Introduction

In the past, mathematics has been concerned largely with sets and functions to which the methods of classical calculus can be applied. Sets or functions that are not sufficiently smooth or regular have tended to be ignored as 'pathological' and not worthy of study. Certainly, they were regarded as individual curiosities and only rarely were thought of as a class to which a general theory might be applicable.

In recent years this attitude has changed. It has been realized that a great deal can be said, and is worth saying, about the mathematics of non-smooth objects. Moreover, irregular sets provide a much better representation of many natural phenomena than do the figures of classical geometry. Fractal geometry provides a general framework for the study of such irregular sets.

We begin by looking briefly at a number of simple examples of fractals, and note some of their features.

The middle third Cantor set is one of the best known and most easily constructed fractals; nevertheless it displays many typical fractal characteristics. It is constructed from a unit interval by a sequence of deletion operations; see figure 0.1. Let E_0 be the interval [0, 1]. (Recall that $[a, b]$ denotes the set of real numbers x such that $a \leqslant x \leqslant b$.) Let E_1 be the set obtained by deleting the middle third of E_0, so that E_1 consists of the two intervals $[0, \frac{1}{3}]$ and $[\frac{2}{3}, 1]$. Deleting the middle thirds of these intervals gives E_2; thus E_2 comprises the four intervals $[0, \frac{1}{9}]$, $[\frac{2}{9}, \frac{1}{3}]$, $[\frac{2}{3}, \frac{7}{9}]$, $[\frac{8}{9}, 1]$. We continue in this way, with E_k obtained by deleting the middle third of each interval in E_{k-1}. Thus E_k consists of 2^k intervals each of length 3^{-k}. The *middle third Cantor set F* consists of the numbers that are in E_k for all k; mathematically, F is the intersection $\bigcap_{k=0}^{\infty} E_k$. The Cantor set F may be thought of as the limit of the sequence of sets E_k as k tends to infinity. It is obviously impossible to draw the set F itself, with its infinitesimal detail, so 'pictures of F' tend to be pictures of one of the E_k, which are a good approximation to F when k is reasonably large; see figure 0.1.

At first glance it might appear that we have removed so much of the interval [0, 1] during the construction of F, that nothing remains. In fact, F is an infinite (and indeed uncountable) set, which contains infinitely many numbers in every neighbourhood of each of its points. The middle third Cantor set F consists

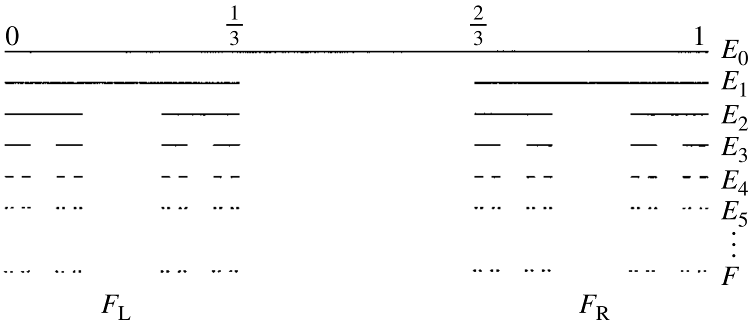


Figure 0.1 Construction of the middle third Cantor set F, by repeated removal of the middle third of intervals. Note that F_L and F_R, the left and right parts of F, are copies of F scaled by a factor $\frac{1}{3}$

precisely of those numbers in [0, 1] whose base-3 expansion does not contain the digit 1, i.e. all numbers $a_1 3^{-1} + a_2 3^{-2} + a_3 3^{-3} + \cdots$ with $a_i = 0$ or 2 for each i. To see this, note that to get E_1 from E_0 we remove those numbers with $a_1 = 1$, to get E_2 from E_1 we remove those numbers with $a_2 = 1$, and so on.

We list some of the features of the middle third Cantor set F; as we shall see, similar features are found in many fractals.

(i) F is self-similar. It is clear that the part of F in the interval $[0, \frac{1}{3}]$ and the part of F in $[\frac{2}{3}, 1]$ are both geometrically similar to F, scaled by a factor $\frac{1}{3}$. Again, the parts of F in each of the four intervals of E_2 are similar to F but scaled by a factor $\frac{1}{9}$, and so on. The Cantor set contains copies of itself at many different scales.

(ii) The set F has a 'fine structure'; that is, it contains detail at arbitrarily small scales. The more we enlarge the picture of the Cantor set, the more gaps become apparent to the eye.

(iii) Although F has an intricate detailed structure, the actual definition of F is very straightforward.

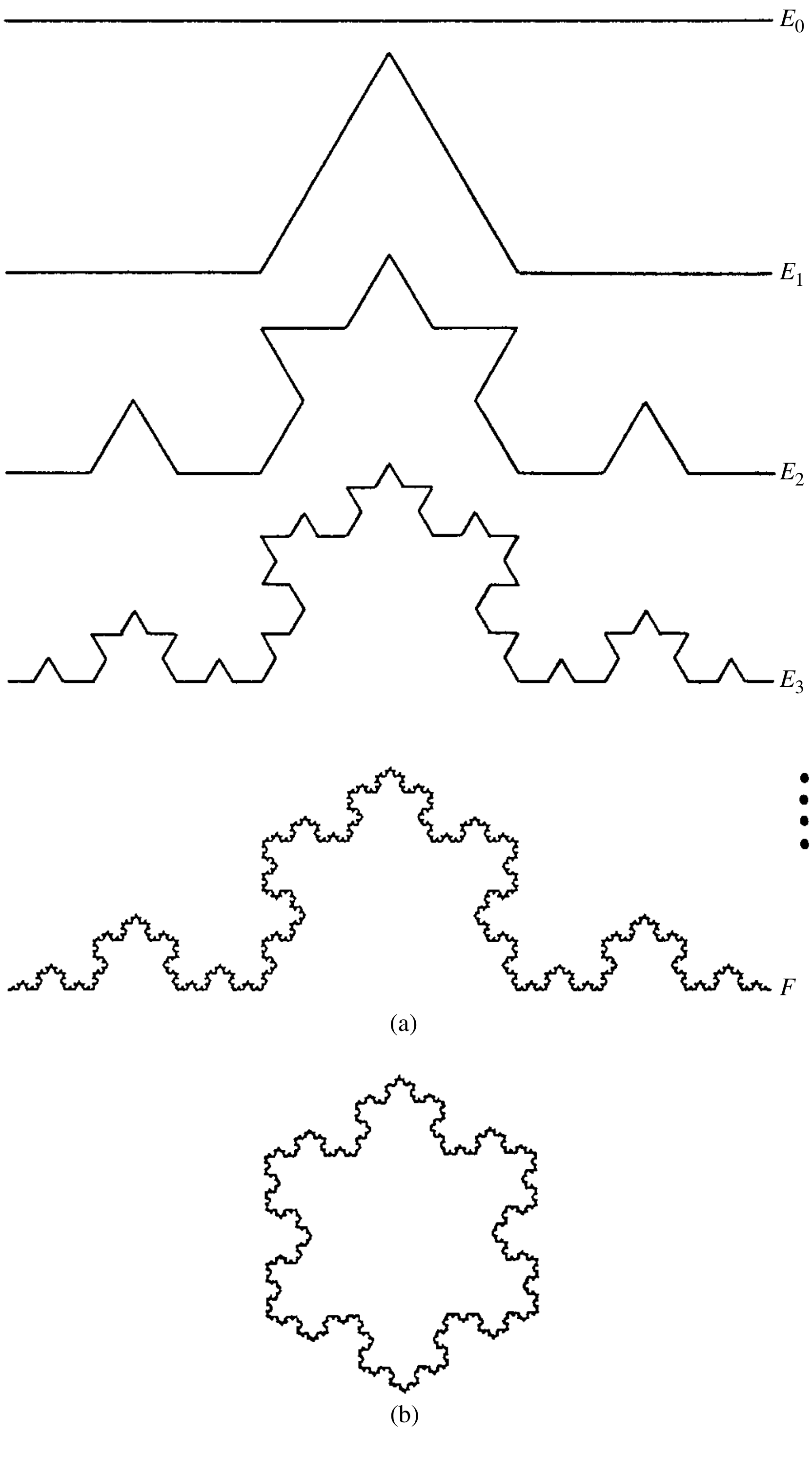
(iv) F is obtained by a recursive procedure. Our construction consisted of repeatedly removing the middle thirds of intervals. Successive steps give increasingly good approximations E_k to the set F.

(v) The geometry of F is not easily described in classical terms: it is not the locus of the points that satisfy some simple geometric condition, nor is it the set of solutions of any simple equation.

(vi) It is awkward to describe the local geometry of F—near each of its points are a large number of other points, separated by gaps of varying lengths.

(vii) Although F is in some ways quite a large set (it is uncountably infinite), its size is not quantified by the usual measures such as length—by any reasonable definition F has length zero.

Our second example, the von Koch curve, will also be familiar to many readers; see figure 0.2. We let E_0 be a line segment of unit length. The set E_1 consists of the four segments obtained by removing the middle third of E_0 and replacing it



(a)

(b)

Figure 0.2 (*a*) Construction of the von Koch curve F. At each stage, the middle third of each interval is replaced by the other two sides of an equilateral triangle. (*b*) Three von Koch curves fitted together to form a snowflake curve

by the other two sides of the equilateral triangle based on the removed segment. We construct E_2 by applying the same procedure to each of the segments in E_1, and so on. Thus E_k comes from replacing the middle third of each straight line segment of E_{k-1} by the other two sides of an equilateral triangle. When k is large, the curves E_{k-1} and E_k differ only in fine detail and as k tends to infinity, the sequence of polygonal curves E_k approaches a limiting curve F, called the *von Koch curve*.

The von Koch curve has features in many ways similar to those listed for the middle third Cantor set. It is made up of four 'quarters' each similar to the whole, but scaled by a factor $\frac{1}{3}$. The fine structure is reflected in the irregularities at all scales; nevertheless, this intricate structure stems from a basically simple construction. Whilst it is reasonable to call F a curve, it is much too irregular to have tangents in the classical sense. A simple calculation shows that E_k is of length $\left(\frac{4}{3}\right)^k$; letting k tend to infinity implies that F has infinite length. On the other hand, F occupies zero area in the plane, so neither length nor area provides a very useful description of the size of F.

Many other sets may be constructed using such recursive procedures. For example, the *Sierpiński triangle* or *gasket* is obtained by repeatedly removing (inverted) equilateral triangles from an initial equilateral triangle of unit side-length; see figure 0.3. (For many purposes, it is better to think of this procedure as repeatedly replacing an equilateral triangle by three triangles of half the height.) A plane analogue of the Cantor set, a 'Cantor dust', is illustrated in figure 0.4. At each stage each remaining square is divided into 16 smaller squares of which four are kept and the rest discarded. (Of course, other arrangements or numbers of squares could be used to get different sets.) It should be clear that such examples have properties similar to those mentioned in connection with the Cantor set and the von Koch curve. The example depicted in figure 0.5 is constructed using two different similarity ratios.

There are many other types of construction, some of which will be discussed in detail later in the book, that also lead to sets with these sorts of properties.

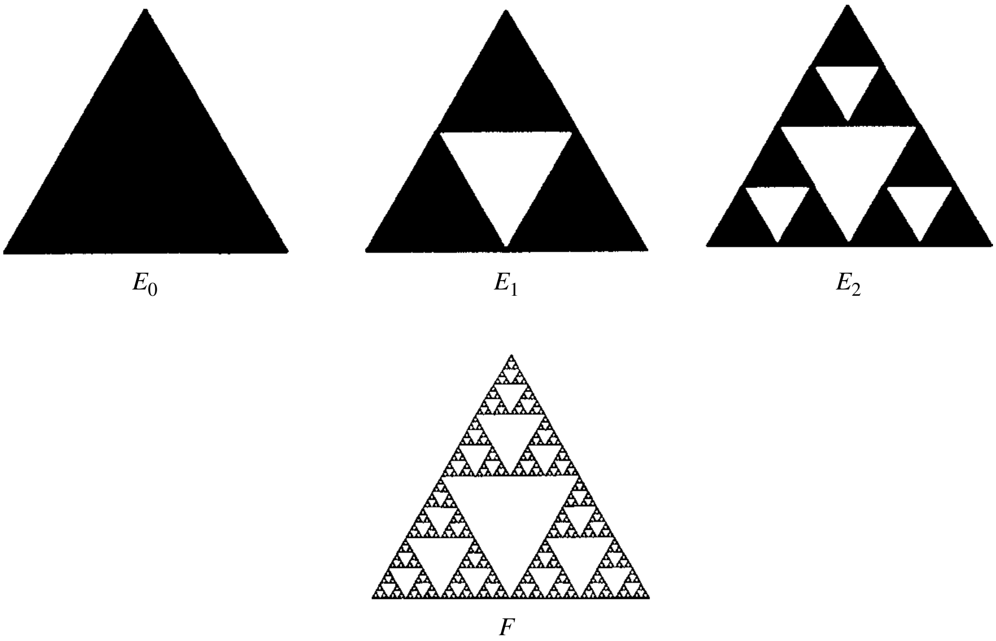


Figure 0.3 Construction of the Sierpiński triangle ($\dim_{\mathrm{H}} F = \dim_{\mathrm{B}} F = \log 3/\log 2$)

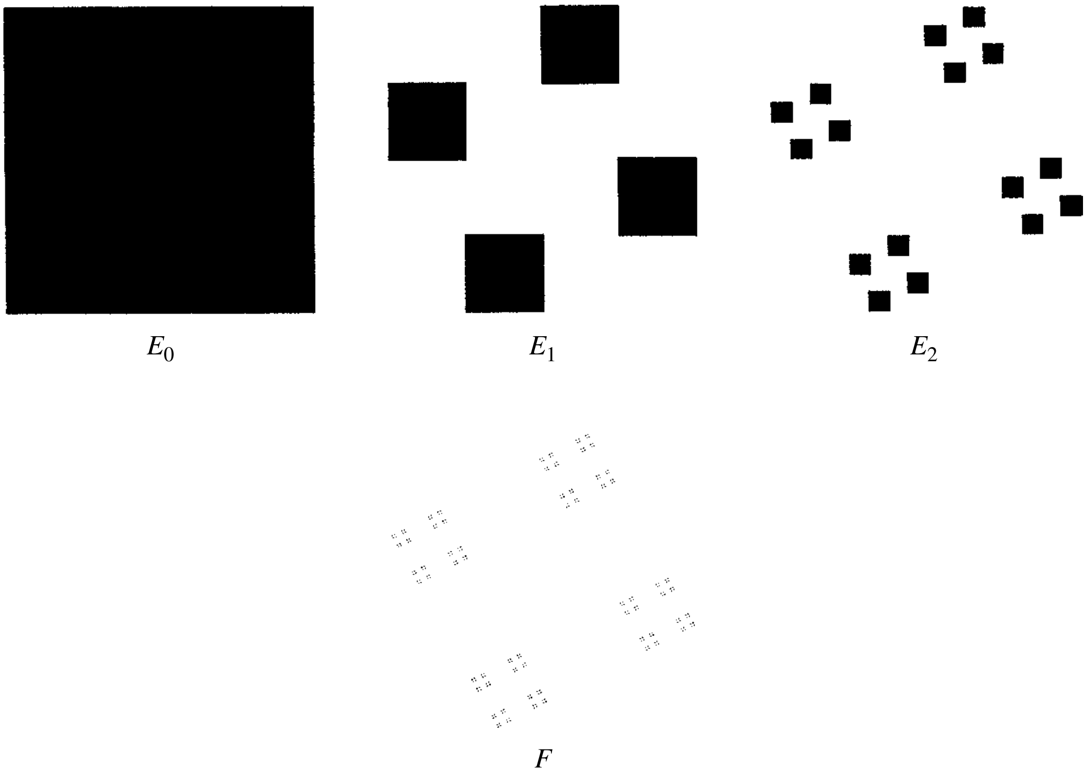


Figure 0.4 Construction of a 'Cantor dust' ($\dim_{\rm H} F = \dim_{\rm B} F = 1$)

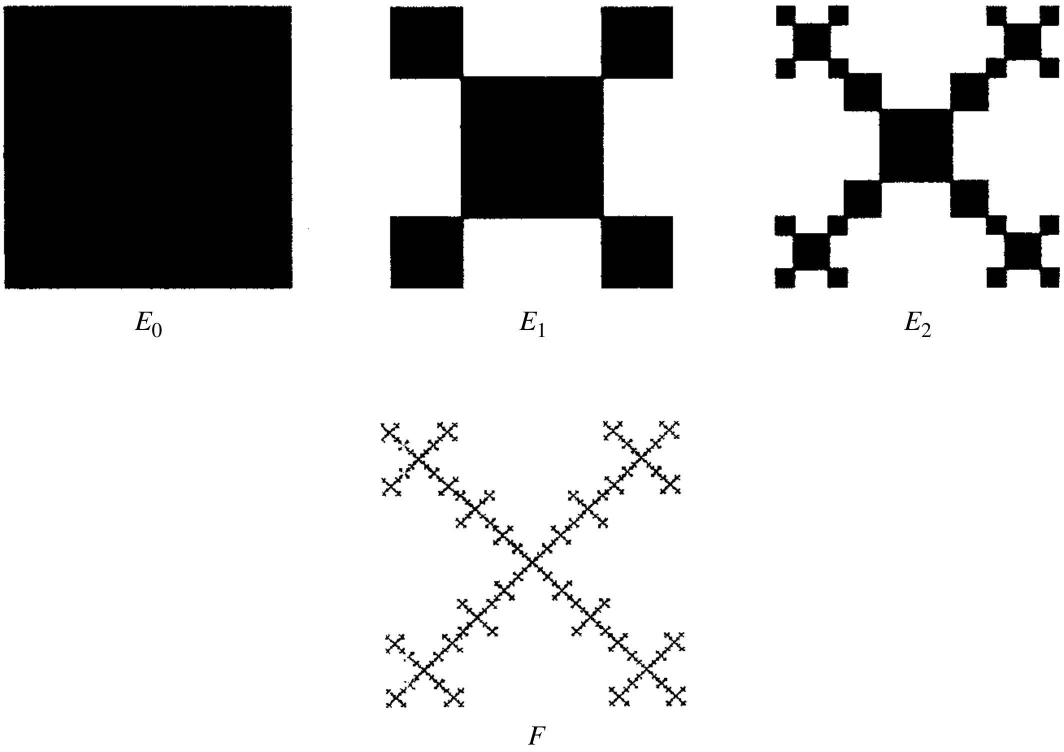


Figure 0.5 Construction of a self-similar fractal with two different similarity ratios

The highly intricate structure of the Julia set illustrated in figure 0.6 stems from the single quadratic function $f(z) = z^2 + c$ for a suitable constant c. Although the set is not strictly self-similar in the sense that the Cantor set and von Koch curve are, it is 'quasi-self-similar' in that arbitrarily small portions of the set can be magnified and then distorted smoothly to coincide with a large part of the set.

Figure 0.7 shows the graph of the function $f(t) = \sum_{k=0}^{\infty}(\frac{3}{2})^{-k/2}\sin((\frac{3}{2})^k t)$; the infinite summation leads to the graph having a fine structure, rather than being a smooth curve to which classical calculus is applicable.

Some of these constructions may be 'randomized'. Figure 0.8 shows a 'random von Koch curve'—a coin was tossed at each step in the construction to determine on which side of the curve to place the new pair of line segments. This random curve certainly has a fine structure, but the strict self-similarity of the von Koch curve has been replaced by a 'statistical self-similarity'.

These are all examples of sets that are commonly referred to as *fractals*. (The word 'fractal' was coined by Mandelbrot in his fundamental essay from the Latin *fractus*, meaning broken, to describe objects that were too irregular to fit into a traditional geometrical setting.) Properties such as those listed for the Cantor set are characteristic of fractals, and it is sets with such properties that we will have in mind throughout the book. Certainly, any fractal worthy of the name will have a fine structure, i.e. detail at all scales. Many fractals have some degree of self-similarity—they are made up of parts that resemble the whole in some way. Sometimes, the resemblance may be weaker than strict geometrical similarity; for example, the similarity may be approximate or statistical.

