functions to be integrated satisfy this condition; this is true of functions that are likely to be encountered in this area of mathematics.

To define integration we first suppose that $f : D \to \mathbb{R}$ is a *simple function*, that is, one that takes only finitely many values $a_1, \ldots, a_k$. We define the *integral with respect to the measure* μ of a non-negative simple function f as

$$\int f \, \mathrm{d}\mu = \sum_{i=1}^{k} a_i \mu\{x : f(x) = a_i\}.$$

The integral of more general functions is defined using approximation by simple functions. If $f : D \to \mathbb{R}$ is a non-negative function, we define its integral as

$$\int f \mathrm{d}\mu = \sup \left\{ \int g \, \mathrm{d}\mu : g \text{ is simple}, 0 \leqslant g \leqslant f \right\}.$$

To complete the definition, if f takes both positive and negative values, we let $f_+(x) = \max\{f(x), 0\}$ and $f_-(x) = \max\{-f(x), 0\}$, so that $f = f_+ - f_-$, and define

$$\int f \mathrm{d}\mu = \int f_+ \mathrm{d}\mu - \int f_- \mathrm{d}\mu$$

provided that both $\int f_+ \, \mathrm{d}\mu$ and $\int f_- \, \mathrm{d}\mu$ are finite.

All the usual properties hold for integrals, for example,

$$\int (f+g)\mathrm{d}\mu = \int f\mathrm{d}\mu + \int g\,\mathrm{d}\mu$$

and

$$\int \lambda f\mathrm{d}\mu = \lambda \int f\mathrm{d}\mu$$

if λ is a scalar. Very useful is the monotone convergence theorem, that is, if $f_k : D \to \mathbb{R}$ is an increasing sequence of non-negative functions converging (pointwise) to f, then

$$\lim_{k\to\infty} \int f_k \mathrm{d}\mu = \int f\mathrm{d}\mu.$$

If A is a Borel subset of D, we define integration over the set A by

$$\int_A f\mathrm{d}\mu = \int f\,\chi_A\,\mathrm{d}\mu$$

where $\chi_A : \mathbb{R}^n \to \mathbb{R}$ is the 'indicator function' of A, defined by $\chi_A(x) = 1$ if x is in A and $\chi_A(x) = 0$ otherwise.

Note that if $f(x) \geqslant 0$ and $\int f\mathrm{d}\mu = 0$, then $f(x) = 0$ for μ-almost all x.

As usual, integration is denoted in various ways, such as $\int f\mathrm{d}\mu, \int f$ or $\int f(x)\mathrm{d}\mu(x)$, depending on the emphasis required. When μ is n-dimensional Lebesgue measure $\mathcal{L}^n$, we usually write $\int f\ \mathrm{d}x$ or $\int f(x)\mathrm{d}x$ in place of $\int f\ \mathrm{d}\mathcal{L}^n$.

On a couple of occasions we shall need to use Egoroff's theorem. Let D be a Borel subset of $\mathbb{R}^n$ and μ a measure with $\mu(D) < \infty$. Let $f_1, f_2, \ldots$ and f be functions from D to $\mathbb{R}$ such that $f_k(x) \to f(x)$ for each x in D. Egoroff's theorem states that for any $\delta > 0$, there is a Borel subset E of D such that $\mu(D\backslash E) < \delta$ and such that the sequence $\{f_k\}$ converges uniformly to f on E, that is, with $\sup_{x\in E}|f_k(x) - f(x)| \to 0$ as $k \to \infty$. For the measures that we shall be concerned with, it may be shown that we can always take the set E to be compact.

1.4 Notes on probability theory

Understanding some of the later chapters of the book requires a basic knowledge of probability theory. We provide a very brief overview of the concepts needed.