Methods of classical geometry and calculus are unsuited to studying fractals and we need alternative techniques. The main tool of fractal geometry is dimension in its many forms. We are familiar enough with the idea that a

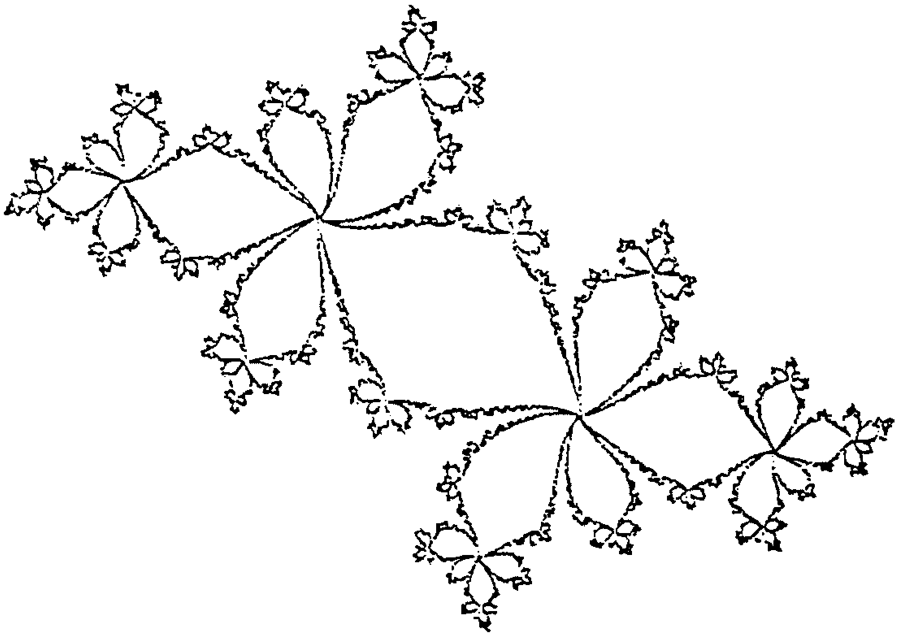


Figure 0.6 A Julia set

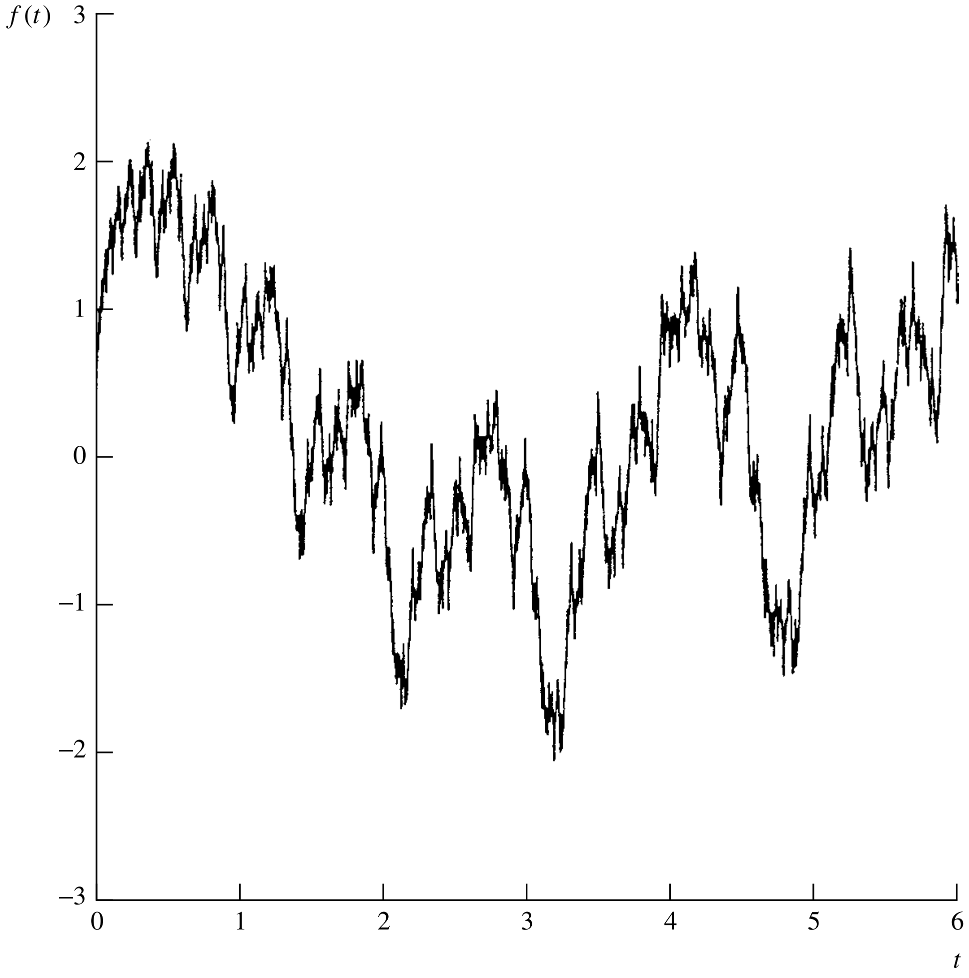


Figure 0.7 Graph of $f(t) = \sum_{k=0}^{\infty}(\frac{3}{2})^{-k/2}\sin((\frac{3}{2})^k t)$

(smooth) curve is a 1-dimensional object and a surface is 2-dimensional. It is less clear that, for many purposes, the Cantor set should be regarded as having dimension $\log 2/\log 3 = 0.631\ldots$ and the von Koch curve as having dimension $\log 4/\log 3 = 1.262\ldots$. This latter number is, at least, consistent with the von Koch curve being 'larger than 1-dimensional' (having infinite length) and 'smaller than 2-dimensional' (having zero area).

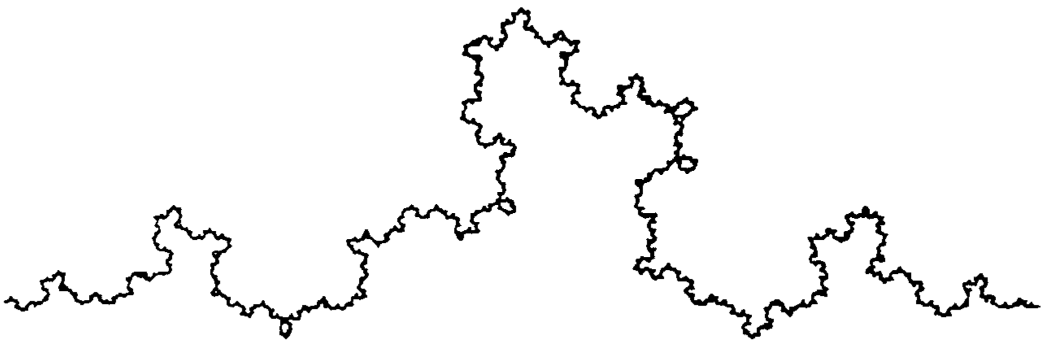


Figure 0.8 A random version of the von Koch curve

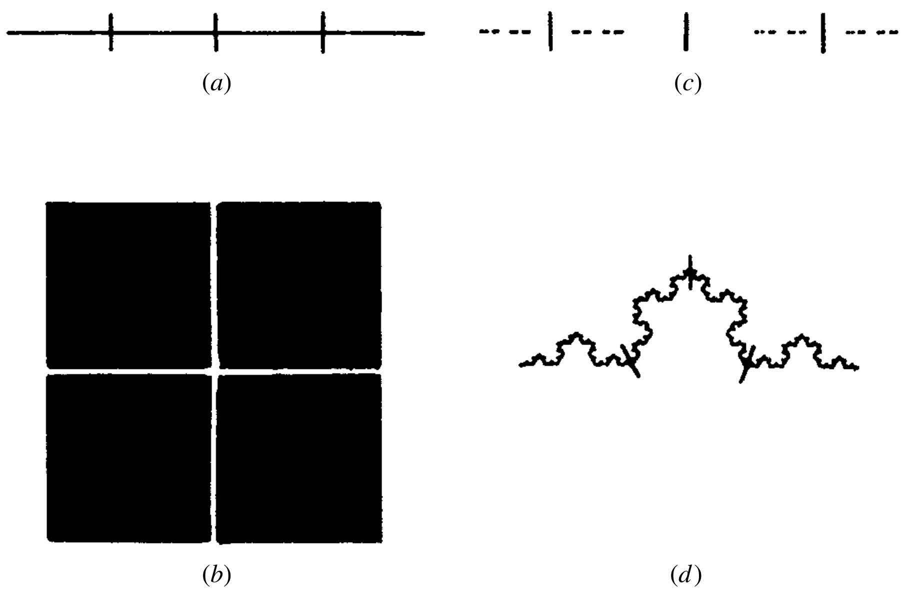


Figure 0.9 Division of certain sets into four parts. The parts are similar to the whole with ratios: $\frac{1}{4}$ for line segment (*a*); $\frac{1}{2}$ for square (*b*); $\frac{1}{9}$ for middle third Cantor set (*c*); $\frac{1}{3}$ for von Koch curve (*d*)

The following argument gives one (rather crude) interpretation of the meaning of these 'dimensions' indicating how they reflect scaling properties and self-similarity. As figure 0.9 indicates, a line segment is made up of four copies of itself, scaled by a factor $\frac{1}{4}$. The segment has dimension $-\log 4/\log\frac{1}{4} = 1$. A square, however, is made up of four copies of itself scaled by a factor $\frac{1}{2}$ (i.e. with half the side length) and has dimension $-\log 4/\log\frac{1}{2} = 2$. In the same way, the von Koch curve is made up of four copies of itself scaled by a factor $\frac{1}{3}$, and has dimension $-\log 4/\log\frac{1}{3} = \log 4/\log 3$, and the Cantor set may be regarded as comprising four copies of itself scaled by a factor $\frac{1}{9}$ and having dimension $-\log 4/\log\frac{1}{9} = \log 2/\log 3$. In general, a set made up of m copies of itself scaled by a factor r might be thought of as having dimension $-\log m/\log r$. The number obtained in this way is usually referred to as the *similarity dimension* of the set.

Unfortunately, similarity dimension is meaningful only for a relatively small class of strictly self-similar sets. Nevertheless, there are other definitions of dimension that are much more widely applicable. For example, Hausdorff dimension and the box-counting dimensions may be defined for any sets, and, in these four examples, may be shown to equal the similarity dimension. The early chapters of the book are concerned with the definition and properties of Hausdorff and other dimensions, along with methods for their calculation. Very roughly, a dimension provides a description of how much space a set fills. It is a measure of the prominence of the irregularities of a set when viewed at very small scales. A dimension contains much information about the geometrical properties of a set.

A word of warning is appropriate at this point. It is possible to define the 'dimension' of a set in many ways, some satisfactory and others less so. It is important to realize that different definitions may give different values of dimension for the same set, and may also have very different properties. Inconsistent usage has sometimes led to considerable confusion. In particular, warning lights flash in my mind (as in the minds of other mathematicians) whenever the term ‘fractal dimension’ is seen. Though some authors attach a precise meaning to this, I have known others interpret it inconsistently in a single piece of work. The reader should always be aware of the definition in use in any discussion.

In his original essay, Mandelbrot defined a fractal to be a set with Hausdorff dimension strictly greater than its topological dimension. (The *topological dimension* of a set is always an integer and is 0 if it is totally disconnected, 1 if each point has arbitrarily small neighbourhoods with boundary of dimension 0, and so on.) This definition proved to be unsatisfactory in that it excluded a number of sets that clearly ought to be regarded as fractals. Various other definitions have been proposed, but they all seem to have this same drawback.

My personal feeling is that the definition of a ‘fractal’ should be regarded in the same way as a biologist regards the definition of ‘life’. There is no hard and fast definition, but just a list of properties characteristic of a living thing, such as the ability to reproduce or to move or to exist to some extent independently of the environment. Most living things have most of the characteristics on the list, though there are living objects that are exceptions to each of them. In the same way, it seems best to regard a fractal as a set that has properties such as those listed below, rather than to look for a precise definition which will almost certainly exclude some interesting cases. From the mathematician’s point of view, this approach is no bad thing. It is difficult to avoid developing properties of dimension other than in a way that applies to ‘fractal’ and ‘non-fractal’ sets alike. For ‘non-fractals’, however, such properties are of little interest—they are generally almost obvious and could be obtained more easily by other methods.

When we refer to a set F as a fractal, therefore, we will typically have the following in mind.

(i) F has a fine structure, i.e. detail on arbitrarily small scales.
(ii) F is too irregular to be described in traditional geometrical language, both locally and globally.
(iii) Often F has some form of self-similarity, perhaps approximate or statistical.
(iv) Usually, the ‘fractal dimension’ of F (defined in some way) is greater than its topological dimension.
(v) In most cases of interest F is defined in a very simple way, perhaps recursively.

What can we say about the geometry of as diverse a class of objects as fractals? Classical geometry gives us a clue. In Part I of this book we study certain analogues of familiar geometrical properties in the fractal situation. The orthogonal projection, or ‘shadow’ of a circle in space onto a plane is, in general, an ellipse. The fractal projection theorems tell us about the ‘shadows’ of a fractal. For many purposes, a tangent provides a good local approximation to a circle.

Though fractals do tend not to have tangents in any sense, it is often possible to say a surprising amount about their local form. Two circles in the plane in 'general position' either intersect in two points or not at all (we regard the case of mutual tangents as 'exceptional'). Using dimension, we can make similar statements about the intersection of fractals. Moving a circle perpendicular to its plane sweeps out a cylinder, with properties that are related to those of the original circle. Similar, and indeed more general, constructions are possible with fractals.

Although classical geometry is of considerable intrinsic interest, it is also called upon widely in other areas of mathematics. For example, circles or parabolae occur as the solution curves of certain differential equations, and a knowledge of the geometrical properties of such curves aids our understanding of the differential equations. In the same way, the general theory of fractal geometry can be applied to the many branches of mathematics in which fractals occur. Various examples of this are given in Part II of the book.

Historically, interest in geometry has been stimulated by its applications to nature. The ellipse assumed importance as the shape of planetary orbits, as did the sphere as the shape of the earth. The geometry of the ellipse and sphere can be applied to these physical situations. Of course, orbits are not quite elliptical, and the earth is not actually spherical, but for many purposes, such as the prediction of planetary motion or the study of the earth's gravitational field, these approximations may be perfectly adequate.

A similar situation pertains with fractals. A glance at the recent physics literature shows the variety of natural objects that are described as fractals—cloud boundaries, topographical surfaces, coastlines, turbulence in fluids, and so on. None of these are actual fractals—their fractal features disappear if they are viewed at sufficiently small scales. Nevertheless, over certain ranges of scale they appear very much like fractals, and at such scales may usefully be regarded as such. The distinction between 'natural fractals' and the mathematical 'fractal sets' that might be used to describe them was emphasized in Mandelbrot's original essay, but this distinction seems to have become somewhat blurred. There are no true fractals in nature. (There are no true straight lines or circles either!)

If the mathematics of fractal geometry is to be really worthwhile, then it should be applicable to physical situations. Considerable progress is being made in this direction and some examples are given towards the end of this book. Although there are natural phenomena that have been explained in terms of fractal mathematics (Brownian motion is a good example), many applications tend to be descriptive rather than predictive. Much of the basic mathematics used in the study of fractals is not particularly new, though much recent mathematics has been specifically geared to fractals. For further progress to be made, development and application of appropriate mathematics remain a high priority.

Notes and references

Unlike the rest of the book, which consists of fairly solid mathematics, this introduction contains some of the author's opinions and prejudices, which may well not be shared by other workers on fractals. *Caveat emptor*!

The foundational treatise on fractals, which may be appreciated at many levels, is the scientific, philosophical and pictorial essay of Mandelbrot (1982) (developed from the original 1975 version), containing a great diversity of natural and mathematical examples. This essay has been the inspiration for much of the work that has been done on fractals.

Many other books have been written on diverse aspects of fractals, and these are cited at the end of the appropriate chapters. Here we mention a selection with a broad coverage. Introductory treatments include Schroeder (1991), Moon (1992), Kaye (1994), Addison (1997) and Lesmoir-Gordon, Rood and Edney (2000). The volume by Peitgen, Jürgens and Saupe (1992) is profusely illustrated with diagrams and examples, and the essays collated by Frame and Mandelbrot (2002) address the role of fractals in mathematics and science education.

The books by Edgar (1990, 1998), Peitgen, Jürgens and Saupe (1992) and Le Méhauté (1991) provide basic mathematical treatments. Falconer (1985a), Mattila (1995), Federer (1996) and Morgan (2000) concentrate on geometric measure theory, Rogers (1998) addresses the general theory of Hausdorff measures, and Wicks (1991) approaches the subject from the standpoint of non-standard analysis. Books with a computational emphasis include Peitgen and Saupe (1988), Devaney and Keen (1989), Hoggar (1992) and Pickover (1998). The sequel to this book, Falconer (1997), contains more advanced mathematical techniques for studying fractals.

Much of interest may be found in proceedings of conferences on fractal mathematics, for example in the volumes edited by Cherbit (1991), Evertsz, Peitgen and Voss (1995) and Novak (1998, 2000). The proceedings edited by Bandt, Graf and Zähle (1995, 2000) concern fractals and probability, those by Lévy Véhel, Lutton and Tricot (1997), Dekking, Lévy Véhel, Lutton and Tricot (1999) address engineering applications. Mandelbrot's 'Selecta' (1997, 1999, 2002) present a wide range of papers with commentaries which provide a fascinating insight into the development and current state of fractal mathematics and science. Edgar (1993) brings together a collection of classic papers on fractal mathematics.

Papers on fractals appear in many journals; in particular the journal *Fractals* covers a wide range of theory and applications.

Part I
FOUNDATIONS

Fractal Geometry: Mathematical Foundations and Application. Second Edition Kenneth Falconer
© 2003 John Wiley & Sons, Ltd ISBNs: 0-470-84861-8 (HB); 0-470-84862-6 (PB)

Chapter 1 Mathematical background

This chapter reviews some of the basic mathematical ideas and notation that will be used throughout the book. Sections 1.1 on set theory and 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. 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; an understanding of 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, etc., and we will occasionally use the coordinate form $x = (x_1, \dots, x_n)$, $y = (y_1, \dots, y_n)$. Addition and scalar multiplication are defined in the usual manner, so that $x + y = (x_1 + y_1, \dots, x_n + y_n)$ and $\lambda x = (\lambda x_1, \dots, \lambda x_n)$, where λ is a real scalar. We use the usual *Euclidean distance* or *metric* on $\mathbb{R}^n$. So if x, 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, we have the triangle inequality $|x + y| \leqslant |x| + |y|$, the reverse triangle inequality $|x - y| \geqslant \big||x| - |y|\big|$, and the metric triangle inequality $|x - y| \leqslant |x - z| + |z - y|$, 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, etc. 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 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,

Fractal Geometry: Mathematical Foundations and Application. Second Edition Kenneth Falconer
© 2003 John Wiley & Sons, Ltd ISBNs: 0-470-84861-8 (HB); 0-470-84862-6 (PB)

and $\mathbb{Z}^+$ are the positive integers. Occasionally 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 $\{x : a \leqslant x < b\}$, etc.

The *coordinate cube* of side $2r$ and centre $x = (x_1, \dots, x_n)$ is the set $\{y = (y_1, \dots, y_n) : |y_i - x_i| \leqslant r$ for all $i = 1, \dots, n\}$. (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 $A_\delta = \{x : |x - y| \leqslant \delta$ for some y in $A\}$; see figure 1.1.

We write $A \cup B$ for the *union* of the sets A and B, i.e. 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 $\{A_\alpha\}$, i.e. 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 $\{(a, b) : a \in A$ 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 $A \times B \subset \mathbb{R}^{n+m}$.

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 $A + B = \{x + y : x \in A$ 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 $x_1, x_2, \dots$ with every element of A appearing at a specific place in the list; otherwise the

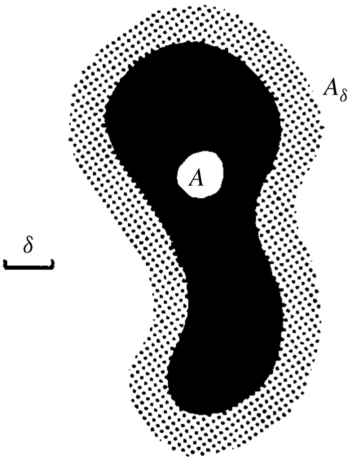


Figure 1.1 A set A and its δ-neighbourhood A_δ

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 the *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, though it is important to realize 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 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 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.

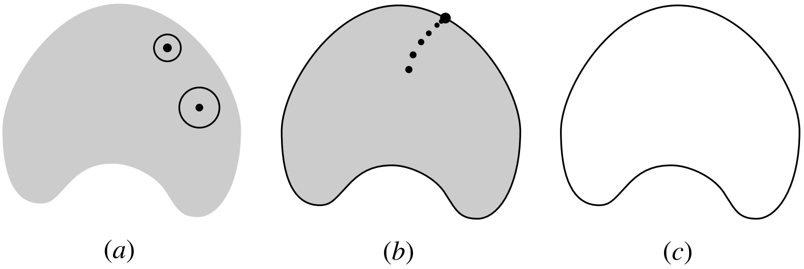


Figure 1.2 (*a*) An open set—there is a ball contained in the set centred at each point of the set. (*b*) A closed set—the limit of any convergent sequence of points from the set lies in the set. (*c*) The boundary of the set in (*a*) or (*b*)

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* subset of A if $B \subset A \subset \overline{B}$, i.e. if there are points of B arbitrarily close to each point of A.

A set A is *compact* if any collection of open sets which 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 though 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:

(a) every open set and every closed set is a Borel set;
(b) 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 in work of the sort described in this book by assuming that all the sets encountered are Borel sets.

1.2 Functions and limits

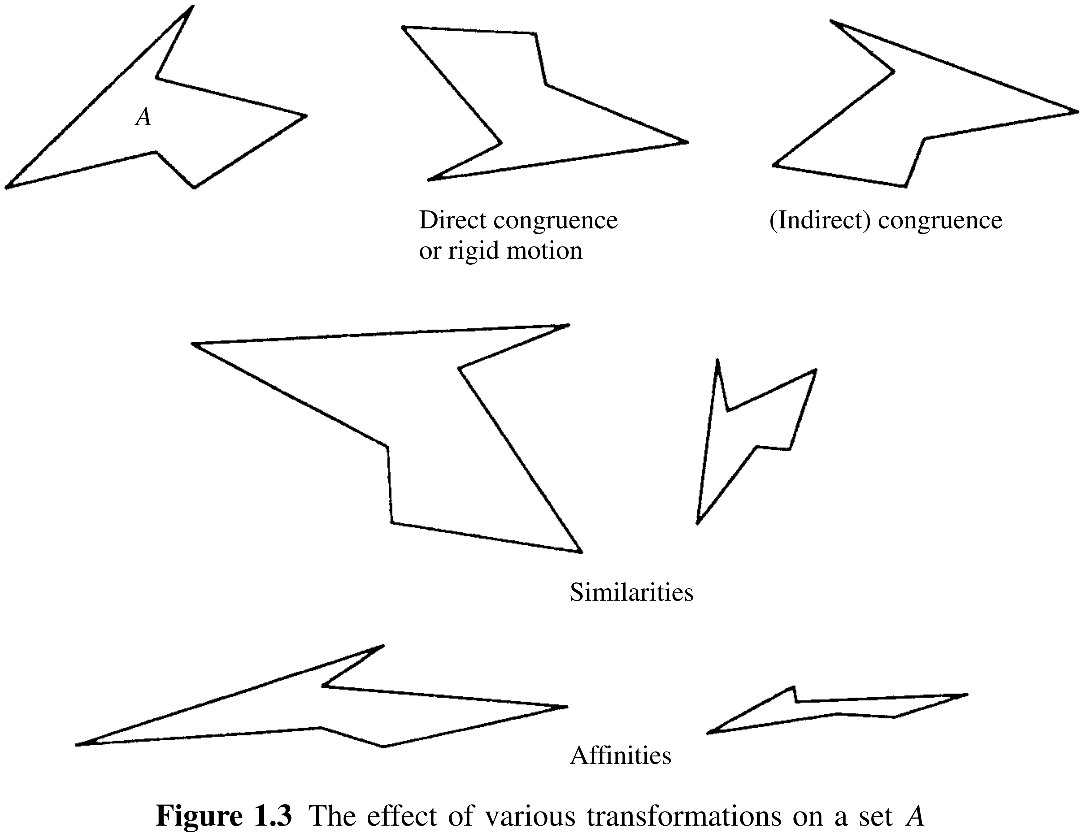
Let X and Y be any sets. A *mapping, function* or *transformation* f from X to Y is a rule or formula that associates a point $f(x)$ of Y with each point x of X.

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, i.e. 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 $x \neq y$, i.e. 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 $f(x) = y$, i.e. 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 x in X and $f(f^{-1}(y)) = y$ for 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, i.e. 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



Direct congruence or rigid motion

(Indirect) congruence

Similarities

Affinities

Figure 1.3 The effect of various transformations on a set A

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 map 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, i.e. 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 point 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. However, if T is orthonormal, then S is a congruence, and if T is a scalar multiple or 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 a *Lipschitz function* 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 a *bi-Lipschitz function* 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.

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

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

Since $\inf\{f(x) : 0 < x < r\}$ 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 in values 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 $x > 0$ 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 often 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

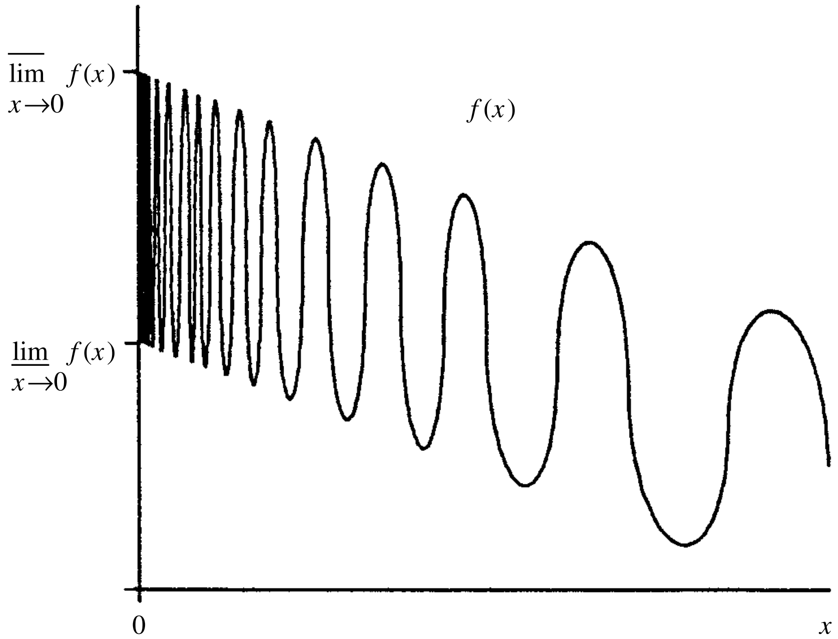


Figure 1.4 The upper and lower limits of a function

have that $f(x) \sim x^s$; in other words that 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 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

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

In particular, the mean value theorem applies: given $a < b$ and f differentiable on $[a, b]$ there exists c with $a < c < b$ such that

$$\frac{f(b) - f(a)}{b - a} = f'(c)$$

(intuitively, any chord of the graph of f is parallel to the slope of f at some intermediate point). A function f is *continuously differentiable* if $f'(x)$ is continuous in x.

More generally, if $f : \mathbb{R}^n \to \mathbb{R}^n$, we say that f is *differentiable* at x with *derivative* 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 $\mathrm{e}^{\log x} = x$, for $x > 0$, and $\log \mathrm{e}^y = y$ 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, since 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), i.e.

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

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 (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 realize 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, since 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,

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

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

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

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

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

More generally, it can be shown that if, for $\delta > 0$, A_δ are Borel sets that are increasing as δ decreases, i.e. $A_{\delta'} \subset A_\delta$ for $0 < \delta < \delta'$, then

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

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 $\mu(\mathbb{R}^n\backslash X) = 0$. The support of a measure is always closed and 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 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 emphasize this intuitive meaning.

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

If $A = \{(x_1, \dots, x_n) \in \mathbb{R}^n : a_i \leqslant x_i \leqslant b_i\}$ is a 'coordinate parallelepiped' in $\mathbb{R}^n$, the n-dimensional volume of A is given by

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

(Of course, vol^1 is length, as in Example 1.3, vol^2 is area and vol^3 is 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

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

where the infimum is taken over all coverings of A by coordinate parallelepipeds A_i. We get that $\mathcal{L}^n(A) = \text{vol}^n(A)$ 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. Define $\mu(A) = \mathcal{L}^1(L \cap A)$ i.e. the 'length' of intersection of A with L. Then μ is a mass distribution with support L, since $\mu(A) = 0$ 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 2.1, are a generalization 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

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

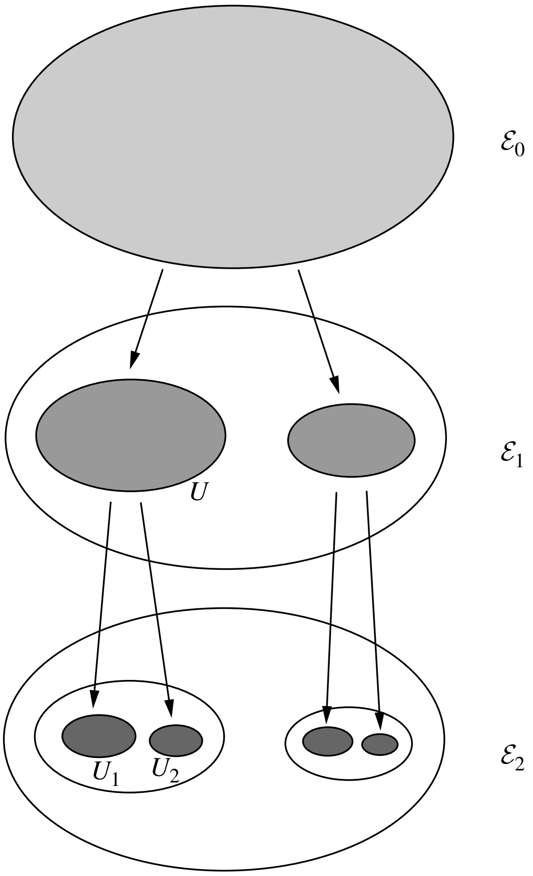


Figure 1.5 Steps in the construction of a mass distribution μ by repeated subdivision. The mass on the sets of $\mathcal{E}_k$ is divided between the sets of $\mathcal{E}_{k+1}$; so, for example, $\mu(U) = \mu(U_1) + \mu(U_2)$

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 $\mathbb{R}^n \backslash E_k$. 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 E to determine $\mu(A)$ for any (Borel) set A. This is indeed the case, as the following proposition states.

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 $\bigcap_{k=1}^{\infty} \overline{E}_k$.

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

$$\mu(A) = \inf\left\{\sum_i \mu(U_i) : A \subset \bigcup_i U_i \text{ and } U_i \in \mathcal{E}\right\}. \tag{1.9}$$

(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; 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. Since $\mu(\mathbb{R}^n \backslash E_k) = 0$, 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

Let $\mathcal{E}_k$ denote the collection of 'binary intervals' of length 2^{-k} 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 $[0, 1]$.

Note on calculation. Clearly, 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; certainly this is true of functions that are likely to be encountered in practice.

To define integration we first suppose that $f : D \to \mathbb{R}$ is a *simple function*, i.e. 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 $\int f_+ \,\mathrm{d}\mu$ and $\int f_- \,\mathrm{d}\mu$ are both 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. We also have the monotone convergence theorem, that 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' such that $\chi_A(x) = 1$ 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, i.e. 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

For an understanding of some of the later chapters of the book, a basic knowledge of probability theory is necessary. We give a brief survey 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 corresponding to the Borel sets.

In our applications later on in the book, we shall be particularly interested in events (on rather large sample spaces) 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, i.e. 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, i.e. $\mathsf{P}(\{\omega : X(\omega) \in E\})$, 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, for example $(X + Y)(\omega) = X(\omega) + Y(\omega)$ 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 $\{(X \in E_k)\}$ 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 both $\mathsf{E}(X_+) < \infty$ and $\mathsf{E}(X_-) < \infty$.

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

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

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

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

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

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

for any real number λ, and

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

provided that X and Y are independent.

If the probability distribution of a random variable is given by an integral, i.e.

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

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

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

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

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

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 $(1/k)S_k$ is the average of the first k trials. The *strong law of large numbers* states that, with probability 1,

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

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,

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

An important aspect of the normal distribution now becomes clear—it is the form of distribution approached by sums of a large number of independent identically distributed random variables.

We may apply these results 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 $(\frac{1}{6})^{-k}$. 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 mathematical texts. Almost any book on mathematical analysis, for example Rudin (1964) or Apostol (1974), contains the basic theory of sets and functions. A thorough treatment of measure and probability theory may be found in Kingman and Taylor (1966), Billingsley (1995) and Edgar (1998). For probability theory, the book by Grimmett and Stirzaker (1992) may be found helpful.

Exercises

The following exercises do no more than emphasize 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) $|x-y| \geqslant \Big| |x| - |y| \Big|$, (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 sets 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 $\{0, 1, \frac{1}{2}, \frac{1}{3}, \frac{1}{4}, \ldots\}$.

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 $\{x \in \mathbb{R} : 0 \leqslant x < 1\}$ 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) $\sin(1/x)$; (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) $f^{-1}(-2)$, (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), (1.2) and (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 $N_k(\omega)$ to equal the number of i with $1 \leqslant i \leqslant k$ for which $\omega \in A_i$. 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.)

Chapter 2 Hausdorff measure and dimension

The notion of dimension is central to fractal geometry. Roughly, dimension indicates how much space a set occupies near to each of its points. Of the wide variety of ‘fractal dimensions’ in use, the definition of Hausdorff, based on a construction of Carathéodory, is the oldest and probably the most important. Hausdorff dimension has the advantage of being defined for any set, and is mathematically convenient, as it is based on measures, which are relatively easy to manipulate. A major disadvantage is that in many cases it is hard to calculate or to estimate by computational methods. However, for an understanding of the mathematics of fractals, familiarity with Hausdorff measure and dimension is essential.

2.1 Hausdorff measure

Recall that if U is any non-empty subset of n-dimensional Euclidean space, $\mathbb{R}^n$, the *diameter* of U is defined as $|U| = \sup\{|x-y| : x, y \in U\}$, i.e. the greatest distance apart of any pair of points in U. If $\{U_i\}$ is a countable (or finite) collection of sets of diameter at most δ that cover F, i.e. $F \subset \bigcup_{i=1}^{\infty} U_i$ with $0 \leqslant |U_i| \leqslant \delta$ for each i, we say that $\{U_i\}$ is a *δ-cover* of F.

Suppose that F is a subset of $\mathbb{R}^n$ and s is a non-negative number. For any $\delta > 0$ we define

$$\mathcal{H}^s_\delta(F) = \inf\left\{\sum_{i=1}^{\infty} |U_i|^s : \{U_i\} \text{ is a } \delta\text{-cover of } F\right\}. \tag{2.1}$$

Thus we look at all covers of F by sets of diameter at most δ and seek to minimize the sum of the sth powers of the diameters (figure 2.1). As δ decreases, the class of permissible covers of F in (2.1) is reduced. Therefore, the infimum $\mathcal{H}^s_\delta(F)$ increases, and so approaches a limit as $\delta \to 0$. We write

$$\mathcal{H}^s(F) = \lim_{\delta\to 0} \mathcal{H}^s_\delta(F). \tag{2.2}$$

Fractal Geometry: Mathematical Foundations and Application. Second Edition Kenneth Falconer
© 2003 John Wiley & Sons, Ltd ISBNs: 0-470-84861-8 (HB); 0-470-84862-6 (PB)

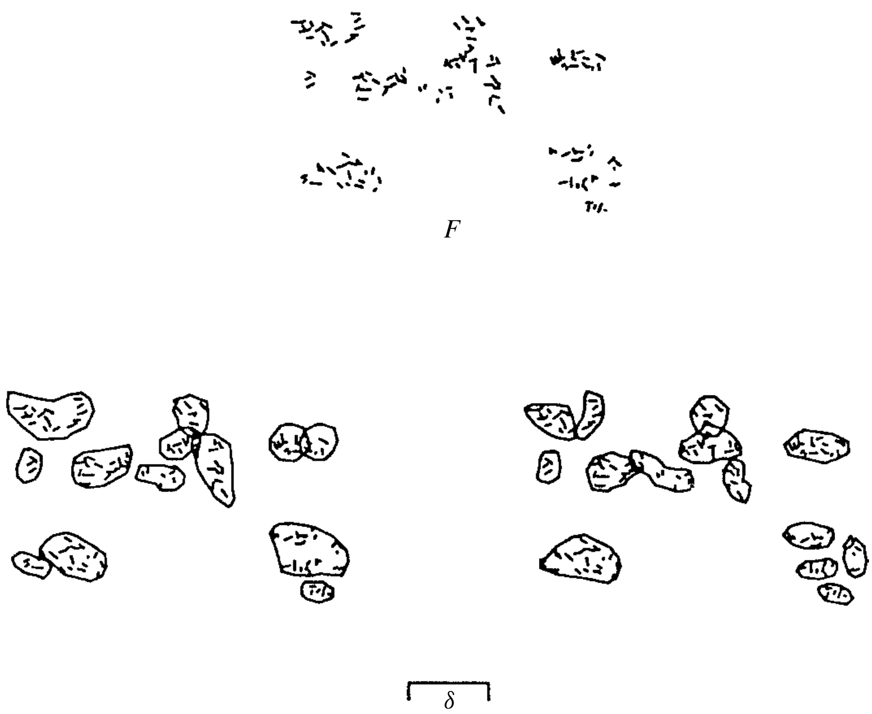


Figure 2.1 A set F and two possible δ-covers for F. The infimum of $\Sigma|U_i|^s$ over all such δ-covers $\{U_i\}$ gives $\mathcal{H}^s_\delta(F)$

This limit exists for any subset F of $\mathbb{R}^n$, though the limiting value can be (and usually is) 0 or ∞. We call $\mathcal{H}^s(F)$ the *s-dimensional Hausdorff measure* of F.

With a certain amount of effort, $\mathcal{H}^s$ may be shown to be a measure; see section 1.3. It is straightforward to show that $\mathcal{H}^s(\emptyset) = 0$, that if E is contained in F then $\mathcal{H}^s(E) \leqslant \mathcal{H}^s(F)$, and that if $\{F_i\}$ is any countable collection of sets, then

$$\mathcal{H}^s\left(\bigcup_{i=1}^{\infty} F_i\right) \leqslant \sum_{i=1}^{\infty} \mathcal{H}^s(F_i). \tag{2.3}$$

It is rather harder to show that there is equality in (2.3) if the $\{F_i\}$ are disjoint Borel sets.

Hausdorff measures generalize the familiar ideas of length, area, volume, etc. It may be shown that, for subsets of $\mathbb{R}^n$, n-dimensional Hausdorff measure is, to within a constant multiple, just n-dimensional Lebesgue measure, i.e. the usual n-dimensional volume. More precisely, if F is a Borel subset of $\mathbb{R}^n$, then

$$\mathcal{H}^n(F) = c_n^{-1}\mathrm{vol}^n(F) \tag{2.4}$$

where c_n is the volume of an n-dimensional ball of diameter 1, so that $c_n = \pi^{n/2}/2^n(n/2)!$ if n is even and $c_n = \pi^{(n-1)/2}((n-1)/2)!/n!$ if n is odd. Similarly,

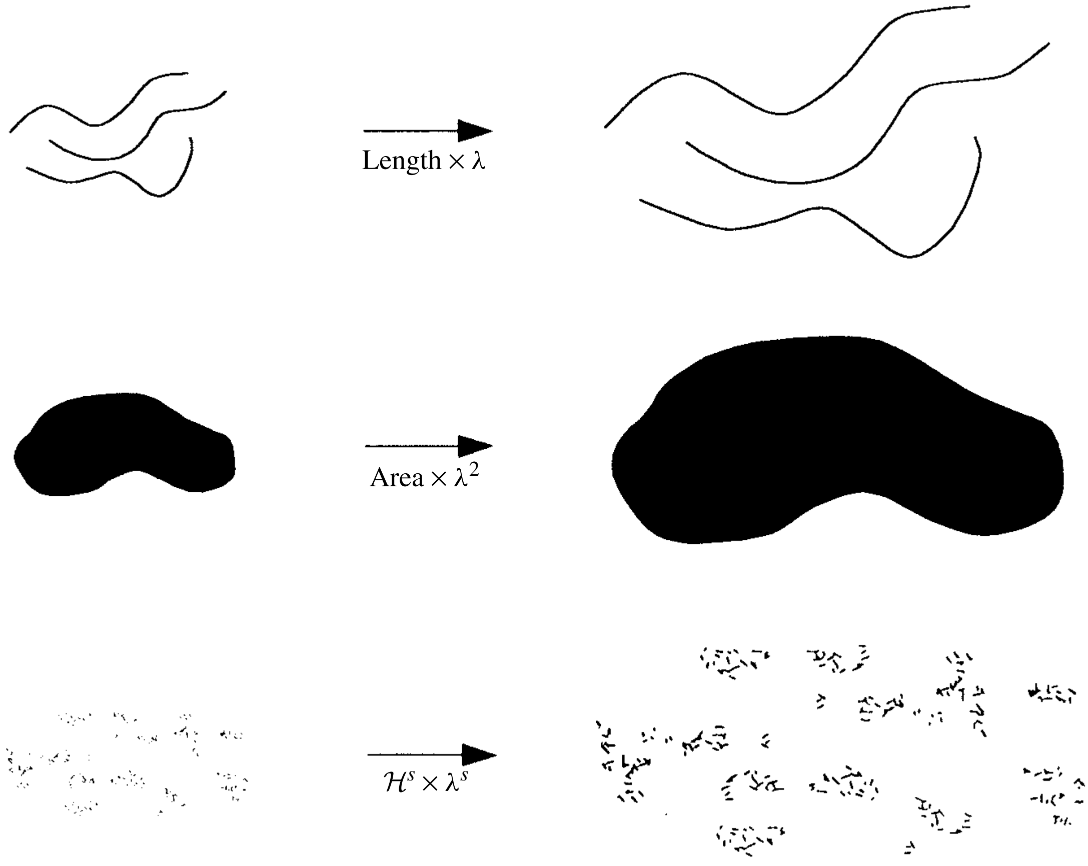


Figure 2.2 Scaling sets by a factor λ increases length by a factor λ, area by a factor λ^2, and s-dimensional Hausdorff measure by a factor λ^s

for 'nice' lower-dimensional subsets of $\mathbb{R}^n$, we have that $\mathcal{H}^0(F)$ is the number of points in F; $\mathcal{H}^1(F)$ gives the length of a smooth curve F; $\mathcal{H}^2(F) = (4/\pi) \times \text{area}(F)$ if F is a smooth surface; $\mathcal{H}^3(F) = (6/\pi) \times \text{vol}(F)$; and $\mathcal{H}^m(F) = c_m^{-1} \times \text{vol}^m(F)$ if F is a smooth m-dimensional submanifold of $\mathbb{R}^n$ (i.e. an m-dimensional surface in the classical sense).

The scaling properties of length, area and volume are well known. On magnification by a factor λ, the length of a curve is multiplied by λ, the area of a plane region is multiplied by λ^2 and the volume of a 3-dimensional object is multiplied by λ^3. As might be anticipated, s-dimensional Hausdorff measure scales with a factor λ^s (figure 2.2). Such scaling properties are fundamental to the theory of fractals.

Scaling property 2.1

Let S be a similarity transformation of scale factor $\lambda > 0$. If $F \subset \mathbb{R}^n$, then

$$\mathcal{H}^s(S(F)) = \lambda^s \mathcal{H}^s(F). \tag{2.5}$$

Proof. If $\{U_i\}$ is a δ-cover of F then $\{S(U_i)\}$ is a $\lambda\delta$-cover of $S(F)$, so

$$\Sigma|S(U_i)|^s = \lambda^s \Sigma |U_i|^s$$

so

$$\mathcal{H}^s_{\lambda\delta}(S(F)) \leqslant \lambda^s \mathcal{H}^s_\delta(F)$$

on taking the infimum. Letting $\delta \to 0$ gives that $\mathcal{H}^s(S(F)) \leqslant \lambda^s \mathcal{H}^s(F)$. Replacing S by S^{-1}, and so λ by $1/\lambda$, and F by $S(F)$ gives the opposite inequality required. □

A similar argument gives the following basic estimate of the effect of more general transformations on the Hausdorff measures of sets.