Probability theory starts with the idea of an *experiment* or *trial*; that is, an action whose outcome is, for all practical purposes, not predetermined. Mathematically, such an experiment is described by a probability space, which has three components: the set of all possible outcomes of the experiment, the list of all the events that may occur as consequences of the experiment and an assessment of likelihood of these events. For example, if a die is thrown, the possible outcomes are $\{1, 2, 3, 4, 5, 6\}$, the list of events includes 'a 3 is thrown', 'an even number is thrown' and 'at least a 4 is thrown'. For a 'fair die', it may be reasonable to assess the six possible outcomes as equally likely.

The set of all possible outcomes of an experiment is called the *sample space*, denoted by Ω. Questions of interest concerning the outcome of an experiment can always be phrased in terms of subsets of Ω; in the above example, 'is an odd number thrown?' asks 'is the outcome in the subset $\{1, 3, 5\}$?' Associating events dependent on the outcome of the experiment with subsets of Ω in this way, it is natural to think of the union $A \cup B$ as 'either A or B occurs', the intersection $A \cap B$ as 'both A and B occur', and the complement $\Omega \backslash A$ as the event 'A does not occur', for any events A and B. In general, there is a collection $\mathcal{F}$ of subsets of Ω that particularly interest us, which we call *events*. In the example of the die, $\mathcal{F}$ would normally be the collection of all subsets of Ω, but in more complicated situations, a relatively small collection of subsets might be relevant. Usually, $\mathcal{F}$ satisfies certain conditions; for example, if the occurrence of an event interests us, then so does its non-occurrence, so if A is in $\mathcal{F}$, we would expect the complement $\Omega \backslash A$ also to be in $\mathcal{F}$. We call a (non-empty) collection $\mathcal{F}$ of subsets of the sample space Ω an *event space* if

$$\Omega \backslash A \in \mathcal{F} \quad \text{whenever } A \in \mathcal{F} \tag{1.10}$$

and

$$\bigcup_{i=1}^{\infty} A_i \in \mathcal{F} \quad \text{whenever } A_i \in \mathcal{F} \quad (1 \leqslant i < \infty). \tag{1.11}$$

It follows from these conditions that $\emptyset$ and Ω are in $\mathcal{F}$ and that $A \backslash B$ and $\bigcap_{i=1}^{\infty} A_i$ are in $\mathcal{F}$ whenever A, B and A_i are in $\mathcal{F}$. As far as our applications are concerned, we do not, in general, specify $\mathcal{F}$ precisely – this avoids technical difficulties connected with the existence of suitable event spaces.

Next, we associate probabilities with the events of $\mathcal{F}$, with $\mathsf{P}(A)$ thought of as the probability, or likelihood, that the event A occurs. We call P a *probability* or *probability measure* if P assigns a number $\mathsf{P}(A)$ to each A in $\mathcal{F}$, such that the following conditions hold:

$$0 \leqslant \mathsf{P}(A) \leqslant 1 \text{ for all } A \in \mathcal{F} \tag{1.12}$$

$$\mathsf{P}(\emptyset) = 0 \text{ and } \mathsf{P}(\Omega) = 1 \tag{1.13}$$

and if $A_1, A_2, \ldots$ are disjoint events in $\mathcal{F}$,

$$\mathsf{P}\left(\bigcup_{i=1}^{\infty} A_i\right) = \sum_{i=1}^{\infty} \mathsf{P}(A_i). \tag{1.14}$$

It should seem natural for any definition of probability to satisfy these conditions.

We call a triple $(\Omega, \mathcal{F}, \mathsf{P})$ a *probability space* if $\mathcal{F}$ is an event space of subsets of Ω and P is a probability measure defined on the sets of $\mathcal{F}$.

For the die-throwing experiment, we might have $\Omega = \{1, 2, 3, 4, 5, 6\}$ with the event space consisting of all subsets of Ω, and with $\mathsf{P}(A) = \frac{1}{6} \times$ number of elements in A. This describes the 'fair die' situation with each outcome equally likely.

Often, Ω is an infinite set. For example, we might have $\Omega = [0, 1]$ and think of a random number drawn from $[0, 1]$ with the probability of the number in a set A as $\mathsf{P}(A) = \text{length}(A)$. Here, the event space might be the Borel subsets of $[0, 1]$.

The resemblance of the definition of probability to the definition of a measure in (1.1)–(1.4) and the use of the term probability measure is no coincidence. Probabilities and measures may be put into the same context, with Ω corresponding to $\mathbb{R}^n$ and with the event space in some ways analogous to the Borel sets.

In our applications later on in the book, we shall be particularly interested in events (on rather large sample spaces such as spaces of continuous functions) that are virtually certain to occur. We say that an event A occurs *with probability* 1 or *almost surely* if $\mathsf{P}(A) = 1$.

Sometimes, we may possess partial information about the outcome of an experiment; for example, we might be told that the number showing on the die is even. This leads us to reassess the probabilities of the various events. If A and B are in $\mathcal{F}$ with $\mathsf{P}(B) > 0$, the *conditional probability of A given B*, denoted by $\mathsf{P}(A|B)$, is defined by

$$\mathsf{P}(A|B) = \frac{\mathsf{P}(A \cap B)}{\mathsf{P}(B)}. \tag{1.15}$$

This is thought of as the probability of A given that the event B is known to occur; as would be expected $\mathsf{P}(B|B) = 1$. It is easy to show that $(\Omega, \mathcal{F}, \mathsf{P}')$ is a probability space, where $\mathsf{P}'(A) = \mathsf{P}(A|B)$. We also have the partition formula: if $B_1, B_2, \ldots$ are disjoint events with $\bigcup_i B_i = \Omega$ and $\mathsf{P}(B_i) > 0$ for all i, then for an event A,

$$\mathsf{P}(A) = \sum_i \mathsf{P}(A|B_i)\mathsf{P}(B_i). \tag{1.16}$$

In the case of the 'fair die' experiment, if B_1 is the event 'an even number is thrown', B_2 is 'an odd number is thrown' and A is 'at least 4 is thrown', then

$$\mathsf{P}(A|B_1) = \mathsf{P}(4 \text{ or } 6 \text{ is thrown})/\mathsf{P}(2, 4 \text{ or } 6 \text{ is thrown}) = \tfrac{2}{6}/\tfrac{3}{6} = \tfrac{2}{3}.$$
$$\mathsf{P}(A|B_2) = \mathsf{P}(5 \text{ is thrown})/\mathsf{P}(1, 3 \text{ or } 5 \text{ is thrown}) = \tfrac{1}{6}/\tfrac{3}{6} = \tfrac{1}{3}$$

from which (1.16) is easily verified.

We think of two events as independent if the occurrence of one does not affect the probability that the other occurs, that is, if $\mathsf{P}(A|B) = \mathsf{P}(A)$ and $\mathsf{P}(B|A) = \mathsf{P}(B)$. Using (1.15), we are led to make the definition that two events A and B in a probability space are *independent* if

$$\mathsf{P}(A \cap B) = \mathsf{P}(A)\mathsf{P}(B). \tag{1.17}$$

More generally, an arbitrary collection of events is independent if for every finite subcollection $\{A_k : k \in J\}$ we have

$$\mathsf{P}\left(\bigcap_{k\in J} A_k\right) = \prod_{k\in J} \mathsf{P}(A_k). \tag{1.18}$$

In the die example, it is easy to see that 'a throw of at least 5' and 'an even number is thrown' are independent events, but 'a throw of at least 4' and 'an even number is thrown' are not.

The idea of a random variable and its expectation (or average or mean) is fundamental to probability theory. Essentially, a random variable X is a real-valued function on a sample space. In the die example, X might represent the score on the die. Alternatively, it might represent the reward for throwing a particular number, for example, $X(\omega) = 0$ if $\omega = 1, 2, 3$, or 4, $X(5) = 1$ and $X(6) = 2$. The outcome of an experiment determines a value of the random variable. The expectation of the random variable is the average of these values weighted according to the likelihood of each outcome.