Proposition 2.2

Let $F \subset \mathbb{R}^n$ and $f : F \to \mathbb{R}^m$ be a mapping such that

$$|f(x) - f(y)| \leqslant c|x-y|^\alpha \quad (x, y \in F) \tag{2.6}$$

for constants $c > 0$ and $\alpha > 0$. Then for each s

$$\mathcal{H}^{s/\alpha}(f(F)) \leqslant c^{s/\alpha} \mathcal{H}^s(F). \tag{2.7}$$

Proof. If $\{U_i\}$ is a δ-cover of F, then, since $|f(F \cap U_i)| \leqslant c|F \cap U_i|^\alpha \leqslant c|U_i|^\alpha$, it follows that $\{f(F \cap U_i)\}$ is an ε-cover of $f(F)$, where $\varepsilon = c\delta^\alpha$. Thus $\sum_i |f(F \cap U_i)|^{s/\alpha} \leqslant c^{s/\alpha} \sum_i |U_i|^s$, so that $\mathcal{H}^{s/\alpha}_\varepsilon(f(F)) \leqslant c^{s/\alpha} \mathcal{H}^s_\delta(F)$. As $\delta \to 0$, so $\varepsilon \to 0$, giving (2.7). □

Condition (2.6) is known as a *Hölder condition of exponent α*; such a condition implies that f is continuous. Particularly important is the case $\alpha = 1$, i.e.

$$|f(x) - f(y)| \leqslant c|x-y| \quad (x, y \in F) \tag{2.8}$$

when f is called a *Lipschitz mapping*, and

$$\mathcal{H}^s(f(F)) \leqslant c^s \mathcal{H}^s(F). \tag{2.9}$$

In particular (2.9) holds for any differentiable function with bounded derivative; such a function is necessarily Lipschitz as a consequence of the mean value theorem. If f is an isometry, i.e. $|f(x) - f(y)| = |x-y|$, then $\mathcal{H}^s(f(F)) = \mathcal{H}^s(F)$. Thus, Hausdorff measures are translation invariant (i.e. $\mathcal{H}^s(F+z) = \mathcal{H}^s(F)$, where $F + z = \{x + z : x \in F\}$), and rotation invariant, as would certainly be expected.

2.2 Hausdorff dimension

Returning to equation (2.1) it is clear that for any given set $F \subset \mathbb{R}^n$ and $\delta < 1$, $\mathcal{H}^s_\delta(F)$ is non-increasing with s, so by (2.2) $\mathcal{H}^s(F)$ is also non-increasing. In fact, rather more is true: if $t > s$ and $\{U_i\}$ is a δ-cover of F we have

$$\sum_i |U_i|^t \leqslant \sum_i |U_i|^{t-s}|U_i|^s \leqslant \delta^{t-s}\sum_i |U_i|^s \tag{2.10}$$

so, taking infima, $\mathcal{H}^t_\delta(F) \leqslant \delta^{t-s}\mathcal{H}^s_\delta(F)$. Letting $\delta \to 0$ we see that if $\mathcal{H}^s(F) < \infty$ then $\mathcal{H}^t(F) = 0$ for $t > s$. Thus a graph of $\mathcal{H}^s(F)$ against s (figure 2.3) shows that there is a critical value of s at which $\mathcal{H}^s(F)$ 'jumps' from ∞ to 0. This critical value is called the *Hausdorff dimension* of F, and written $\dim_{\rm H} F$; it is defined for *any* set $F \subset \mathbb{R}^n$. (Note that some authors refer to Hausdorff dimension as *Hausdorff–Besicovitch dimension*.) Formally

$$\dim_{\rm H} F = \inf\{s \geqslant 0 : \mathcal{H}^s(F) = 0\} = \sup\{s : \mathcal{H}^s(F) = \infty\} \tag{2.11}$$

(taking the supremum of the empty set to be 0), so that

$$\mathcal{H}^s(F) = \begin{cases} \infty & \text{if } 0 \leqslant s < \dim_{\rm H} F \\ 0 & \text{if } s > \dim_{\rm H} F. \end{cases} \tag{2.12}$$

If $s = \dim_{\rm H} F$, then $\mathcal{H}^s(F)$ may be zero or infinite, or may satisfy

$$0 < \mathcal{H}^s(F) < \infty.$$

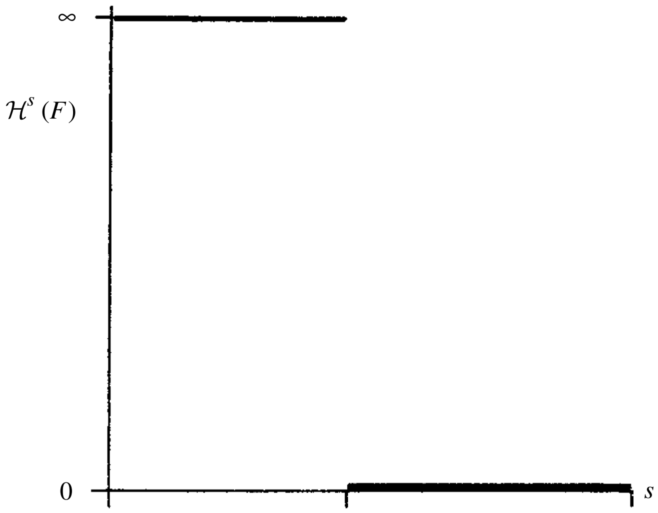


Figure 2.3 Graph of $\mathcal{H}^s(F)$ against s for a set F. The Hausdorff dimension is the value of s at which the 'jump' from ∞ to 0 occurs

A Borel set satisfying this last condition is called an *s-set*. Mathematically, s-sets are by far the most convenient sets to study, and fortunately they occur surprisingly often.

For a very simple example, let F be a flat disc of unit radius in $\mathbb{R}^3$. From familiar properties of length, area and volume, $\mathcal{H}^1(F) = \text{length}\,(F) = \infty$, $0 < \mathcal{H}^2(F) = (4/\pi) \times \text{area}\,(F) = 4 < \infty$ and $\mathcal{H}^3(F) = (6/\pi) \times \text{vol}(F) = 0$. Thus $\dim_{\text{H}} F = 2$, with $\mathcal{H}^s(F) = \infty$ if $s < 2$ and $\mathcal{H}^s(F) = 0$ if $s > 2$.

Hausdorff dimension satisfies the following properties (which might well be expected to hold for any reasonable definition of dimension).

Monotonicity. If $E \subset F$ then $\dim_{\text{H}} E \leqslant \dim_{\text{H}} F$. This is immediate from the measure property that $\mathcal{H}^s(E) \leqslant \mathcal{H}^s(F)$ for each s.

Countable stability. If $F_1, F_2, \ldots$ is a (countable) sequence of sets then $\dim_{\text{H}} \bigcup_{i=1}^{\infty} F_i = \sup_{1 \leqslant i < \infty}\{\dim_{\text{H}} F_i\}$. Certainly, $\dim_{\text{H}} \bigcup_{i=1}^{\infty} F_i \geqslant \dim_{\text{H}} F_j$ for each j from the monotonicity property. On the other hand, if $s > \dim_{\text{H}} F_i$ for all i, then $\mathcal{H}^s(F_i) = 0$, so that $\mathcal{H}^s(\bigcup_{i=1}^{\infty} F_i) = 0$, giving the opposite inequality.

Countable sets. If F is countable then $\dim_{\text{H}} F = 0$. For if F_i is a single point, $\mathcal{H}^0(F_i) = 1$ and $\dim_{\text{H}} F_i = 0$, so by countable stability $\dim_{\text{H}} \bigcup_{i=1}^{\infty} F_i = 0$.

Open sets. If $F \subset \mathbb{R}^n$ is open, then $\dim_{\text{H}} F = n$. For since F contains a ball of positive n-dimensional volume, $\dim_{\text{H}} F \geqslant n$, but since F is contained in countably many balls, $\dim_{\text{H}} F \leqslant n$ using countable stability and monotonicity.

Smooth sets. If F is a smooth (i.e. continuously differentiable) m-dimensional submanifold (i.e. m-dimensional surface) of $\mathbb{R}^n$ then $\dim_{\text{H}} F = m$. In particular smooth curves have dimension 1 and smooth surfaces have dimension 2. Essentially, this may be deduced from the relationship between Hausdorff and Lebesgue measures, see also Exercise 2.7.

The transformation properties of Hausdorff dimension follow immediately from the corresponding ones for Hausdorff measures given in Proposition 2.2.

Proposition 2.3

Let $F \subset \mathbb{R}^n$ *and suppose that* $f : F \to \mathbb{R}^m$ *satisfies a Hölder condition*

$$|f(x) - f(y)| \leqslant c|x - y|^{\alpha} \qquad (x, y \in F).$$

Then $\dim_{\text{H}} f(F) \leqslant (1/\alpha)\dim_{\text{H}} F$.

Proof. If $s > \dim_{\text{H}} F$ then by Proposition 2.2 $\mathcal{H}^{s/\alpha}(f(F)) \leqslant c^{s/\alpha}\mathcal{H}^s(F) = 0$, implying that $\dim_{\text{H}} f(F) \leqslant s/\alpha$ for all $s > \dim_{\text{H}} F$. □

Corollary 2.4

(*a*) *If* $f : F \to \mathbb{R}^m$ *is a Lipschitz transformation* (*see* (2.8)) *then* $\dim_{\text{H}} f(F) \leqslant \dim_{\text{H}} F$.

(*b*) *If* $f : F \to \mathbb{R}^m$ *is a bi-Lipschitz transformation, i.e.*

$$c_1|x-y| \leqslant |f(x)-f(y)| \leqslant c_2|x-y| \qquad (x, y \in F) \qquad (2.13)$$

where $0 < c_1 \leqslant c_2 < \infty$, *then* $\dim_{\mathrm{H}} f(F) = \dim_{\mathrm{H}} F$.

Proof. Part (a) follows from Proposition 2.3 taking $\alpha = 1$. Applying this to $f^{-1} : f(F) \to F$ gives the other inequality required for (b). □

This corollary reveals a fundamental property of Hausdorff dimension: *Hausdorff dimension is invariant under bi-Lipschitz transformations*. Thus if two sets have different dimensions there cannot be a bi-Lipschitz mapping from one onto the other. This is reminiscent of the situation in topology where various 'invariants' (such as homotopy or homology groups) are set up to distinguish between sets that are not homeomorphic: if the topological invariants of two sets differ then there cannot be a homeomorphism (continuous one-to-one mapping with continuous inverse) between the two sets.

In topology two sets are regarded as 'the same' if there is a homeomorphism between them. One approach to fractal geometry is to regard two sets as 'the same' if there is a bi-Lipschitz mapping between them. Just as topological invariants are used to distinguish between non-homeomorphic sets, we may seek parameters, including dimension, to distinguish between sets that are not bi-Lipschitz equivalent. Since bi-Lipschitz transformations (2.13) are necessarily homeomorphisms, topological parameters provide a start in this direction, and Hausdorff dimension (and other definitions of dimension) provide further distinguishing characteristics between fractals.

In general, the dimension of a set alone tells us little about its topological properties. However, any set of dimension less than 1 is necessarily so sparse as to be totally disconnected; that is, no two of its points lie in the same connected component.

Proposition 2.5

A set $F \subset \mathbb{R}^n$ *with* $\dim_{\mathrm{H}} F < 1$ *is totally disconnected.*

Proof. Let x and y be distinct points of F. Define a mapping $f : \mathbb{R}^n \to [0, \infty)$ by $f(z) = |z-x|$. Since f does not increase distances, as $|f(z)-f(w)| = \big||z-x| - |w-x|\big| \leqslant |(z-x)-(w-x)| = |z-w|$, we have from Corollary 2.4(a) that $\dim_{\mathrm{H}} f(F) \leqslant \dim_{\mathrm{H}} F < 1$. Thus $f(F)$ is a subset of $\mathbb{R}$ of $\mathcal{H}^1$-measure or length zero, and so has a dense complement. Choosing r with $r \notin f(F)$ and $0 < r < f(y)$ it follows that

$$F = \{z \in F : |z-x| < r\} \cup \{z \in F : |z-x| > r\}.$$

Thus F is contained in two disjoint open sets with x in one set and y in the other, so that x and y lie in different connected components of F. □

2.3 Calculation of Hausdorff dimension—simple examples

This section indicates how to calculate the Hausdorff dimension of some simple fractals such as some of those mentioned in the Introduction. Other methods will be encountered throughout the book. It is important to note that most dimension calculations involve an upper estimate and a lower estimate, which are hopefully equal. Each of these estimates usually involves a geometric observation followed by a calculation.

Example 2.6

Let F be the Cantor dust constructed from the unit square as in figure 0.4. (*At each stage of the construction the squares are divided into* 16 *squares with a quarter of the side length, of which the same pattern of four squares is retained.*) *Then* $1 \leqslant \mathcal{H}^1(F) \leqslant \sqrt{2}$, *so* $\dim_{\mathrm{H}} F = 1$.

Calculation. Observe that E_k, the kth stage of the construction, consists of 4^k squares of side 4^{-k} and thus of diameter $4^{-k}\sqrt{2}$. Taking the squares of E_k as a δ-cover of F where $\delta = 4^{-k}\sqrt{2}$, we get an estimate $\mathcal{H}^1_\delta(F) \leqslant 4^k 4^{-k}\sqrt{2}$ for the infimum in (2.1). As $k \to \infty$ so $\delta \to 0$ giving $\mathcal{H}^1(F) \leqslant \sqrt{2}$.

For the lower estimate, let proj denote orthogonal projection onto the x-axis. Orthogonal projection does not increase distances, i.e. $|\mathrm{proj}\, x - \mathrm{proj}\, y| \leqslant |x - y|$ if $x, y \in \mathbb{R}^2$, so proj is a Lipschitz mapping. By virtue of the construction of F, the projection or 'shadow' of F on the x-axis, proj F, is the unit interval $[0, 1]$. Using (2.9)

$$1 = \text{length}\,[0, 1] = \mathcal{H}^1([0, 1]) = \mathcal{H}^1(\mathrm{proj}\, F) \leqslant \mathcal{H}^1(F). \qquad \square$$

Note that the same argument and result hold for a set obtained by repeated division of squares into m^2 squares of side length $1/m$ of which one square in each column is retained.

This trick of using orthogonal projection to get a lower estimate of Hausdorff measure only works in special circumstances and is not the basis of a more general method. Usually we need to work rather harder!

Example 2.7

Let F be the middle third Cantor set (*see figure* 0.1). *If* $s = \log 2/\log 3 = 0.6309\ldots$ *then* $\dim_{\mathrm{H}} F = s$ and $\frac{1}{2} \leqslant \mathcal{H}^s(F) \leqslant 1$.

Heuristic calculation. The Cantor set F splits into a left part $F_{\mathrm{L}} = F \cap [0, \frac{1}{3}]$ and a right part $F_{\mathrm{R}} = F \cap [\frac{2}{3}, 1]$. Clearly both parts are geometrically similar to F but scaled by a ratio $\frac{1}{3}$, and $F = F_{\mathrm{L}} \cup F_{\mathrm{R}}$ with this union disjoint. Thus for any s

$$\mathcal{H}^s(F) = \mathcal{H}^s(F_{\mathrm{L}}) + \mathcal{H}^s(F_{\mathrm{R}}) = (\tfrac{1}{3})^s\mathcal{H}^s(F) + (\tfrac{1}{3})^s\mathcal{H}^s(F)$$

by the scaling property 2.1 of Hausdorff measures. *Assuming that at the critical value* $s = \dim_{\mathrm{H}} F$ *we have* $0 < \mathcal{H}^s(F) < \infty$ (a big assumption, but one that can be justified) we may divide by $\mathcal{H}^s(F)$ to get $1 = 2(\frac{1}{3})^s$ or $s = \log 2/\log 3$.

Rigorous calculation. We call the intervals that make up the sets E_k in the construction of F *level-k intervals*. Thus E_k consists of 2^k level-k intervals each of length 3^{-k}.

Taking the intervals of E_k as a 3^{-k}-cover of F gives that $\mathcal{H}^s_{3^{-k}}(F) \leqslant 2^k 3^{-ks} = 1$ if $s = \log 2/\log 3$. Letting $k \to \infty$ gives $\mathcal{H}^s(F) \leqslant 1$.

To prove that $\mathcal{H}^s(F) \geqslant \frac{1}{2}$ we show that

$$\sum |U_i|^s \geqslant \tfrac{1}{2} = 3^{-s} \tag{2.14}$$

for any cover $\{U_i\}$ of F. Clearly, it is enough to assume that the $\{U_i\}$ are intervals, and by expanding them slightly and using the compactness of F, we need only verify (2.14) if $\{U_i\}$ is a finite collection of closed subintervals of [0, 1]. For each U_i, let k be the integer such that

$$3^{-(k+1)} \leqslant |U_i| < 3^{-k}. \tag{2.15}$$

Then U_i can intersect at most one level-k interval since the separation of these level-k intervals is at least 3^{-k}. If $j \geqslant k$ then, by construction, U_i intersects at most $2^{j-k} = 2^j 3^{-sk} \leqslant 2^j 3^s |U_i|^s$ level-j intervals of E_j, using (2.15). If we choose j large enough so that $3^{-(j+1)} \leqslant |U_i|$ for all U_i, then, since the $\{U_i\}$ intersect all 2^j basic intervals of length 3^{-j}, counting intervals gives $2^j \leqslant \sum_i 2^j 3^s |U_i|^s$, which reduces to (2.14). □

With extra effort, the calculation can be adapted to show that $\mathcal{H}^s(F) = 1$.

It is already becoming apparent that calculation of Hausdorff measures and dimensions can be a little involved, even for simple sets. Usually it is the lower estimate that is awkward to obtain.

The 'heuristic' method of calculation used in Example 2.7 gives the right answer for the dimension of many self-similar sets. For example, the von Koch curve is made up of four copies of itself scaled by a factor $\frac{1}{3}$, and hence has dimension $\log 4/\log 3$. More generally, if $F = \bigcup_{i=1}^m F_i$, where each F_i is geometrically similar to F but scaled by a factor c_i then, provided that the F_i do not overlap 'too much', the heuristic argument gives $\dim_{\mathrm{H}} F$ as the number s satisfying $\sum_{i=1}^m c_i^s = 1$. The validity of this formula is discussed fully in Chapter 9.

*2.4 Equivalent definitions of Hausdorff dimension

It is worth pointing out that there are other classes of covering set that define measures leading to Hausdorff dimension. For example, we could use coverings by spherical balls: letting

$$\mathcal{B}^s_\delta(F) = \inf\{\Sigma|B_i|^s : \{B_i\} \text{ is a } \delta\text{-cover of } F \text{ by balls}\} \tag{2.16}$$

we obtain a measure $\mathcal{B}^s(F) = \lim_{\delta\to 0}\mathcal{B}^s_\delta(F)$ and a 'dimension' at which $\mathcal{B}^s(F)$ jumps from ∞ to 0. Clearly $\mathcal{H}^s_\delta(F) \leqslant \mathcal{B}^s_\delta(F)$ since any δ-cover of F by balls is a permissible covering in the definition of $\mathcal{H}^s_\delta$. Also, if $\{U_i\}$ is a δ-cover of F, then $\{B_i\}$ is a 2δ-cover, where, for each i, B_i is chosen to be some ball containing U_i and of radius $|U_i| \leqslant \delta$. Thus $\Sigma|B_i|^s \leqslant \Sigma(2|U_i|)^s = 2^s\Sigma|U_i|^s$, and taking infima gives $\mathcal{B}^s_{2\delta}(F) \leqslant 2^s\mathcal{H}^s_\delta(F)$. Letting $\delta \to 0$ it follows that $\mathcal{H}^s(F) \leqslant \mathcal{B}^s(F) \leqslant 2^s\mathcal{H}^s(F)$. In particular, this implies that the values of s at which $\mathcal{H}^s$ and $\mathcal{B}^s$ jump from ∞ to 0 are the same, so that the dimensions defined by the two measures are equal.

It is easy to check that we get the same values for Hausdorff measure and dimension if in (2.1) we use δ-covers of just open sets or just closed sets. Moreover, if F is compact, then, by expanding the covering sets slightly to open sets, and taking a finite subcover, we get the same value of $\mathcal{H}^s(F)$ if we merely consider δ-covers by finite collections of sets.

Net measures are another useful variant. For the sake of simplicity let F be a subset of the interval $[0, 1)$. A *binary interval* is an interval of the form $[r2^{-k}, (r+1)2^{-k})$ where $k = 0, 1, 2, \ldots$ and $r = 0, 1, \ldots, 2^k - 1$. We define

$$\mathcal{M}^s_\delta(F) = \inf\{\Sigma|U_i|^s : \{U_i\} \text{ is a } \delta\text{-cover of } F \text{ by binary intervals}\} \tag{2.17}$$

leading to the *net measures*

$$\mathcal{M}^s(F) = \lim_{\delta\to 0}\mathcal{M}^s_\delta(F). \tag{2.18}$$

Since any interval $U \subset [0, 1)$ is contained in two consecutive binary intervals each of length at most $2|U|$ we see, in just the same way as for the measure $\mathcal{B}^s$, that

$$\mathcal{H}^s(F) \leqslant \mathcal{M}^s(F) \leqslant 2^{s+1}\mathcal{H}^s(F). \tag{2.19}$$

It follows that the value of s at which $\mathcal{M}^s(F)$ jumps from ∞ to 0 equals the Hausdorff dimension of F, i.e. both definitions of measure give the same dimension.