The precise definition of a random variable requires a little care. We say that X is a *random variable* on a probability space $(\Omega, \mathcal{F}, \mathsf{P})$ if $X : \Omega \to \mathbb{R}$ is a function such that $X^{-1}((-\infty, a])$ is an event in $\mathcal{F}$ for each real number a; in other words, the set of ω in Ω with $X(\omega) \leqslant a$ is in the event space. This condition is equivalent to saying that $X^{-1}(E)$ is in $\mathcal{F}$ for any Borel set E. In particular, for any such E, the probability that the random variable X takes a value in E, that is, $\mathsf{P}(\{\omega : X(\omega) \in E\})$, is defined. It may be shown that $\mathsf{P}(\{\omega : X(\omega) \in E\})$ is determined for all Borel sets E from a knowledge of $\mathsf{P}(\{\omega : X(\omega) \leqslant a\})$ for each real number a. Note that it is usual to abbreviate expressions such as $\mathsf{P}(\{\omega : X(\omega) \in E\})$ to $\mathsf{P}(X \in E)$.

It is not difficult to show that if X and Y are random variables on $(\Omega, \mathcal{F}, \mathsf{P})$ and λ is a real number, then $X + Y, X - Y, XY$ and λX are all random variables (these are defined in the obvious way, e.g. $(X + Y)(\omega) = X(\omega) + Y(\omega)$ for each $\omega \in \Omega$). Moreover, if $X_1, X_2, \ldots$ is a sequence of random variables with $X_k(\omega)$ increasing and bounded for each ω, then $\lim_{k\to\infty} X_k$ is a random variable.

A collection of random variables $\{X_k\}$ is *independent* if, for any Borel sets E_k, the events $\{(X_k \in E_k)\}$ are independent in the sense of (1.18); that is, if, for every finite set of indices J,

$$\mathsf{P}(X_k \in E_k \text{ for all } k \in J) = \prod_{k \in J} \mathsf{P}(X_k \in E_k).$$

Intuitively, X and Y are independent if the probability of Y taking any particular value is unaffected by a knowledge of the value of X. Consider the probability space representing two successive throws of a die, with sample space $\{(x, y) : x, y = 1, 2, \ldots, 6\}$ and probability measure P defined by $\mathsf{P}\{(x, y)\} = \frac{1}{36}$ for each pair (x, y). If X and Y are the random variables given by the scores on successive throws, then X and Y are independent, modelling the assumption that one throw does not affect the other. However, X and $X + Y$ are not independent – this reflects that the bigger the score for the first throw, the greater the chance of a high total score.

The formal definition of the expectation of a random variable is analogous to the definition of the integral of a function; indeed, expectation is really the integral of the random variable with respect to the probability measure. Let X be a random variable on a probability space $(\Omega, \mathcal{F}, \mathsf{P})$. First suppose that $X(\omega) \geqslant 0$ for all ω in Ω and that X takes only finitely many values $x_1, \ldots, x_k$; we call such a random

variable *simple*. We define the *expectation, mean* or *average* $\mathsf{E}(X)$ of X as

$$\mathsf{E}(X) = \sum_{i=1}^{k} x_i \mathsf{P}(X = x_i). \tag{1.19}$$

The expectation of an arbitrary random variable is defined using approximation by simple random variables. Thus for a non-negative random variable X

$$\mathsf{E}(X) = \sup\{\mathsf{E}(Y) : Y \text{ is a simple random variable with } 0 \leqslant Y(\omega) \leqslant X(\omega) \text{ for all } \omega \in \Omega\}.$$

Lastly, if X takes both positive and negative values, we let $X_+ = \max\{X, 0\}$ and $X_- = \max\{-X, 0\}$, so that $X = X_+ - X_-$, and define

$$\mathsf{E}(X) = \mathsf{E}(X_+) - \mathsf{E}(X_-)$$

provided that $\mathsf{E}(X_+) < \infty$ and $\mathsf{E}(X_-) < \infty$.

The random variable X representing the score of a fair die is a simple random variable, because $X(\omega)$ takes just the values $1, \ldots, 6$. Thus

$$\mathsf{E}(X) = \sum_{i=1}^{6} \left(i \times \tfrac{1}{6} \right) = 3\tfrac{1}{2}.$$

Expectation satisfies certain basic properties, analogous to those for the integral. If $X_1, X_2, \ldots$ are random variables, then

$$\mathsf{E}(X_1 + X_2) = \mathsf{E}(X_1) + \mathsf{E}(X_2)$$

and, more generally,

$$\mathsf{E}\left(\sum_{i=1}^{k} X_i \right) = \sum_{i=1}^{k} \mathsf{E}(X_i).$$

If λ is a constant,

$$\mathsf{E}(\lambda X) = \lambda \mathsf{E}(X)$$

and if the sequence of non-negative random variables $X_1, X_2, \ldots$ is increasing with $X = \lim_{k\to\infty} X_k$ a (finite) random variable, then

$$\lim_{k\to\infty} \mathsf{E}(X_k) = \mathsf{E}(X).$$

Provided that X_1 and X_2 are independent, we also have

$$\mathsf{E}(X_1 X_2) = \mathsf{E}(X_1)\mathsf{E}(X_2).$$

Thus, if X_i represents that kth throw of a fair die in a sequence of throws, the expectation of the sum of the first k throws is $\mathsf{E}(X_1 + \cdots + X_k) = \mathsf{E}(X_1) + \cdots + \mathsf{E}(X_k) = 3\frac{1}{2} \times k$.

We define the *conditional expectation* $\mathsf{E}(X|B)$ of X given an event B with $\mathsf{P}(B) > 0$ in a similar way but starting with

$$\mathsf{E}(X|B) = \sum_{i=1}^{k} x_i \mathsf{P}(X = x_i|B) \tag{1.20}$$

in place of (1.19). We get a partition formula resembling (1.16)

$$\mathsf{E}(X) = \sum_i \mathsf{E}(X|B_i)\mathsf{P}(B_i), \tag{1.21}$$

where $B_1, B_2, \ldots$ are disjoint events with $\bigcup_i B_i = \Omega$ and $\mathsf{P}(B_i) > 0$.

It is often useful to have an indication of the fluctuation of a random variable across a sample space. Thus we introduce the *variance* of the random variable X as

$$\begin{aligned}\mathrm{var}(X) &= \mathsf{E}((X - \mathsf{E}(X))^2) \\ &= \mathsf{E}(X^2) - \mathsf{E}(X)^2\end{aligned}$$

by a simple calculation. Using the properties of expectation, we get

$$\mathrm{var}(\lambda X) = \lambda^2 \mathrm{var}(X),$$

for any real number λ, and

$$\mathrm{var}(X + Y) = \mathrm{var}(X) + \mathrm{var}(Y)$$

provided that X and Y are independent.

If the probability distribution of a random variable is given by an integral, that is,

$$\mathsf{P}(X \leqslant x) = \int_{-\infty}^{x} f(u)\, \mathrm{d}u, \tag{1.22}$$

the function f is called the *probability density function* for X. It may be shown from the definition of expectation that

$$\mathsf{E}(X) = \int_{-\infty}^{\infty} u f(u)\, \mathrm{d}u$$

and

$$\mathsf{E}(X^2) = \int_{-\infty}^{\infty} u^2 f(u)\, \mathrm{d}u$$

which allows $\mathrm{var}(X) = \mathsf{E}(X^2) - \mathsf{E}(X)^2$ to be calculated.