For certain purposes net measures are much more convenient than Hausdorff measures. This is because two binary intervals are either disjoint or one of them is contained in the other, allowing any cover of binary intervals to be reduced to a cover of *disjoint* binary intervals.

*2.5 Finer definitions of dimension

It is sometimes desirable to have a sharper indication of dimension than just a number. To achieve this let $h : \mathbb{R}^+ \to \mathbb{R}^+$ be a function that is increasing and continuous, which we call a *dimension function* or *gauge function*. Analogously to (2.1) we define

$$\mathcal{H}^h_\delta(F) = \inf\{\Sigma h(|U_i|) : \{U_i\} \text{ is a } \delta\text{-cover of } F\} \tag{2.20}$$

for F a subset of $\mathbb{R}^n$. This leads to a measure, taking $\mathcal{H}^h(F) = \lim_{\delta\to 0}\mathcal{H}^h_\delta(F)$. (If $h(t) = t^s$ this is the usual definition of s-dimensional Hausdorff measure.) If h and g are dimension functions such that $h(t)/g(t) \to 0$ as $t \to 0$ then, by an argument similar to (2.10), we get that $\mathcal{H}^h(F) = 0$ whenever $\mathcal{H}^g(F) < \infty$. Thus partitioning the dimension functions into those for which $\mathcal{H}^h$ is finite and those for which it is infinite gives a more precise indication of the 'dimension' of F than just the number $\dim_{\mathrm{H}} F$.

An important example of this is Brownian motion in $\mathbb{R}^3$ (see Chapter 16 for further details). It may be shown that (with probability 1) a Brownian path has Hausdorff dimension 2 but with $\mathcal{H}^2$-measure equal to 0. More refined calculations show that such a path has positive and finite $\mathcal{H}^h$-measure, where $h(t) = t^2 \log\log(1/t)$. Although Brownian paths have dimension 2, the dimension is, in a sense, logarithmically smaller than 2.

2.6 Notes and references

The idea of defining measures using covers of sets was introduced by Carathéodory (1914). Hausdorff (1919) used this method to define the measures that now bear his name, and showed that the middle third Cantor set has positive and finite measure of dimension $\log 2/\log 3$. Properties of Hausdorff measures have been developed ever since, not least by Besicovitch and his students.

Technical aspects of Hausdorff measures and dimensions are discussed in rather more detail in Falconer (1985a), and in greater generality in the books of Rogers (1998), Federer (1996) and Mattila (1995). Merzenich and Staiger (1994) relate Hausdorff dimension to formal languages and automata theory.

Exercises

2.1 Verify that the value of $\mathcal{H}^s(F)$ is unaltered if, in (2.1), we only consider δ-covers by sets $\{U_i\}$ that are all closed.

2.2 Show that $\mathcal{H}^0(F)$ equals the number of points in the set F.

2.3 Verify from the definition that $\mathcal{H}^s(\emptyset) = 0$, that $\mathcal{H}^s(E) \subset \mathcal{H}^s(F)$ if $E \subset F$, and that $\mathcal{H}^s(\bigcup_{i=1}^\infty F_i) \leqslant \sum_{i=1}^\infty \mathcal{H}^s(F_i)$.

2.4 Let F be the closed interval $[0, 1]$. Show that $\mathcal{H}^s(F) = \infty$ if $0 \leqslant s < 1$, that $\mathcal{H}^s(F) = 0$ if $s > 1$, and that $0 < \mathcal{H}^1(F) < \infty$.

2.5 Let $f : \mathbb{R} \to \mathbb{R}$ be a differentiable function with continuous derivative. Show that $\dim_{\mathrm{H}} f(F) \leqslant \dim_{\mathrm{H}} F$ for any set F. (Consider the case of F bounded first and show that f is Lipschitz on F.)

2.6 Let $f : \mathbb{R} \to \mathbb{R}$ be the function $f(x) = x^2$, and let F be any subset of $\mathbb{R}$. Show that $\dim_{\mathrm{H}} f(F) = \dim_{\mathrm{H}} F$.

2.7 Let $f : [0, 1] \to \mathbb{R}$ be a Lipschitz function. Writing $\operatorname{graph} f = \{(x, f(x)) : 0 \leqslant x \leqslant 1\}$, show that $\dim_{\mathrm{H}} \operatorname{graph} f = 1$. Note, in particular, that this is true if f is continuously differentiable, see Exercise 1.13.

2.8 What is the Hausdorff dimension of the sets $\{0, 1, 2, 3, \ldots\}$ and $\{0, 1, \frac{1}{2}, \frac{1}{3}, \frac{1}{4}, \ldots\}$ in $\mathbb{R}$?

2.9 Let F be the set consisting of the numbers between 0 and 1 whose decimal expansions do not contain the digit 5. Use a 'heuristic' argument to show that $\dim_{\mathrm{H}} F = \log 9/\log 10$. Can you prove this by a rigorous argument? Generalize this result.

2.10 Let F consist of the points $(x, y) \in \mathbb{R}^2$ such that the decimal expansions of neither x or y contain the digit 5. Use a 'heuristic' argument to show that $\dim_{\mathrm{H}} F = 2 \log 9/\log 10$.

2.11 Use a 'heuristic' argument to show that the Hausdorff dimension of the set depicted in figure 0.5 is given by the solution of the equation $4(\frac{1}{4})^s + (\frac{1}{2})^s = 1$. By solving a quadratic equation in $(\frac{1}{2})^s$, find an explicit expression for s.

2.12 Let F be the set of real numbers with base-3 expansion $b_m b_{m-1} \cdots b_1 \cdot a_1 a_2 \cdots$ with none of the digits b_i or a_i equal to 1. (Thus F is constructed by a Cantor-like process extending outwards as well as inwards.) What is the Hausdorff dimension of F?

2.13 What is the Hausdorff dimension of the set of numbers x with base-3 expansion $0 \cdot a_1 a_2 \cdots$ for which there is a positive integer k (which may depend on x) such that $a_i \neq 1$ for all $i \geqslant k$?

2.14 Let F be the middle-λ Cantor set (obtained by removing a proportion $0 < \lambda < 1$ from the middle of intervals). Use a 'heuristic argument' to show that $\dim_{\mathrm{H}} F = \log 2/\log(2/(1 - \lambda))$. Now let $E = F \times F \subset \mathbb{R}^2$. Show in the same way that $\dim_{\mathrm{H}} E = 2 \log 2/\log(2/(1 - \lambda))$.

2.15 Show that there is a totally disconnected subset of the plane of Hausdorff dimension s for every $0 \leqslant s \leqslant 2$. (Modify the construction of the Cantor dust in figure 0.4.)

2.16 Let S be the unit circle in the plane, with points on S parameterized by the angle θ subtended at the centre with a fixed axis, so that θ_1 and θ_2 represent the same point if and only if θ_1 and θ_2 differ by a multiple of 2π, in the usual way. Let $F = \{\theta \in S : 0 \leqslant 3^k\theta \leqslant \pi (\bmod 2\pi)$ for all $k = 1, 2, \ldots\}$. Show that $\dim_{\mathrm{H}} F = \log 2/\log 3$.

2.17 Show that if h and g are dimension functions such that $h(t)/g(t) \to 0$ as $t \to 0$ then $\mathcal{H}^h(F) = 0$ whenever $\mathcal{H}^g(F) < \infty$.

Chapter 3 Alternative definitions of dimension

Hausdorff dimension, discussed in the last chapter, is the principal definition of dimension that we shall work with. However, other definitions are in widespread use, and it is appropriate to examine some of these and their inter-relationship. Not all definitions are generally applicable—some only describe particular classes of set, such as curves.

Fundamental to most definitions of dimension is the idea of 'measurement at scale δ'. For each δ, we measure a set in a way that ignores irregularities of size less than δ, and we see how these measurements behave as $\delta \to 0$. For example, if F is a plane curve, then our measurement, $M_\delta(F)$, might be the number of steps required by a pair of dividers set at length δ to traverse F. A dimension of F is then determined by the power law (if any) obeyed by $M_\delta(F)$ as $\delta \to 0$. If

$$M_\delta(F) \sim c\delta^{-s} \tag{3.1}$$

for constants c and s, we might say that F has 'divider dimension' s, with c regarded as the 's-dimensional length' of F. Taking logarithms

$$\log M_\delta(F) \simeq \log c - s \log \delta \tag{3.2}$$

in the sense that the difference of the two sides tends to 0 with δ, and

$$s = \lim_{\delta \to 0} \frac{\log M_\delta(F)}{-\log \delta}. \tag{3.3}$$

These formulae are appealing for computational or experimental purposes, since s can be estimated as minus the gradient of a log–log graph plotted over a suitable range of δ; see figure 3.1. Of course, for real phenomena, we can only work with a finite range of δ; theory and experiment diverge before an atomic scale is

Fractal Geometry: Mathematical Foundations and Application. Second Edition Kenneth Falconer
© 2003 John Wiley & Sons, Ltd ISBNs: 0-470-84861-8 (HB); 0-470-84862-6 (PB)

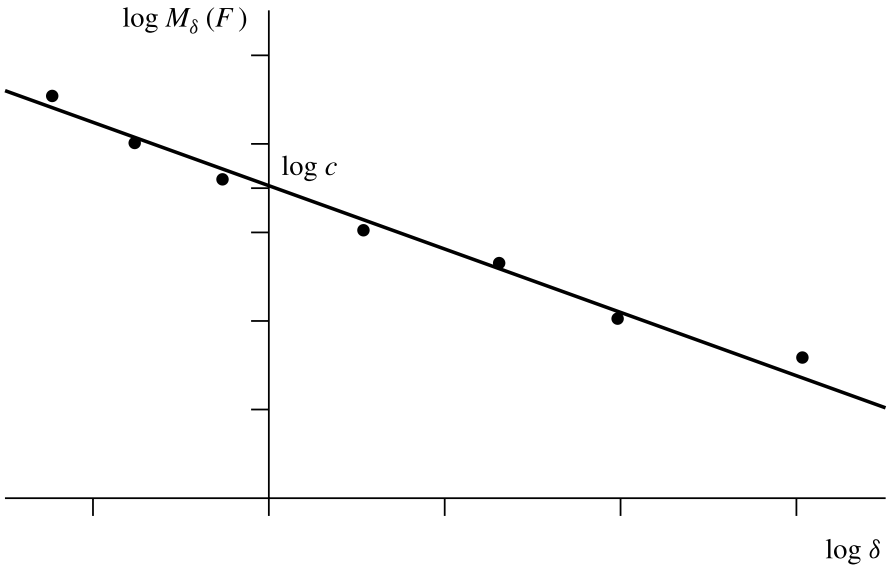


Figure 3.1 Empirical estimation of a dimension of a set F, on the power-law assumption $M_\delta(F) \sim c\delta^{-s}$

reached. For example, if F is the coastline of Britain, plotting a log–log graph for δ between 20 m and 200 km gives the divider dimension of F about 1.2.

There may be no exact power law for $M_\delta(F)$, and the closest we can get to (3.3) are the lower and upper limits.

For the value of s given by (3.1) to behave like a dimension, the method of measurement needs to scale with the set, so that doubling the size of F and at the same time doubling the scale at which measurement takes place does not affect the answer; that is, we require $M_\delta(\delta F) = M_1(F)$ for all δ. If we modify our example and redefine $M_\delta(F)$ to be the sum of the divider step lengths then $M_\delta(F)$ is homogeneous of degree 1, i.e. $M_\delta(\delta F) = \delta^1 M_1(F)$ for $\delta > 0$, and this must be taken into account when defining the dimension. In general, if $M_\delta(F)$ is homogeneous of degree d, that is $M_\delta(\delta F) = \delta^d M_1(F)$, then a power law of the form $M_\delta(F) \sim c\delta^{d-s}$ corresponds to a dimension s.

There are no hard and fast rules for deciding whether a quantity may reasonably be regarded as a dimension. There are many definitions that do not fit exactly into the above, rather simplified, scenario. The factors that determine the acceptability of a definition of a dimension are recognized largely by experience and intuition. In general one looks for some sort of scaling behaviour, a naturalness of the definition in the particular context and properties typical of dimensions such as those discussed below.

A word of warning: as we shall see, apparently similar definitions of dimension can have widely differing properties. It should not be assumed that different definitions give the same value of dimension, even for ‘nice’ sets. Such assumptions have led to major misconceptions and confusion in the past. It is necessary to derive the properties of any ‘dimension’ from its definition. The properties of Hausdorff dimension (on which we shall largely concentrate in the later chapters of this book) do not necessarily all hold for other definitions.

What are the desirable properties of a ‘dimension’? Those derived in the last chapter for Hausdorff dimension are fairly typical.

Monotonicity. If $E \subset F$ then $\dim_{\mathrm{H}} E \leqslant \dim_{\mathrm{H}} F$.

Stability. $\dim_{\mathrm{H}}(E \cup F) = \max(\dim_{\mathrm{H}} E, \dim_{\mathrm{H}} F)$.

Countable stability. $\dim_{\mathrm{H}} \left(\bigcup_{i=1}^{\infty} F_i\right) = \sup_{1 \leqslant i < \infty} \dim_{\mathrm{H}} F_i$.

Geometric invariance. $\dim_{\mathrm{H}} f(F) = \dim_{\mathrm{H}} F$ if f is a transformation of $\mathbb{R}^n$ such as a translation, rotation, similarity or affinity.

Lipschitz invariance. $\dim_{\mathrm{H}} f(F) = \dim_{\mathrm{H}} F$ if f is a bi-Lipschitz transformation.

Countable sets. $\dim_{\mathrm{H}} F = 0$ if F is finite or countable.

Open sets. If F is an open subset of $\mathbb{R}^n$ then $\dim_{\mathrm{H}} F = n$.

Smooth manifolds. $\dim_{\mathrm{H}} F = m$ if F is a smooth m-dimensional manifold (curve, surface, etc.).

All definitions of dimension are monotonic, most are stable, but, as we shall see, some common definitions fail to exhibit countable stability and may have countable sets of positive dimension. All the usual dimensions are Lipschitz invariant, and, therefore, geometrically invariant. The ‘open sets’ and ‘smooth manifolds’ properties ensure that the dimension is an extension of the classical definition. Note that different definitions of dimension can provide different information about which sets are Lipschitz equivalent.

3.1 Box-counting dimensions

Box-counting or box dimension is one of the most widely used dimensions. Its popularity is largely due to its relative ease of mathematical calculation and empirical estimation. The definition goes back at least to the 1930s and it has been variously termed Kolmogorov entropy, entropy dimension, capacity dimension (a term best avoided in view of potential theoretic associations), metric dimension, logarithmic density and information dimension. We shall always refer to box or box-counting dimension to avoid confusion.

Let F be any non-empty bounded subset of $\mathbb{R}^n$ and let $N_\delta(F)$ be the smallest number of sets of diameter at most δ which can cover F. The *lower* and *upper box-counting dimensions* of F respectively are defined as

$$\underline{\dim}_{\mathrm{B}} F = \underline{\lim}_{\delta \to 0} \frac{\log N_\delta(F)}{-\log \delta} \tag{3.4}$$

$$\overline{\dim}_{\mathrm{B}} F = \overline{\lim}_{\delta \to 0} \frac{\log N_\delta(F)}{-\log \delta}. \tag{3.5}$$

If these are equal we refer to the common value as the *box-counting dimension* or *box dimension* of F

$$\dim_{\mathrm{B}} F = \lim_{\delta \to 0} \frac{\log N_\delta(F)}{-\log \delta}. \tag{3.6}$$

Here, and throughout the book, we assume that $\delta > 0$ is sufficiently small to ensure that $-\log\delta$ and similar quantities are strictly positive. To avoid problems with '$\log 0$' or '$\log\infty$' we generally consider box dimension only for non-empty bounded sets. In developing the general theory of box dimensions we assume that sets considered are non-empty and bounded.

There are several equivalent definitions of box dimension that are sometimes more convenient to use. Consider the collection of cubes in the δ-coordinate mesh of $\mathbb{R}^n$, i.e. cubes of the form

$$[m_1\delta, (m_1+1)\delta] \times \cdots \times [m_n\delta, (m_n+1)\delta]$$

where $m_1, \ldots, m_n$ are integers. (Recall that a 'cube' is an interval in $\mathbb{R}^1$ and a square in $\mathbb{R}^2$.) Let $N'_\delta(F)$ be the number of δ-mesh cubes that intersect F. They obviously provide a collection of $N'_\delta(F)$ sets of diameter $\delta\sqrt{n}$ that cover F, so

$$N_{\delta\sqrt{n}}(F) \leqslant N'_\delta(F).$$

If $\delta\sqrt{n} < 1$ then

$$\frac{\log N_{\delta\sqrt{n}}(F)}{-\log(\delta\sqrt{n})} \leqslant \frac{\log N'_\delta(F)}{-\log\sqrt{n} - \log\delta}$$

so taking limits as $\delta \to 0$

$$\underline{\dim}_{\mathrm{B}} F \leqslant \varliminf_{\delta\to 0} \frac{\log N'_\delta(F)}{-\log\delta} \tag{3.7}$$

and

$$\overline{\dim}_{\mathrm{B}} F \leqslant \varlimsup_{\delta\to 0} \frac{\log N'_\delta(F)}{-\log\delta}. \tag{3.8}$$

On the other hand, any set of diameter at most δ is contained in 3^n mesh cubes of side δ (by choosing a cube containing some point of the set together with its neighbouring cubes). Thus

$$N'_\delta(F) \leqslant 3^n N_\delta(F)$$

and taking logarithms and limits as $\delta \to 0$ leads to the opposite inequalities to (3.7) and (3.8). Hence to find the box dimensions (3.4)–(3.6), we can equally well take $N_\delta(F)$ to be the number of mesh cubes of side δ that intersect F.

This version of the definitions is widely used empirically. To find the box dimension of a plane set F we draw a mesh of squares or boxes of side δ and count the number $N_\delta(F)$ that overlap the set for various small δ (hence the name 'box-counting'). The dimension is the logarithmic rate at which $N_\delta(F)$ increases as $\delta \to 0$, and may be estimated by the gradient of the graph of $\log N_\delta(F)$ against $-\log\delta$.

This definition gives an interpretation of the meaning of box dimension. The number of mesh cubes of side δ that intersect a set is an indication of how spread out or irregular the set is when examined at scale δ. The dimension reflects how rapidly the irregularities develop as $\delta \to 0$.

Another frequently used definition of box dimension is obtained by taking $N_\delta(F)$ in (3.4)–(3.6) to be the smallest number of *arbitrary* cubes of side δ required to cover F. The equivalence of this definition follows as in the mesh cube case, noting that any cube of side δ has diameter $\delta\sqrt{n}$, and that any set of diameter of at most δ is contained in a cube of side δ.

Similarly, we get exactly the same values if in (3.4)–(3.6) we take $N_\delta(F)$ as the smallest number of closed balls of radius δ that cover F.

A less obviously equivalent formulation of box dimension has the *largest* number of *disjoint* balls of radius δ with centres in F. Let this number be $N'_\delta(F)$, and let $B_1, \dots, B_{N'_\delta(F)}$ be disjoint balls centred in F and of radius δ. If x belongs to F then x must be within distance δ of one of the B_i, otherwise the ball of centre x and radius δ can be added to form a larger collection of disjoint balls. Thus the $N'_\delta(F)$ balls concentric with the B_i but of radius 2δ (diameter 4δ) cover F, giving

$$N_{4\delta}(F) \leqslant N'_\delta(F). \tag{3.9}$$

Suppose also that $B_1, \dots, B_{N'_\delta(F)}$ are disjoint balls of radii δ with centres in F. Let $U_1, \dots, U_k$ be any collection of sets of diameter at most δ which cover F. Since the U_j must cover the centres of the B_i, each B_i must contain at least one of the U_j. As the B_i are disjoint there are at least as many U_j as B_i. Hence

$$N'_\delta(F) \leqslant N_\delta(F). \tag{3.10}$$

Taking logarithms and limits of (3.9) and (3.10) shows that the values of (3.4)–(3.6) are unaltered if $N_\delta(F)$ is replaced by this $N'_\delta(F)$.

These various definitions are summarized below and in figure 3.2.