Note that the density function tells us about the distribution of the random variable X without reference to the underlying probability space, which, for many purposes, is irrelevant. We may express the probability that X belongs to any Borel set E in terms of the density function as

$$\mathsf{P}(X \in E) = \int_E f(u)\, \mathrm{d}u.$$

We say that a random variable X has *uniform distribution* on the interval $[a, b]$ if

$$\mathsf{P}(X \leqslant x) = \frac{1}{b-a} \int_a^x \mathrm{d}u \qquad (a \leqslant x \leqslant b). \tag{1.23}$$

Thus, the probability of X lying in a subinterval of $[a, b]$ is proportional to the length of the interval. In this case, we get that $\mathsf{E}(X) = \frac{1}{2}(a+b)$ and $\mathrm{var}(X) = \frac{1}{12}(b-a)^2$.

A random variable X has *normal* or *Gaussian distribution* of mean m and variance σ^2 if

$$\mathsf{P}(X \leqslant x) = \frac{1}{\sigma\sqrt{2\pi}} \int_{-\infty}^x \exp\left(\frac{-(u-m)^2}{2\sigma^2}\right) \mathrm{d}u. \tag{1.24}$$

It may be verified by integration that $\mathsf{E}(X) = m$ and $\mathrm{var}(X) = \sigma^2$. If X_1 and X_2 are independent normally distributed random variables of means m_1 and m_2 and variances σ_1^2 and σ_2^2, respectively, then $X_1 + X_2$ is normal with mean $m_1 + m_2$ and variance $\sigma_1^2 + \sigma_2^2$, and λX_1 is normal with mean λm_1 and variance $\lambda^2\sigma_1^2$, for any real number λ.

If we throw a fair die a large number of times, we might expect the average score thrown to be very close to $3\frac{1}{2}$, the expectation or mean outcome of each throw. Moreover, the larger the number of throws, the closer the average should be to the mean. This 'law of averages' is made precise as the strong law of large numbers.

Let $(\Omega, \mathcal{F}, \mathsf{P})$ be a probability space. Let $X_1, X_2, \ldots$ be random variables that are independent and that have identical distribution (i.e. for every set E, $\mathsf{P}(X_i \in E)$ is the same for all i), with expectation m and variance σ^2, both assumed finite. For each k, we may form the random variable $S_k = X_1 + \cdots + X_k$, so that the random variable S_k/k is the average of the first k trials. The *strong law of large numbers* states that, with probability 1, this average approaches the mean, that is,

$$\lim_{k\to\infty} \frac{1}{k} S_k = m. \tag{1.25}$$

We can also say a surprising amount about the distribution of the random variable S_k when k is large. It may be shown that S_k has approximately the normal distribution with mean km and variance $k\sigma^2$. This is the content of the *central limit theorem*, which states that for every real number x,

$$\mathsf{P}\left(\frac{S_k - km}{\sigma\sqrt{k}} \leqslant x\right) \to \int_{-\infty}^x \frac{1}{\sqrt{2\pi}} \exp\left(-\tfrac{1}{2}u^2\right) \mathrm{d}u \quad \text{as } k \to \infty, \tag{1.26}$$

that is, $(S_k - km)/(\sigma\sqrt{k})$ converges (in some sense) to a normal distribution. This is one reason why the normal distribution is so important – it is the form of distribution approached by sums of a large number of independent identically distributed random variables.

We may apply these remarks to the experiment consisting of an infinite sequence of die throws. Let Ω be the set of all infinite sequences $\{\omega = (\omega_1, \omega_2, \ldots) : \omega_i = 1, 2, \ldots, 6\}$ (we think of ω_i as the outcome of the kth

throw). It is possible to define an event space $\mathcal{F}$ and probability measure P in such a way that for any given k and sequence $\omega_1, \ldots, \omega_k$ ($\omega_i = 1, 2, \ldots, 6$), the event 'the first k throws are $\omega_1, \ldots, \omega_k$' is in $\mathcal{F}$ and has probability $\left(\frac{1}{6}\right)^k$. Let X_k be the random variable given by the outcome of the kth throw, so that $X_k(\omega) = \omega_k$. It is easy to see that the X_k are independent and identically distributed, with mean $m = 3\frac{1}{2}$ and variance $2\frac{11}{12}$. The strong law of large numbers tells us that, with probability 1, the average of the first k throws, S_k/k converges to $3\frac{1}{2}$ as k tends to infinity, and the central limit theorem tells us that when k is large, the sum S_k is approximately normally distributed, with mean $3\frac{1}{2} \times k$ and variance $2\frac{11}{12} \times k$. Thus, if we repeat the experiment of throwing k dice a large number of times, the sum of the k throws will have a distribution close to the normal distribution, in the sense of (1.26).

1.5 Notes and references

The material outlined in this chapter is covered at various levels of sophistication in numerous undergraduate and graduate mathematical texts. Any of the many books on mathematical analysis, for example, the classics by Apostol (1974); Rudin (1976) or Howie (2001), contain the basic theory of sets and functions. For thorough treatments of measure theory see, for example, Taylor (1973a); Edgar (1998); Capinski and Kopp (2007) or Tao (2011). For probability theory, Grimmett and Stirzaker (2001) or Billingsley (2012) may be helpful.

Exercises

The following exercises do no more than emphasise some of the many facts that have been mentioned in this chapter.

1.1 Verify that for $x, y, z \in \mathbb{R}^n$, (i) $|x + y| \leqslant |x| + |y|$, (ii) $|x - y| \geqslant ||x| - |y||$ and (iii) $|x - y| \leqslant |x - z| + |z - y|$.

1.2 Show from the definition of δ-neighbourhood that $A_{\delta+\delta'} = (A_\delta)_{\delta'}$.

1.3 Show that a (non-empty) set is bounded if and only if it is contained in some ball $B(0, r)$ with centre the origin.

1.4 Determine which of the following subsets of $\mathbb{R}$ are open and which are closed. In each case, determine the interior and closure of the set. (i) A non-empty finite set A, (ii) the interval $(0, 1)$, (iii) the interval $[0, 1]$, (iv) the interval $[0, 1)$, (v) the set $\left\{0, 1, \frac{1}{2}, \frac{1}{3}, \frac{1}{4}, \ldots\right\}$.

1.5 Show that the middle third Cantor set, Figure 0.1, is compact and totally disconnected. What is its interior, closure and boundary?

1.6 Show that the union of any collection of open subsets of $\mathbb{R}^n$ is open and that the intersection of any finite collection of open sets is open. Show that

a subset of $\mathbb{R}^n$ is closed if and only if its complement is open and hence deduce the corresponding result for unions and intersections of closed sets.

1.7 Show that if $A_1 \supset A_2 \supset \cdots$ is a decreasing sequence of non-empty compact subsets of $\mathbb{R}^n$ then $\bigcap_{k=1}^{\infty} A_k$ is a non-empty compact set.

1.8 Show that the half-open interval $[0,1) = \{x \in \mathbb{R} : 0 \leqslant x < 1\}$ is a Borel subset of $\mathbb{R}$.

1.9 Let F be the set of numbers in $[0,1]$ whose decimal expansions contain the digit 5 infinitely many times. Show that F is a Borel set.

1.10 Show that the coordinate transformation of the plane

$$\begin{pmatrix} x_1 \\ x_2 \end{pmatrix} \mapsto \begin{pmatrix} c\cos\theta & -c\sin\theta \\ c\sin\theta & c\cos\theta \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} + \begin{pmatrix} a_1 \\ a_2 \end{pmatrix}$$

is a similarity of ratio c, and describe the transformation geometrically.

1.11 Find $\underline{\lim}_{x\to 0} f(x)$ and $\overline{\lim}_{x\to 0} f(x)$ where $f : \mathbb{R}^+ \to \mathbb{R}$ is given by (i) $\sin x$; (ii) $\sin(1/x)$ and (iii) $x^2 + (3+x)\sin(1/x)$.

1.12 Let $f, g : [0,1] \to \mathbb{R}$ be Lipschitz functions. Show that the functions defined on $[0,1]$ by $f(x) + g(x)$ and $f(x)g(x)$ are also Lipschitz.