Equivalent definitions 3.1

The lower and upper box-counting dimensions of a subset F of $\mathbb{R}^n$ are given by

$$\underline{\dim}_{\mathrm{B}} F = \underline{\lim}_{\delta\to 0} \frac{\log N_\delta(F)}{-\log\delta} \tag{3.11}$$

$$\overline{\dim}_{\mathrm{B}} F = \overline{\lim}_{\delta\to 0} \frac{\log N_\delta(F)}{-\log\delta} \tag{3.12}$$

and the box-counting dimension of F by

$$\dim_{\mathrm{B}} F = \lim_{\delta\to 0} \frac{\log N_\delta(F)}{-\log\delta} \tag{3.13}$$

(*if this limit exists*), *where* $N_\delta(F)$ *is any of the following*:

(i) *the smallest number of closed balls of radius δ that cover F*;
(ii) *the smallest number of cubes of side δ that cover F*;
(iii) *the number of δ-mesh cubes that intersect F*;
(iv) *the smallest number of sets of diameter at most δ that cover F*;
(v) *the largest number of disjoint balls of radius δ with centres in F.*

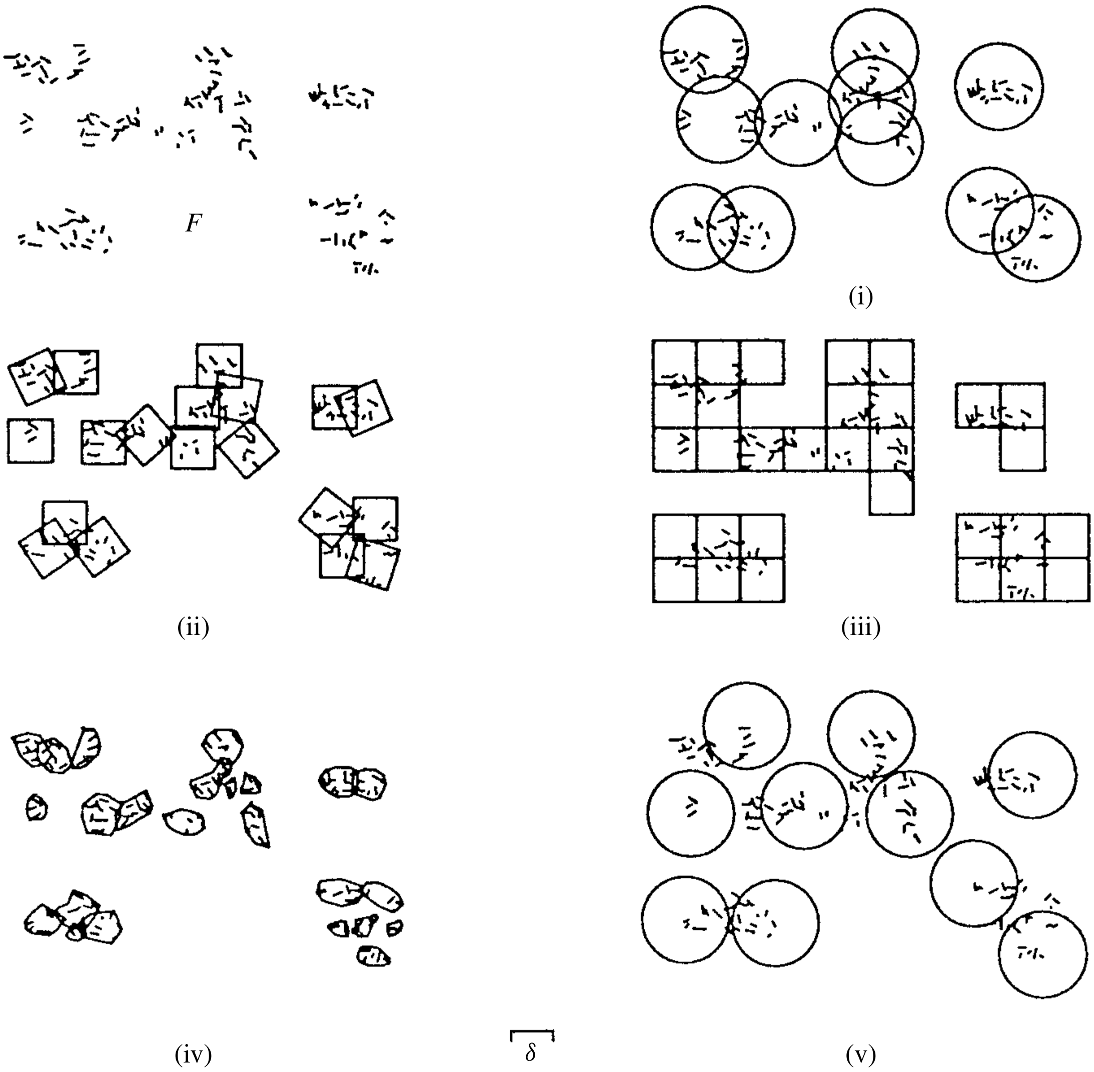


Figure 3.2 Five ways of finding the box dimension of F; see Equivalent definitions 3.1. The number $N_\delta(F)$ is taken to be: (i) the least number of closed balls of radius δ that cover F; (ii) the least number of cubes of side δ that cover F; (iii) the number of δ-mesh cubes that intersect F; (iv) the least number of sets of diameter at most δ that cover F; (v) the greatest number of disjoint balls of radius δ with centres in F

This list could be extended further; in practice one adopts the definition most convenient for a particular application.

It is worth noting that, in (3.11)–(3.13), it is enough to consider limits as δ tends to 0 through any decreasing sequence δ_k such that $\delta_{k+1} \geqslant c\delta_k$ for some constant $0 < c < 1$; in particular for $\delta_k = c^k$. To see this, note that if $\delta_{k+1} \leqslant \delta < \delta_k$, then, with $N_\delta(F)$ the least number of sets in a δ-cover of F,

$$\frac{\log N_\delta(F)}{-\log\delta} \leqslant \frac{\log N_{\delta_{k+1}}(F)}{-\log\delta_k} = \frac{\log N_{\delta_{k+1}}(F)}{-\log\delta_{k+1} + \log(\delta_{k+1}/\delta_k)} \leqslant \frac{\log N_{\delta_{k+1}}(F)}{-\log\delta_{k+1} + \log c}$$

and so

$$\overline{\lim_{\delta\to 0}}\frac{\log N_\delta(F)}{-\log\delta} \leqslant \overline{\lim_{k\to\infty}}\frac{\log N_{\delta_k}(F)}{-\log\delta_k}. \tag{3.14}$$

The opposite inequality is trivial; the case of lower limits may be dealt with in the same way.

There is an equivalent definition of box dimension of a rather different form that is worth mentioning. Recall that the *δ-neighbourhood* F_δ of a subset F of $\mathbb{R}^n$ is

$$F_\delta = \{x \in \mathbb{R}^n : |x-y| \leqslant \delta \text{ for some } y \in F\} \tag{3.15}$$

i.e. the set of points within distance δ of F. We consider the rate at which the n-dimensional volume of F_δ shrinks as $\delta \to 0$. In $\mathbb{R}^3$, if F is a single point then F_δ is a ball with $\text{vol}(F_\delta) = \frac{4}{3}\pi\delta^3$, if F is a segment of length l then F_δ is 'sausage-like' with $\text{vol}(F_\delta) \sim \pi l\delta^2$, and if F is a flat set of area a then F_δ is essentially a thickening of F with $\text{vol}(F_\delta) \sim 2a\delta$. In each case, $\text{vol}(F_\delta) \sim c\delta^{3-s}$ where the integer s is the dimension of F, so that exponent of δ is indicative of the dimension. The coefficient c of δ^{3-s}, known as the *Minkowski content* of F, is a measure of the length, area or volume of the set as appropriate.

This idea extends to fractional dimensions. If F is a subset of $\mathbb{R}^n$ and, for some s, $\text{vol}^n(F_\delta)/\delta^{n-s}$ tends to a positive finite limit as $\delta \to 0$ where vol^n denotes n-dimensional volume, then it makes sense to regard F as s-dimensional. The limiting value is called the *s-dimensional content* of F—a concept of slightly restricted use since it is not necessarily additive on disjoint subsets, i.e. is not a measure. Even if this limit does not exist, we may be able to extract the critical exponent of δ and this turns out to be related to the box dimension.

Proposition 3.2

If F is a subset of $\mathbb{R}^n$, then

$$\underline{\dim}_{\text{B}} F = n - \overline{\lim_{\delta\to 0}}\frac{\log \text{vol}^n(F_\delta)}{\log\delta}$$

$$\overline{\dim}_{\text{B}} F = n - \underline{\lim}_{\delta\to 0}\frac{\log \text{vol}^n(F_\delta)}{\log\delta}$$

where F_δ is the δ-neighbourhood of F.

Proof. If F can be covered by $N_\delta(F)$ balls of radius $\delta < 1$ then F_δ can be covered by the concentric balls of radius 2δ. Hence

$$\text{vol}^n(F_\delta) \leqslant N_\delta(F)c_n(2\delta)^n$$

where c_n is the volume of the unit ball in $\mathbb{R}^n$. Taking logarithms,

$$\frac{\log \mathrm{vol}^n(F_\delta)}{-\log\delta} \leqslant \frac{\log 2^n c_n + n\log\delta + \log N_\delta(F)}{-\log\delta},$$

so

$$\varliminf_{\delta\to 0}\frac{\log \mathrm{vol}^n(F_\delta)}{-\log\delta} \leqslant -n + \underline{\dim}_{\mathrm{B}} F \tag{3.16}$$

with a similar inequality for the upper limits. On the other hand if there are $N_\delta(F)$ disjoint balls of radius δ with centres in F, then by adding their volumes,

$$N_\delta(F)c_n\delta^n \leqslant \mathrm{vol}^n(F_\delta).$$

Taking logarithms and letting $\delta \to 0$ gives the opposite inequality to (3.16), using Equivalent definition 3.1(v). □

In the context of Proposition 3.2, box dimension is sometimes referred to as *Minkowski dimension* or *Minkowski–Bouligand dimension.*

It is important to understand the relationship between box-counting dimension and Hausdorff dimension. If F can be covered by $N_\delta(F)$ sets of diameter δ, then, from definition (2.1),

$$\mathcal{H}^s_\delta(F) \leqslant N_\delta(F)\delta^s.$$

If $1 < \mathcal{H}^s(F) = \lim_{\delta\to 0}\mathcal{H}^s_\delta(F)$ then $\log N_\delta(F) + s\log\delta > 0$ if δ is sufficiently small. Thus $s \leqslant \varliminf_{\delta\to 0}\log N_\delta(F)/-\log\delta$ so

$$\dim_{\mathrm{H}} F \leqslant \underline{\dim}_{\mathrm{B}} F \leqslant \overline{\dim}_{\mathrm{B}} F \tag{3.17}$$

for every $F \subset \mathbb{R}^n$. We do *not* in general get equality here. Although Hausdorff and box dimensions are equal for many ‘reasonably regular’ sets, there are plenty of examples where this inequality is strict.

Roughly speaking (3.6) says that $N_\delta(F) \simeq \delta^{-s}$ for small δ, where $s = \dim_{\mathrm{B}} F$. More precisely, it says that

$$N_\delta(F)\delta^s \to \infty \qquad \text{if } s < \dim_{\mathrm{B}} F$$

and

$$N_\delta(F)\delta^s \to 0 \qquad \text{if } s > \dim_{\mathrm{B}} F.$$

But

$$N_\delta(F)\delta^s = \inf\left\{\sum_i \delta^s : \{U_i\} \text{ is a (finite) } \delta\text{-cover of } F\right\},$$

which should be compared with

$$\mathcal{H}^s_\delta(F) = \inf\left\{\sum_i |U_i|^s : \{U_i\} \text{ is a } \delta\text{-cover of } F\right\},$$

which occurs in the definitions of Hausdorff measure and dimension. In calculating Hausdorff dimension, we assign different weights $|U_i|^s$ to the covering sets U_i, whereas for the box dimensions we use the same weight δ^s for each covering set. Box dimensions may be thought of as indicating the efficiency with which a set may be covered by small sets of equal size, whereas Hausdorff dimension involves coverings by sets of small but perhaps widely varying size.

There is a temptation to introduce the quantity $v(F) = \underline{\lim}_{\delta\to 0} N_\delta(F)\delta^s$, but this does *not* give a measure on subsets of $\mathbb{R}^n$. As we shall see, one consequence of this is that box dimensions have a number of unfortunate properties, and can be awkward to handle mathematically.

Since box dimensions are determined by coverings by sets of equal size they tend to be easier to calculate than Hausdorff dimensions. Just as with Hausdorff dimension, calculations of box dimension usually involve finding a lower bound and an upper bound separately, each bound depending on a geometric observation followed by an algebraic estimate.

Example 3.3

Let F be the middle third Cantor set (figure 0.1). Then $\underline{\dim}_{\mathrm{B}} F = \overline{\dim}_{\mathrm{B}} F = \log 2/\log 3$.

Calculation. The obvious covering by the 2^k level-k intervals of E_k of length 3^{-k} gives that $N_\delta(F) \leqslant 2^k$ if $3^{-k} < \delta \leqslant 3^{-k+1}$. From (3.5)

$$\overline{\dim}_{\mathrm{B}} F = \overline{\lim_{\delta\to 0}} \frac{\log N_\delta(F)}{-\log\delta} \leqslant \overline{\lim_{k\to\infty}} \frac{\log 2^k}{\log 3^{k-1}} = \frac{\log 2}{\log 3}.$$

On the other hand, any interval of length δ with $3^{-k-1} \leqslant \delta < 3^{-k}$ intersects at most one of the level-k intervals of length 3^{-k} used in the construction of F. There are 2^k such intervals so at least 2^k intervals of length δ are required to cover F. Hence $N_\delta(F) \geqslant 2^k$ leading to $\underline{\dim}_{\mathrm{B}} F \geqslant \log 2/\log 3$. □

Thus, at least for the Cantor set, $\dim_{\mathrm{H}} F = \dim_{\mathrm{B}} F$.

3.2 Properties and problems of box-counting dimension

The following elementary properties of box dimension mirror those of Hausdorff dimension, and may be verified in much the same way.

(i) A smooth m-dimensional submanifold of $\mathbb{R}^n$ has $\dim_{\mathrm{B}} F = m$.

(ii) $\underline{\dim}_{\mathrm{B}}$ and $\overline{\dim}_{\mathrm{B}}$ are monotonic.

(iii) $\overline{\dim}_{\mathrm{B}}$ is *finitely* stable, i.e.

$$\overline{\dim}_{\mathrm{B}}(E \cup F) = \max\{\overline{\dim}_{\mathrm{B}} E, \overline{\dim}_{\mathrm{B}} F\};$$

the corresponding identity does *not* hold for $\underline{\dim}_{\mathrm{B}}$.

(iv) $\underline{\dim}_{\mathrm{B}}$ and $\overline{\dim}_{\mathrm{B}}$ are bi-Lipschitz invariant. This is so because, if $|f(x) - f(y)| \leqslant c|x - y|$ and F can be covered by $N_\delta(F)$ sets of diameter at most δ, then the $N_\delta(F)$ images of these sets under f form a cover of $f(F)$ by sets of diameter at most $c\delta$, thus $\dim_{\mathrm{B}} f(F) \leqslant \dim_{\mathrm{B}} F$. Similarly, box dimensions behave just like Hausdorff dimensions under bi-Lipschitz and Hölder transformations.

We now start to encounter the disadvantages of box-counting dimension. The next proposition is at first appealing, but has undesirable consequences.

Proposition 3.4

Let $\overline{F}$ denote the closure of F (i.e. the smallest closed subset of $\mathbb{R}^n$ containing F). Then

$$\underline{\dim}_{\mathrm{B}} \overline{F} = \underline{\dim}_{\mathrm{B}} F$$

and

$$\overline{\dim}_{\mathrm{B}} \overline{F} = \overline{\dim}_{\mathrm{B}} F.$$

Proof. Let $B_1, \ldots, B_k$ be a finite collection of closed balls of radii δ. If the closed set $\bigcup_{i=1}^k B_i$ contains F, it also contains $\overline{F}$. Hence the smallest number of closed balls of radius δ that cover F equals the smallest number required to cover the larger set $\overline{F}$. The result follows. □

An immediate consequence of this is that if F is a dense subset of an open region of $\mathbb{R}^n$ then $\underline{\dim}_{\mathrm{B}} F = \overline{\dim}_{\mathrm{B}} F = n$. For example, let F be the (countable) set of rational numbers between 0 and 1. Then $\overline{F}$ is the entire interval [0, 1], so that $\underline{\dim}_{\mathrm{B}} F = \overline{\dim}_{\mathrm{B}} F = 1$. Thus countable sets can have non-zero box dimension. Moreover, the box-counting dimension of each rational number regarded as a one-point set is clearly zero, but the countable union of these singleton sets has dimension 1. Consequently, it is not generally true that $\dim_{\mathrm{B}} \bigcup_{i=1}^\infty F_i = \sup_i \dim_{\mathrm{B}} F_i$.

This severely limits the usefulness of box dimension—introducing a small, i.e. countable, set of points can play havoc with the dimension. We might hope to salvage something by restricting attention to closed sets, but difficulties still remain.

Example 3.5

$F = \{0, 1, \frac{1}{2}, \frac{1}{3}, \ldots\}$ *is a compact set with* $\dim_{\mathrm{B}} F = \frac{1}{2}$.

Calculation. Let $0 < \delta < \frac{1}{2}$ and let k be the integer satisfying $1/(k-1)k > \delta \geqslant 1/k(k+1)$. If $|U| \leqslant \delta$, then U can cover at most one of the points $\{1, \frac{1}{2}, \dots, 1/k\}$ since $1/(k-1) - 1/k = 1/(k-1)k > \delta$. Thus at least k sets of diameter δ are required to cover F, so $N_\delta(F) \geqslant k$ giving

$$\frac{\log N_\delta(F)}{-\log\delta} \geqslant \frac{\log k}{\log k(k+1)}.$$

Letting $\delta \to 0$ so $k \to \infty$ gives $\underline{\dim}_{\mathrm{B}} F \geqslant \frac{1}{2}$. On the other hand, if $\frac{1}{2} > \delta > 0$, take k such that $1/(k-1)k > \delta \geqslant 1/k(k+1)$. Then $(k+1)$ intervals of length δ cover $[0, 1/k]$, leaving $k-1$ points of F which can be covered by another $k-1$ intervals. Thus $N_\delta(F) \leqslant 2k$, so

$$\frac{\log N_\delta(F)}{-\log\delta} \leqslant \frac{\log(2k)}{\log k(k-1)}$$

giving

$$\overline{\dim}_{\mathrm{B}} F \leqslant \tfrac{1}{2}. \qquad \square$$

No-one would regard this set, with all but one of its points isolated, as a fractal, yet it has large box dimension.

Nevertheless, as well as being convenient in practice, box dimensions are very useful in theory. If, as often happens, it can be shown that a set has equal box and Hausdorff dimensions, the interplay between these definitions can be used to powerful effect.

*3.3 Modified box-counting dimensions

There are ways of overcoming the difficulties of box dimension outlined in the last section. However, they may not at first seem appealing since they re-introduce all the difficulties of calculation associated with Hausdorff dimension and more.

For F a subset of $\mathbb{R}^n$ we can try to decompose F into a countable number of pieces $F_1, F_2, \dots$ in such a way that the largest piece has as small a dimension as possible. This idea leads to the following *modified box-counting dimensions*:

$$\underline{\dim}_{\mathrm{MB}} F = \inf\left\{\sup_i \underline{\dim}_{\mathrm{B}} F_i : F \subset \bigcup_{i=1}^{\infty} F_i\right\} \tag{3.18}$$

$$\overline{\dim}_{\mathrm{MB}} F = \inf\left\{\sup_i \overline{\dim}_{\mathrm{B}} F_i : F \subset \bigcup_{i=1}^{\infty} F_i\right\}. \tag{3.19}$$

(In both cases the infimum is over all possible countable covers $\{F_i\}$ of F.) Clearly $\underline{\dim}_{\mathrm{MB}} F \leqslant \underline{\dim}_{\mathrm{B}} F$ and $\overline{\dim}_{\mathrm{MB}} F \leqslant \overline{\dim}_{\mathrm{B}} F$. However, we now have that

$\underline{\dim}_{\mathrm{MB}} F = \overline{\dim}_{\mathrm{MB}} F = 0$ if F is countable—just take the F_i to be one-point sets. Moreover, for any subset F of $\mathbb{R}^n$,

$$0 \leqslant \dim_{\mathrm{H}} F \leqslant \underline{\dim}_{\mathrm{MB}} F \leqslant \overline{\dim}_{\mathrm{MB}} F \leqslant \overline{\dim}_{\mathrm{B}} F \leqslant n. \tag{3.20}$$

It is easy to see that $\underline{\dim}_{\mathrm{MB}}$ and $\overline{\dim}_{\mathrm{MB}}$ recover all the desirable properties of a dimension, but they can be hard to calculate. However, there is a useful test for compact sets to have equal box and modified box dimensions. It applies to sets that might be described as 'dimensionally homogeneous'.

Proposition 3.6

Let $F \subset \mathbb{R}^n$ be compact. Suppose that

$$\overline{\dim}_{\mathrm{B}}(F \cap V) = \overline{\dim}_{\mathrm{B}} F \tag{3.21}$$

for all open sets V that intersect F. Then $\overline{\dim}_{\mathrm{B}} F = \overline{\dim}_{\mathrm{MB}} F$. A similar result holds for lower box-counting dimensions.

Proof. Let $F \subset \bigcup_{i=1}^{\infty} F_i$ with each F_i closed. A version of Baire's category theorem (which may be found in any text on basic general topology, and which we quote without proof) states that there is an index i and an open set $V \subset \mathbb{R}^n$ such that $F \cap V \subset F_i$. For this i, $\overline{\dim}_{\mathrm{B}} F_i = \overline{\dim}_{\mathrm{B}} F$. Using (3.19) and Proposition 3.4

$$\begin{aligned}\overline{\dim}_{\mathrm{MB}} F &= \inf\left\{\sup \overline{\dim}_{\mathrm{B}} F_i : F \subset \bigcup_{i=1}^{\infty} F_i \text{ where the } F_i \text{ are closed sets}\right\}\\ &\geqslant \overline{\dim}_{\mathrm{B}} F.\end{aligned}$$

The opposite inequality is contained in (3.20). A similar argument deals with the lower dimensions. □

For an application, let F be a compact set with a high degree of self-similarity, for instance the middle third Cantor set or von Koch curve. If V is any open set that intersects F, then $F \cap V$ contains a geometrically similar copy of F which must have upper box dimension equal to that of F, so that (3.21) holds, leading to equal box and modified box dimensions.

*3.4 Packing measures and dimensions

Unlike Hausdorff dimension, neither the box dimensions or modified box dimensions are defined in terms of measures, and this can present difficulties in their theoretical development. Nevertheless, the circle of ideas in the last section may be completed in a way that is, at least mathematically, elegant. Recall that Hausdorff dimension may be defined using economical coverings by small balls (2.16) whilst $\underline{\dim}_{\rm B}$ may be defined using economical coverings by small balls of equal radius (Equivalent definition 3.1(i)). On the other hand $\overline{\dim}_{\rm B}$ may be thought of as a dimension that depends on packings by disjoint balls of equal radius that are as dense as possible (Equivalent definition 3.1(v)). Coverings and packings play a dual role in many areas of mathematics and it is therefore natural to try to look for a dimension that is defined in terms of dense packings by disjoint balls of differing small radii.

We try to follow the pattern of definition of Hausdorff measure and dimension. For $s \geqslant 0$ and $\delta > 0$, let

$$\mathcal{P}^s_\delta(F) = \sup\left\{\sum_i |B_i|^s : \{B_i\} \text{ is a collection of disjoint balls of radii at most } \delta \text{ with centres in } F\right\}. \tag{3.22}$$

Since $\mathcal{P}^s_\delta(F)$ decreases with δ, the limit

$$\mathcal{P}^s_0(F) = \lim_{\delta\to 0}\mathcal{P}^s_\delta(F) \tag{3.23}$$

exists. At this point we meet the problems encountered with box-counting dimensions. By considering countable dense sets it is easy to see that $\mathcal{P}^s_0(F)$ is not a measure. Hence we modify the definition to

$$\mathcal{P}^s(F) = \inf\left\{\sum_i \mathcal{P}^s_0(F_i) : F \subset \bigcup_{i=1}^{\infty} F_i\right\}. \tag{3.24}$$

It may be shown that $\mathcal{P}^s(F)$ is a measure on $\mathbb{R}^n$, known as the *s-dimensional packing measure*. We may define the *packing dimension* in the natural way:

$$\dim_{\rm P} F = \sup\{s : \mathcal{P}^s(F) = \infty\} = \inf\{s : \mathcal{P}^s(F) = 0\}. \tag{3.25}$$

The underlying measure structure immediately implies monotonicity: that $\dim_{\rm P} E \leqslant \dim_{\rm P} F$ if $E \subset F$. Moreover, for a countable collection of sets $\{F_i\}$,

$$\dim_{\rm P}\left(\bigcup_{i=1}^{\infty} F_i\right) = \sup_i \dim_{\rm P} F_i, \tag{3.26}$$

since if $s > \dim_{\rm P} F_i$ for all i, then $\mathcal{P}^s(\bigcup_i F_i) \leqslant \sum_i \mathcal{P}^s(F_i) = 0$ implying $\dim_{\rm P}\left(\bigcup_i F_i\right) \leqslant s$.

We now investigate the relationship of packing dimension with other definitions of dimension and verify the surprising fact that packing dimension is just the same as the modified upper box dimension.

Lemma 3.7

$$\dim_{\mathrm{P}} F \leqslant \overline{\dim}_{\mathrm{B}} F. \tag{3.27}$$

Proof. If $\dim_{\mathrm{P}} F = 0$, the result is obvious. Otherwise choose any t and s with $0 < t < s < \dim_{\mathrm{P}} F$. Then $\mathcal{P}^s(F) = \infty$, so $\mathcal{P}_0^s(F) = \infty$. Thus, given $0 < \delta \leqslant 1$, there are disjoint balls $\{B_i\}$, of radii at most δ with centres in F, such that $1 < \Sigma_{i=1}^{\infty} |B_i|^s$. Suppose that, for each k, exactly n_k of these balls satisfy $2^{-k-1} < |B_i| \leqslant 2^{-k}$; then

$$1 < \sum_{k=0}^{\infty} n_k 2^{-ks}. \tag{3.28}$$

There must be some k with $n_k > 2^{kt}(1 - 2^{t-s})$, otherwise the sum in (3.28) is at most $\Sigma_{k=0}^{\infty} 2^{kt-ks}(1 - 2^{t-s}) = 1$, by summing the geometric series. These n_k balls all contain balls of radii $2^{-k-2} \leqslant \delta$ centred in F. Hence if $N_\delta(F)$ denotes the greatest number of disjoint balls of radius δ with centres in F, then

$$N_{2^{-k-2}}(F)(2^{-k-2})^t \geqslant n_k (2^{-k-2})^t > 2^{-2t}(1 - 2^{t-s})$$

where $2^{-k-2} < \delta$. It follows that $\overline{\lim}_{\delta \to 0} N_\delta(F)\delta^t > 0$, so that $\overline{\dim}_{\mathrm{B}} F \geqslant t$ using Equivalent definition 3.1(v). This is true for any $0 < t < \dim_{\mathrm{P}} F$ so (3.27) follows. □

Proposition 3.8

If $F \subset \mathbb{R}^n$ *then* $\dim_{\mathrm{P}} F = \overline{\dim}_{\mathrm{MB}} F$.

Proof. If $F \subset \bigcup_{i=1}^{\infty} F_i$ then, by (3.26) and (3.27),

$$\dim_{\mathrm{P}} F \leqslant \sup_i \dim_{\mathrm{P}} F_i \leqslant \sup_i \overline{\dim}_{\mathrm{B}} F_i.$$

Definition (3.19) now gives that $\dim_{\mathrm{P}} F \leqslant \overline{\dim}_{\mathrm{MB}} F$.

Conversely, if $s > \dim_{\mathrm{P}} F$ then $\mathcal{P}^s(F) = 0$, so that $F \subset \bigcup_i F_i$ for a collection of sets F_i with $\mathcal{P}_0^s(F_i) < \infty$ for each i, by (3.24). Hence, for each i, if δ is small enough, then $\mathcal{P}_\delta^s(F_i) < \infty$, so by (3.22) $N_\delta(F_i)\delta^s$ is bounded as $\delta \to 0$, where $N_\delta(F_i)$ is the largest number of disjoint balls of radius δ with centres in F_i. By Equivalent definition 3.1(v) $\overline{\dim}_{\mathrm{B}} F_i \leqslant s$ for each i, giving that $\overline{\dim}_{\mathrm{MB}} F \leqslant s$ by (3.19), as required. □

We have established the following relations:

$$\dim_{\mathrm{H}} F \leqslant \underline{\dim}_{\mathrm{MB}} F \leqslant \overline{\dim}_{\mathrm{MB}} F = \dim_{\mathrm{P}} F \leqslant \overline{\dim}_{\mathrm{B}} F. \tag{3.29}$$

Suitable examples show that none of the inequalities can be replaced by equality.

As with Hausdorff dimension, packing dimension permits the use of powerful measure theoretic techniques in its study. The introduction of packing measures (remarkably some 60 years after Hausdorff measures) has led to a greater understanding of the geometric measure theory of fractals, with packing measures behaving in a way that is 'dual' to Hausdorff measures in many respects. Indeed corresponding results for Hausdorff and packing measures are often presented side by side. Nevertheless, one cannot pretend that packing measures and dimensions are easy to work with or to calculate; the extra step (3.24) in their definition makes them more awkward to use than the Hausdorff analogues.

This situation is improved slightly by the equality of packing dimension and the modified upper box dimension. It is improved considerably for compact sets with 'local' dimension constant throughout—a situation that occurs frequently in practice, in particular in sets with some kind of self-similarity.

Corollary 3.9

Let $F \subset \mathbb{R}^n$ be compact and such that

$$\overline{\dim}_{\mathrm{B}}(F \cap V) = \overline{\dim}_{\mathrm{B}} F \tag{3.30}$$

for all open sets V that intersect F. Then $\dim_{\mathrm{P}} F = \overline{\dim}_{\mathrm{B}} F$.

Proof. This is immediate from Propositions 3.6 and 3.8. □

The nicest case, of course, is of fractals with equal Hausdorff and upper box dimensions, in which case equality holds throughout (3.29)—we shall see many such examples later on. However, even the much weaker condition $\dim_{\mathrm{H}} F = \dim_{\mathrm{P}} F$, though sometimes hard to prove, eases analysis of F.

3.5 Some other definitions of dimension

A wide variety of other definitions of dimension have been introduced, many of them only of limited applicability, but nonetheless useful in their context.

The special form of curves gives rise to the several definitions of dimension. We define a *curve* or *Jordan curve* C to be the image of an interval $[a, b]$ under a continuous bijection $f : [a, b] \to \mathbb{R}^n$. (Thus, we restrict attention to curves that are non-self-intersecting.) If C is a curve and $\delta > 0$, we define $M_\delta(C)$ to be the maximum number of points $x_0, x_1, \dots, x_m$, on the curve C, in that order, such that $|x_k - x_{k-1}| = \delta$ for $k = 1, 2, \dots, m$. Thus $(M_\delta(C) - 1)\delta$ may be thought of as the 'length' of the curve C measured using a pair of dividers with points set at a distance δ apart. The *divider dimension* is defined as

$$\lim_{\delta \to 0} \frac{\log M_\delta(C)}{-\log \delta} \tag{3.31}$$

assuming the limit exists (otherwise we may define upper and lower divider dimensions using upper and lower limits). It is easy to see that the divider dimension of a curve is at least equal to the box dimension (assuming that they both exist) and in simple self-similar examples, such as the von Koch curve, they are equal. The assertion that the coastline of Britain has dimension 1.2 is usually made with the divider dimension in mind—this empirical value comes from estimating the ratio in (3.31) for values of δ between about 20 m and 200 km.

A variant of Hausdorff dimension may be defined for curves by using intervals of the curves themselves as covering sets. Thus we look at $\inf\left\{\sum_{i=1}^{m} |f[t_{i-1}, t_i]|^s\right\}$ where the infimum is over all dissections $a = t_0 < t_1 < \cdots < t_m = b$ such that the diameters $|f([t_{i-1}, t_i])|$ are all at most δ. We let δ tend to 0 and deem the value of s at which this limit jumps from ∞ to 0 to be the dimension. For self-similar examples such as the von Koch curve, this equals the Hausdorff dimension, but for 'squeezed' curves, such as graphs of certain functions (see Chapter 11) we may get a somewhat larger value.

Sometimes, we are interested in the dimension of a fractal F that is the boundary of a set A. We can define the box dimension of F in the usual way, but sometimes it is useful to take special account of the distinction between A and its complement. Thus the following variation of the 's-dimensional content' definition of box dimension, in which we take the volume of the set of points within distance δ of F that are contained in A is sometimes useful. We define the *one-sided dimension* of the boundary F of a set A in $\mathbb{R}^n$ as

$$n - \lim_{\delta \to 0} \frac{\log \mathrm{vol}^n(F_\delta \cap A)}{\log \delta} \tag{3.32}$$

where F_δ is the δ-neighbourhood of F (compare Proposition 3.2). This definition has applications to the surface physics of solids where it is the volume very close to the surface that is important and also to partial differential equations in domains with fractal boundaries.

It is sometimes possible to define dimension in terms of the complement of a set. Suppose F is obtained by removal of a sequence of intervals $I_1, I_2, \ldots$ from, say, the unit interval [0, 1], as, for example, in the Cantor set construction. We may define a dimension as the number s_0 such that the series

$$\sum_{j=1}^{\infty} |I_j|^s \text{ converges if } s < s_0 \text{ and diverges if } s > s_0; \tag{3.33}$$

the number s_0 is called the *critical exponent* of the series. For the middle third Cantor set, this series is $\sum_{k=1}^{\infty} 2^{k-1} 3^{-ks}$, giving $s_0 = \log 2/\log 3$, equal to the Hausdorff and box dimensions in this case. In general, s_0 equals the upper box dimension of F.

Dimension prints provide an interesting variation on Hausdorff dimension of a rather different nature. Dimension prints may be thought of as a sort of 'fingerprint' that enables sets with differing characteristics to be distinguished, even though they may have the same Hausdorff dimension. In particular they reflect non-isotropic features of a set.

We restrict attention to subsets of the plane, in which case the dimension print will also be planar. The definition of dimension prints is very similar to that of Hausdorff dimension but coverings by rectangles are used with side lengths replacing diameters. Let U be a rectangle (the sides need not be parallel to the coordinate axes) and let $a(U) \geqslant b(U)$ be the lengths of the sides of U. Let s, t be non-negative numbers. For F a subset of $\mathbb{R}^2$, let

$$\mathcal{H}_\delta^{s,t}(F) = \inf\left\{\sum_i a(U_i)^s b(U_i)^t : \{U_i\} \text{ is a } \delta\text{-cover of } F \text{ by rectangles}\right\}.$$

In the usual way, we get measures of 'Hausdorff type', $\mathcal{H}^{s,t}$, by letting $\delta \to 0$:

$$\mathcal{H}^{s,t}(F) = \lim_{\delta\to 0} \mathcal{H}_\delta^{s,t}(F).$$

(Note that $\mathcal{H}^{s,0}$ is just a minor variant of s-dimensional Hausdorff measure where only rectangles are allowed in the δ-covers.) The *dimension print*, print F, of F is defined to be the set of non-negative pairs (s, t) for which $\mathcal{H}^{s,t}(F) > 0$.

Using standard properties of measures, it is easy to see that we have monotonicity

$$\text{print}\, F_1 \subset \text{print}\, F_2 \quad \text{if } F_1 \subset F_2 \tag{3.34}$$

and countable stability

$$\text{print}\left(\bigcup_{i=1}^{\infty} F_i\right) = \bigcup_{i=1}^{\infty} \text{print}\, F_i. \tag{3.35}$$

Moreover, if (s, t) is a point in print F and (s', t') satisfies

$$\begin{aligned} s' + t' &\leqslant s + t \\ t' &\leqslant t \end{aligned} \tag{3.36}$$

then (s', t') is also in print F.

Unfortunately, dimension prints are not particularly easy to calculate. We display a few known examples in figure 3.3. Notice that the Hausdorff dimension of a set is given by the point where the edge of its print intersects the x-axis.

Dimension prints are a useful and appealing extension of the idea of Hausdorff dimension. Notice how the prints in the last two cases distinguish between two sets of Hausdorff (or box) dimension $1\frac{1}{2}$, one of which is dust-like, the other stratified.

One disadvantage of dimension prints defined in this way is that they are *not* Lipschitz invariants. The straight line segment and smooth convex curve are bi-Lipschitz equivalent, but their prints are different. In the latter case the dimension

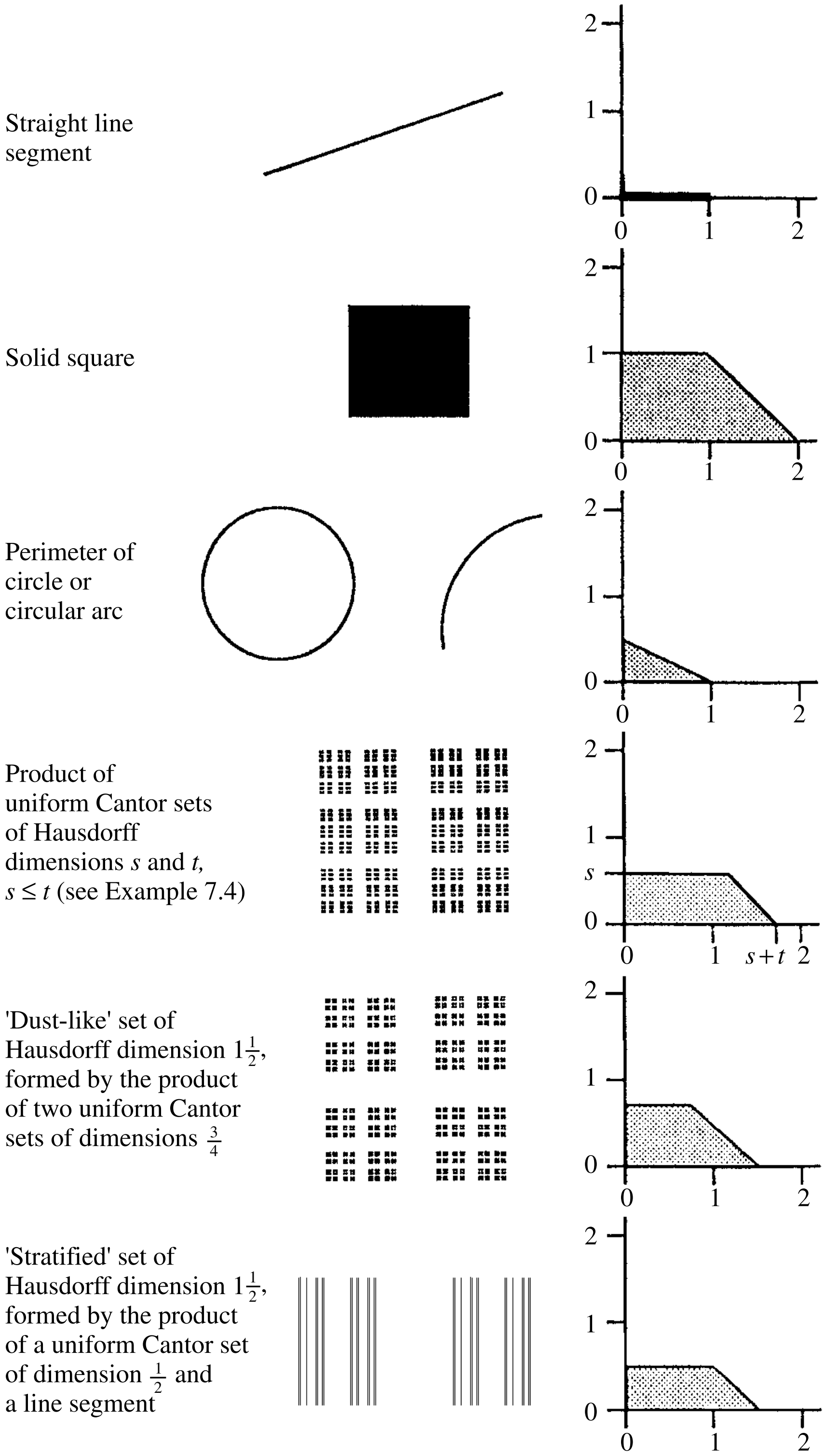

Straight line segment
Solid square
Perimeter of circle or circular arc
Product of uniform Cantor sets of Hausdorff dimensions s and t, $s \le t$ (see Example 7.4)
'Dust-like' set of Hausdorff dimension $1\frac{1}{2}$, formed by the product of two uniform Cantor sets of dimensions $\frac{3}{4}$
'Stratified' set of Hausdorff dimension $1\frac{1}{2}$, formed by the product of a uniform Cantor set of dimension $\frac{1}{2}$ and a line segment
0
1
2
s
$s+t$

Figure 3.3 A selection of dimension prints of plane sets

print takes into account the curvature. It would be possible to avoid this difficulty by redefining print F as the set of (s, t) such that $\mathcal{H}^{s,t}(F') > 0$ for all bi-Lipschitz images F' of F. This would restore Lipschitz invariance of the prints, but would add further complications to their calculation.

Of course, it would be possible to define dimension prints by analogy with box dimensions rather than Hausdorff dimensions, using covers by equal rectangles. Calculations still seem awkward.

3.6 Notes and references

Many different definitions of 'fractal dimension' are scattered throughout the mathematical literature. The origin of box dimension seems hard to trace—it seems certain that it must have been considered by the pioneers of Hausdorff measure and dimension, and was probably rejected as being less satisfactory from a mathematical viewpoint. Bouligand adapted the Minkowski content to non-integral dimensions in 1928, and the more usual definition of box dimension was given by Pontrjagin and Schnirelman in 1932.

Packing measures and dimensions are much more recent, introduced by Tricot (1982). Their similarities and contrasts to Hausdorff measures and dimensions have proved an important theoretical tool. Packing measures and box and packing dimensions are discussed in Mattila (1995) and Edgar (1998). Dimensions of curves are considered by Tricot (1995).

Dimension prints are an innovation of Rogers (1988, 1998).

Exercises

3.1 Let $f : F \to \mathbb{R}^n$ be a Lipschitz function. Show that $\underline{\dim}_{\mathrm{B}} f(F) \leqslant \underline{\dim}_{\mathrm{B}} F$ and $\overline{\dim}_{\mathrm{B}} f(F) \leqslant \overline{\dim}_{\mathrm{B}} F$. More generally, show that if f satisfies a Hölder condition $|f(x) - f(y)| \leqslant c|x - y|^\alpha$ where $c > 0$ and $0 < \alpha \leqslant 1$ then $\underline{\dim}_{\mathrm{B}} f(F) \leqslant \frac{1}{\alpha}\underline{\dim}_{\mathrm{B}} f(F)$.

3.2 Verify directly from the definitions that Equivalent definitions 3.1(i) and (iii) give the same values for box dimension.

3.3 Let F consist of those numbers in [0, 1] whose decimal expansions do not contain the digit 5. Find $\dim_{\mathrm{B}} F$, showing that this box dimension exists.

3.4 Verify that the Cantor dust depicted in figure 0.4 has box dimension 1 (take E_0 to have side length 1).

3.5 Use Equivalent definition 3.1(iv) to check that the upper box dimension of the von Koch curve is at most $\log 4/\log 3$ and 3.1(v) to check that the lower box dimension is at least this value.

3.6 Use convenient parts of Equivalent definition 3.1 to find the box dimension of the Sierpiński triangle in figure 0.3.

3.7 Let F be the middle third Cantor set. For $0 < \delta < 1$, find the length of the δ-neighbourhood F_δ of F, and hence find the box dimension of F using Proposition 3.2.