1.13 Let $f : \mathbb{R} \to \mathbb{R}$ be differentiable with $|f'(x)| \leqslant c$ for all x. Show, using the mean value theorem, that f is a Lipschitz function.

1.14 Show that every Lipschitz function $f : \mathbb{R} \to \mathbb{R}$ is continuous.

1.15 Let $f : \mathbb{R} \to \mathbb{R}$ be given by $f(x) = x^2 + x$. Find (i) $f^{-1}(2)$, (ii) $f^{-1}(-2)$ and (iii) $f^{-1}([2,6])$.

1.16 Show that $f(x) = x^2$ is Lipschitz on $[0,2]$, bi-Lipschitz on $[1,2]$ and not Lipschitz on $\mathbb{R}$.

1.17 Show that if E is a compact subset of $\mathbb{R}^n$ and $f : E \to \mathbb{R}^n$ is continuous, then $f(E)$ is compact.

1.18 Let $A_1, A_2, \ldots$ be a decreasing sequence of Borel subsets of $\mathbb{R}^n$ and let $A = \bigcap_{k=1}^{\infty} A_k$. If μ is a measure on $\mathbb{R}^n$ with $\mu(A_1) < \infty$, show using (1.6) that $\mu(A_k) \to \mu(A)$ as $k \to \infty$.

1.19 Show that the point mass concentrated at a (see Example 1.2) is a measure.

1.20 Show how to define a mass distribution on the middle third Cantor set, Figure 0.1, in as uniform a way as possible.

1.21 Verify that Lebesgue measure satisfies (1.1)–(1.3).

1.22 Let $f : [0,1] \to \mathbb{R}$ be a continuous function. For A a subset of $\mathbb{R}^2$ define $\mu(A) = \mathcal{L}\{x : (x, f(x)) \in A\}$, where $\mathcal{L}$ is Lebesgue measure. Show that μ is a mass distribution on $\mathbb{R}^2$ supported by the graph of f.

1.23 Let D be a Borel subset of $\mathbb{R}^n$ and let μ be a measure on D with $\mu(D) < \infty$. Let $f_k : D \to \mathbb{R}$ be a sequence of functions such that $f_k(x) \to f(x)$ for all x in D. Prove Egoroff's theorem: that given $\varepsilon > 0$, there exists a Borel subset A of D with $\mu(D \backslash A) < \varepsilon$ such that $f_k(x)$ converges to $f(x)$ uniformly for x in A.

1.24 Prove that if μ is a measure on D and $f : D \to \mathbb{R}$ satisfies $f(x) \geqslant 0$ for all x in D and $\int_D f \, \mathrm{d}\mu = 0$ then $f(x) = 0$ for μ-almost all x.

1.25 If X is a random variable show that $\mathsf{E}((X - \mathsf{E}(X))^2) = \mathsf{E}(X^2) - \mathsf{E}(X)^2$ (these numbers equalling the variance of X).

1.26 Verify that if X has the uniform distribution on $[a, b]$ (see (1.23)), then $\mathsf{E}(X) = \frac{1}{2}(a + b)$ and $\mathrm{var}(X) = (b - a)^2/12$.

1.27 Let $A_1, A_2, \ldots$ be a sequence of independent events in some probability space such that $\mathsf{P}(A_k) = p$ for all k, where $0 < p < 1$. Let N_k be the random variable defined by taking N_k to equal the number of i with $1 \leqslant i \leqslant k$ for which A_i occurs. Use the strong law of large numbers to show that, with probability 1, $N_k/k \to p$ as $k \to \infty$. Deduce that the proportion of successes in a sequence of independent trials converges to the probability of success of each trial.

1.28 A fair die is thrown 6000 times. Use the central limit theorem to estimate the probability that at least 1050 sixes are thrown. (A numerical method will be needed if the integral obtained is to be evaluated.)