3.8 Construct a set F for which $\underline{\dim}_{\mathrm{B}} F < \overline{\dim}_{\mathrm{B}} F$. (Hint: let $k_n = 10^n$, and adapt the Cantor set construction by deleting, at the kth stage, the middle $\frac{1}{3}$ of intervals if $k_{2n} < k \leqslant k_{2n+1}$, but the middle $\frac{3}{5}$ of intervals if $k_{2n-1} < k \leqslant k_{2n}$.)

3.9 Verify that $\overline{\dim}_{\mathrm{B}}(E \cup F) = \max\{\overline{\dim}_{\mathrm{B}} E, \overline{\dim}_{\mathrm{B}} F\}$ for bounded $E, F \subset \mathbb{R}$.

3.10 Find subsets E and F of $\mathbb{R}$ such that $\underline{\dim}_{\mathrm{B}}(E \cup F) > \max\{\underline{\dim}_{\mathrm{B}} E, \underline{\dim}_{\mathrm{B}} F\}$. (Hint: consider two sets of the form indicated in Exercise 3.8.)

3.11 What are the Hausdorff and box dimensions of the set $\left\{0, 1, \frac{1}{4}, \frac{1}{9}, \frac{1}{16}, \dots\right\}$?

3.12 Find two disjoint Borel subsets E and F of $\mathbb{R}$ such that $\mathcal{P}_0^s(E \cup F) \neq \mathcal{P}_0^s(E) + \mathcal{P}_0^s(F)$.

3.13 What is the packing dimension of the von Koch curve?

3.14 Find the divider dimension (3.31) of the von Koch curve.

3.15 Show that the divider dimension (3.31) of a curve is greater than or equal to its box dimension, assuming that they both exist.

3.16 Let $0 < \lambda < 1$ and let F be the 'middle λ Cantor set' obtained by repeated removal of the middle proportion λ from intervals. Show that the dimension of F defined by (3.33) in terms of removed intervals equals the Hausdorff and box dimensions of F.

3.17 Verify properties (3.34)–(3.36) of dimension prints. Given an example of a set with a non-convex dimension print.

Chapter 4 Techniques for calculating dimensions

A direct attempt at calculating the dimensions, in particular the Hausdorff dimension, of almost any set will convince the reader of the practical limitations of working from the definitions. Rigorous dimension calculations often involve pages of complicated manipulations and estimates that provide little intuitive enlightenment.

In this chapter we bring together some of the basic techniques that are available for dimension calculations. Other methods, that are applicable in more specific cases, will be found throughout the book.

4.1 Basic methods

As a general rule, we get upper bounds for Hausdorff measures and dimensions by finding effective coverings by small sets, and lower bounds by putting measures or mass distributions on the set. For most fractals 'obvious' upper estimates of dimension may be obtained using natural coverings by small sets.

Proposition 4.1

Suppose F can be covered by n_k sets of diameter at most δ_k with $\delta_k \to 0$ as $k \to \infty$. Then

$$\dim_{\mathrm{H}} F \leqslant \underline{\dim}_{\mathrm{B}} F \leqslant \varliminf_{k\to\infty} \frac{\log n_k}{-\log \delta_k}.$$

Moreover, if $n_k\delta_k^s$ remains bounded as $k \to \infty$, then $\mathcal{H}^s(F) < \infty$. If $\delta_k \to 0$ but $\delta_{k+1} \geqslant c\delta_k$ for some $0 < c < 1$, then

$$\overline{\dim}_{\mathrm{B}} F \leqslant \varlimsup_{k\to\infty} \frac{\log n_k}{-\log \delta_k}.$$

Proof. The inequalities for the box-counting dimension are immediate from the definitions and the remark at (3.14). That $\dim_{\mathrm{H}} F \leqslant \underline{\dim}_{\mathrm{B}} F$ is in (3.17), and if

Fractal Geometry: Mathematical Foundations and Application. Second Edition Kenneth Falconer

ISBNs: 0-470-84861-8 (HB); 0-470-84862-6 (PB)

$n_k\delta_k^s$ is bounded then $\mathcal{H}^s_{\delta_k}(F) \leqslant n_k\delta_k^s$, so $\mathcal{H}^s_{\delta_k}(F)$ tends to a finite limit $\mathcal{H}^s(F)$ as $k \to \infty$. □

Thus, as we have seen already (Example 2.7), in the case of the middle third Cantor set the natural coverings by 2^k intervals of length 3^{-k} give $\dim_{\mathrm{H}} F \leqslant \underline{\dim}_{\mathrm{B}} F \leqslant \overline{\dim}_{\mathrm{B}} F \leqslant \log 2/\log 3$.

Surprisingly often, the 'obvious' upper bound for the Hausdorff dimension of a set turns out to be the actual value. However, demonstrating this can be difficult. To obtain an upper bound it is enough to evaluate sums of the form $\sum |U_i|^s$ for *specific* coverings $\{U_i\}$ of F, whereas for a lower bound we must show that $\sum |U_i|^s$ is greater than some positive constant for *all* δ-coverings of F. Clearly an enormous number of such coverings are available. In particular, when working with Hausdorff dimension as opposed to box dimension, consideration must be given to covers where some of the U_i are very small and others have relatively large diameter—this prohibits sweeping estimates for $\sum |U_i|^s$ such as those available for upper bounds.

One way of getting around these difficulties is to show that no *individual* set U can cover too much of F compared with its size measured as $|U|^s$. Then if $\{U_i\}$ covers the whole of F the sum $\sum |U_i|^s$ cannot be too small. The usual way to do this is to concentrate a suitable mass distribution μ on F and compare the mass $\mu(U)$ covered by U with $|U|^s$ for each U. (Recall that a mass distribution on F is a measure with support contained in F such that $0 < \mu(F) < \infty$, see Section 1.3.)

Mass distribution principle 4.2

Let μ be a mass distribution on F and suppose that for some s there are numbers $c > 0$ and $\varepsilon > 0$ such that

$$\mu(U) \leqslant c|U|^s \tag{4.1}$$

for all sets U with $|U| \leqslant \varepsilon$. Then $\mathcal{H}^s(F) \geqslant \mu(F)/c$ and

$$s \leqslant \dim_{\mathrm{H}} F \leqslant \underline{\dim}_{\mathrm{B}} F \leqslant \overline{\dim}_{\mathrm{B}} F.$$

Proof. If $\{U_i\}$ is any cover of F then

$$0 < \mu(F) \leqslant \mu\left(\bigcup_i U_i\right) \leqslant \sum_i \mu(U_i) \leqslant c \sum_i |U_i|^s \tag{4.2}$$

using properties of a measure and (4.1).

Taking infima, $\mathcal{H}^s_\delta(F) \geqslant \mu(F)/c$ if δ is small enough, so $\mathcal{H}^s(F) \geqslant \mu(F)/c$. Since $\mu(F) > 0$ we get $\dim_{\mathrm{H}} F \geqslant s$. □

Notice that the conclusion $\mathcal{H}^s(F) \geqslant \mu(F)/c$ remains true if μ is a mass distribution on $\mathbb{R}^n$ and F is any subset.

The Mass distribution principle 4.2 gives a quick lower estimate for the Hausdorff dimension of the middle third Cantor set F (figure 0.1). Let μ be the natural mass distribution on F, so that each of the 2^k kth level intervals of length 3^{-k} in E_k in the construction of F carry a mass 2^{-k}. (We imagine that we start with unit mass on E_0 and repeatedly divide the mass on each interval of E_k between its two subintervals in E_{k+1}; see Proposition 1.7.) Let U be a set with $|U| < 1$ and let k be the integer such that $3^{-(k+1)} \leqslant |U| < 3^{-k}$. Then U can intersect at most one of the intervals of E_k, so

$$\mu(U) \leqslant 2^{-k} = (3^{\log 2/\log 3})^{-k} = (3^{-k})^{\log 2/\log 3} \leqslant (3|U|)^{\log 2/\log 3}$$

and hence $\mathcal{H}^{\log 2/\log 3}(F) \geqslant 3^{-\log 2/\log 3} = \frac{1}{2}$ by the mass distribution principle, giving $\dim_{\mathrm{H}} F \geqslant \log 2/\log 3$.

Example 4.3

Let $F_1 = F \times [0, 1] \subset \mathbb{R}^2$ *be the product of the middle third Cantor set F and the unit interval. Then, setting* $s = 1 + \log 2/\log 3$, *we have* $\dim_{\mathrm{B}} F_1 = \dim_{\mathrm{H}} F_1 = s$, *with* $0 < \mathcal{H}^s(F_1) < \infty$.

Calculation. For each k, there is a covering of F by 2^k intervals of length 3^{-k}. A column of 3^k squares of side 3^{-k} (diameter $3^{-k}\sqrt{2}$) covers the part of F_1 above each such interval, so taking these all together, F_1 may be covered by $2^k 3^k$ squares of side 3^{-k}. Thus $\mathcal{H}^s_{3^{-k}\sqrt{2}}(F_1) \leqslant 3^k 2^k (3^{-k}\sqrt{2})^s = (3 \cdot 2 \cdot 3^{-1-\log 2/\log 3})^k 2^{s/2} = 2^{s/2}$, so $\mathcal{H}^s(F_1) \leqslant 2^{s/2}$ and $\dim_{\mathrm{H}} F_1 \leqslant \underline{\dim}_{\mathrm{B}} F_1 \leqslant \overline{\dim}_{\mathrm{B}} F_1 \leqslant s$.

We define a mass distribution μ on F_1 by taking the natural mass distribution on F described above (each kth level interval of F of side 3^{-k} having mass 2^{-k}) and 'spreading it' uniformly along the intervals above F. Thus if U is a rectangle, with sides parallel to the coordinate axes, of height $h \leqslant 1$, above a kth level interval of F, then $\mu(U) = h2^{-k}$. Any set U is contained in a square of side $|U|$ with sides parallel to the coordinate axes. If $3^{-(k+1)} \leqslant |U| < 3^{-k}$ then U lies above at most one kth level interval of F of side 3^{-k}, so

$$\mu(U) \leqslant |U| 2^{-k} \leqslant |U| 3^{-k \log 2/\log 3} \leqslant |U|(3|U|)^{\log 2/\log 3} = 3^{\log 2/\log 3}|U|^s = 2|U|^s.$$

By the Mass distribution principle 4.2, $\mathcal{H}^s(F_1) > \frac{1}{2}$. □

Note that the method of Examples 4.2 and 4.3 extends to a wide variety of self-similar sets. Indeed, Theorem 9.3 may be regarded as a generalization of this calculation.

Notice that in this example the dimension of the product of two sets equals the sum of the dimensions of the sets. We study this in greater depth in Chapter 7.

The following *general construction* of a subset of $\mathbb{R}$ may be thought of as a generalization of the Cantor set construction. Let $[0, 1] = E_0 \supset E_1 \supset E_2 \supset \ldots$ be a decreasing sequence of sets, with each E_k a union of a finite number of disjoint closed intervals (called kth *level basic intervals*), with each interval of E_k containing at least two intervals of E_{k+1}, and the maximum length of kth level intervals tending to 0 as $k \to \infty$. Then the set

$$F = \bigcap_{k=0}^{\infty} E_k \tag{4.3}$$

is a totally disconnected subset of [0, 1] which is generally a fractal (figure 4.1).

Obvious upper bounds for the dimension of F are available by taking the intervals of E_k as covering intervals, for each k, but, as usual, lower bounds are harder to find. Note that, in the following examples, the upper estimates for $\dim_{\mathrm{H}} F$ depend on the number and size of the basic intervals, whilst the lower estimates depend on their spacing. For these to be equal, the $(k+1)$th level intervals must be 'nearly uniformly distributed' inside the kth level intervals.

Example 4.4

Let s be a number strictly between 0 *and* 1. *Assume that in the general construction* (4.3) *for each kth level interval I, the* $(k+1)$*th level intervals* $I_1, \ldots, I_m$ $(m \geqslant 2)$ *contained in I are of equal length and equally spaced, the lengths being given by*

$$|I_i|^s = \frac{1}{m}|I|^s \qquad (1 \leqslant i \leqslant m) \tag{4.4}$$

with the left-hand ends of I_1 *and I coinciding, and the right-hand ends of* I_m *and I coinciding. Then* $\dim_{\mathrm{H}} F = s$ *and* $0 < \mathcal{H}^s(F) < \infty$. (*Notice that m may be different for different intervals I in the construction, so that the kth level intervals may have widely differing lengths.*)

Calculation. With I, I_i, as above,

$$|I|^s = \sum_{i=1}^{m} |I_i|^s. \tag{4.5}$$

Applying this inductively to the kth level intervals for successive k, we have, for each k, that $1 = \sum |I_i|^s$, where the sum is over all the kth level intervals I_i.

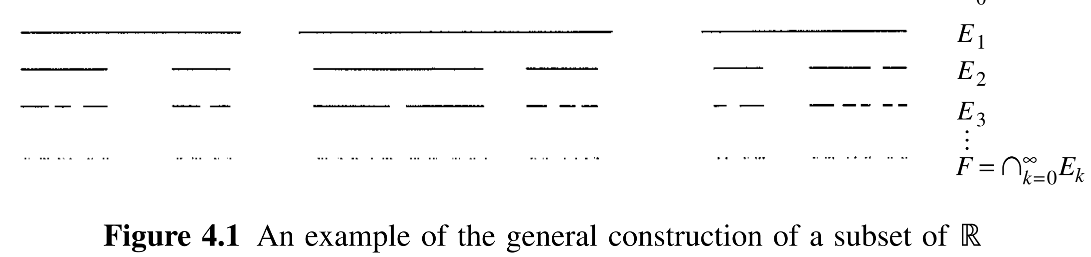


Figure 4.1 An example of the general construction of a subset of $\mathbb{R}$

The kth level intervals cover F; since the maximum interval length tends to 0 as $k \to \infty$, we have $\mathcal{H}^s_\delta(F) \leqslant 1$ for sufficiently small δ, giving $\mathcal{H}^s(F) \leqslant 1$.

Now distribute a mass μ on F in such a way that $\mu(I) = |I|^s$ whenever I is any level k interval. Thus, starting with unit mass on [0, 1] we divide this equally between each level 1 interval, the mass on each of these intervals being divided equally between each level 2 subinterval, and so on; see Proposition 1.7. Equation (4.5) ensures that we get a mass distribution on F with $\mu(I) = |I|^s$ for every basic interval. We estimate $\mu(U)$ for an interval U with endpoints in F. Let I be the smallest basic interval that contains U; suppose that I is a kth level interval, and let $I_i, \ldots, I_m$ be the $(k+1)$th level intervals contained in I. Then U intersects a number $j \geqslant 2$ of the I_i, otherwise U would be contained in a smaller basic interval. The spacing between consecutive I_i is

$$\begin{aligned}(|I| - m|I_i|)/(m-1) &= |I|(1 - m|I_i|/|I|)/(m-1)\\ &= |I|(1 - m^{1-1/s})/(m-1)\\ &\geqslant c_s|I|/m\end{aligned}$$

using (4.4) and that $m \geqslant 2$ and $0 < s < 1$, where $c_s = (1 - 2^{1-1/s})$. Thus

$$|U| \geqslant \frac{j-1}{m}c_s|I| \geqslant \frac{j}{2m}c_s|I|.$$

By (4.4)

$$\begin{aligned}\mu(U) &\leqslant j\mu(I_i) = j|I_i|^s = \frac{j}{m}|I|^s\\ &\leqslant 2^s c_s^{-s}\left(\frac{j}{m}\right)^{1-s}|U|^s \leqslant 2^s c_s^{-s}|U|^s.\end{aligned} \tag{4.6}$$

This is true for any interval U with endpoints in F, and so for any set U (by applying (4.6) to the smallest interval containing $U \cap F$). By the Mass distribution principle 4.2, $\mathcal{H}^s(F) > 0$. □

A more careful estimate of $\mu(U)$ in Example 4.4 leads to $\mathcal{H}^s(F) = 1$.

We call the sets obtained when m is kept constant throughout the construction of Example 4.4 *uniform Cantor sets*; see figure 4.2. These provide a natural generalization of the middle third Cantor set.

Example 4.5. Uniform Cantor sets

Let $m \geqslant 2$ be an integer and $0 < r < 1/m$. Let F be the set obtained by the construction in which each basic interval I is replaced by m equally spaced subintervals of lengths $r|I|$, the ends of I coinciding with the ends of the extreme subintervals. Then $\dim_{\mathrm{H}} F = \dim_{\mathrm{B}} F = \log m/-\log r$, *and* $0 < \mathcal{H}^{\log m/-\log r}(F) < \infty$.

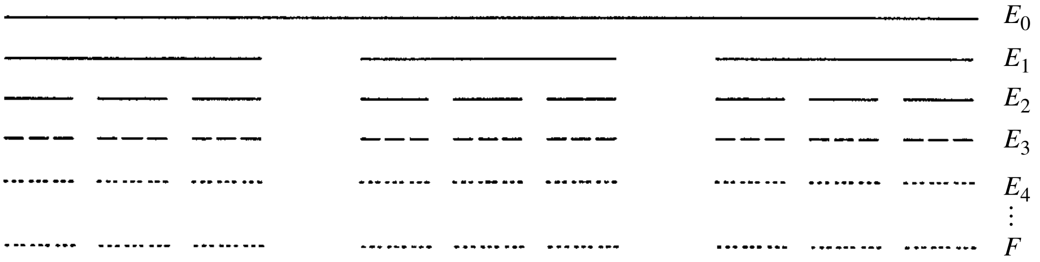


Figure 4.2 A uniform Cantor set (Example 4.5) with $m = 3, r = \frac{4}{15}$, $\dim_{\mathrm{H}} F = \dim_{\mathrm{B}} F = \log 3/-\log\frac{4}{15} = 0.831\ldots$

Calculation. The set F is obtained on taking m constant and $s = \log m/(-\log r)$ in Example 4.4. Equation (4.4) becomes $(r|I|)^s = (1/m)|I|^s$, which is satisfied identically, so $\dim_{\mathrm{H}} F = s$. For the box dimension, note that F is covered by the m^k kth level intervals of length r^{-k} for each k, leading to $\overline{\dim}_{\mathrm{B}} F \leqslant \log m/-\log r$ in the usual way. □

The *middle λ Cantor set* is obtained by repeatedly removing a proportion $0 < \lambda < 1$ from the middle of intervals, starting with the unit interval. This is a special case of a uniform Cantor set, having $m = 2$ and $r = \frac{1}{2}(1-\lambda)$ and thus Hausdorff and box dimensions $\log 2/\log(2/(1-\lambda))$.

The next example is another case of the general construction.

Example 4.6

Suppose in the general construction (4.3) *each* $(k-1)$*th level interval contains at least* $m_k \geqslant 2$ *kth level intervals* $(k = 1, 2, \ldots)$ *which are separated by gaps of at least* ε_k*, where* $0 < \varepsilon_{k+1} < \varepsilon_k$ *for each k. Then*

$$\dim_{\mathrm{H}} F \geqslant \varliminf_{k\to\infty} \frac{\log(m_1 \cdots m_{k-1})}{-\log(m_k \varepsilon_k)}. \tag{4.7}$$

Calculation. We may assume that the right hand side of (4.7) is positive, otherwise (4.7) is obvious. We may assume that each $(k-1)$th level interval contains exactly m_k kth level intervals; if not we may throw out excess intervals to get smaller sets E_k and F for which this is so. We may define a mass distribution μ on F by assigning a mass of $(m_1 \cdots m_k)^{-1}$ to each of the $m_1 \cdots m_k$ kth level intervals.

Let U be an interval with $0 < |U| < \varepsilon_1$; we estimate $\mu(U)$. Let k be the integer such that $\varepsilon_k \leqslant |U| < \varepsilon_{k-1}$. The number of kth level intervals that intersect U is

(i) at most m_k since U intersects at most one $(k-1)$th level interval
(ii) at most $(|U|/\varepsilon_k) + 1 \leqslant 2|U|/\varepsilon_k$ since the kth level intervals have gaps of at least ε_k between them.

Each kth level interval supports mass $(m_1 \cdots m_k)^{-1}$ so that

$$\begin{aligned}\mu(U) &\leqslant (m_1 \cdots m_k)^{-1} \min\{2|U|/\varepsilon_k, m_k\}\\ &\leqslant (m_1 \cdots m_k)^{-1}(2|U|/\varepsilon_k)^s m_k^{1-s}\end{aligned}$$

for every $0 \leqslant s \leqslant 1$.

Hence
$$\frac{\mu(U)}{|U|^s} \leqslant \frac{2^s}{(m_1 \cdots m_{k-1}) m_k^s \varepsilon_k^s}.$$

If

$$s < \varliminf_{k\to\infty} \log(m_1 \cdots m_{k-1})/-\log(m_k\varepsilon_k)$$

then $(m_1 \cdots m_{k-1}) m_k^s \varepsilon_k^s > 1$ for large k, so $\mu(U) \leqslant 2^s |U|^s$, and $\dim_{\mathrm{H}} F \geqslant s$ by Principle 4.2, giving (4.7). □

Now suppose that in Example 4.6 the kth level intervals are all of length δ_k, and that each $(k-1)$th level interval contains exactly m_k kth level intervals, which are 'roughly equally spaced' in the sense that $m_k\varepsilon_k \geqslant c\delta_{k-1}$, where $c > 0$ is a constant. Then (4.7) becomes

$$\dim_{\mathrm{H}} F \geqslant \varliminf_{k\to\infty} \frac{\log(m_1 \cdots m_{k-1})}{-\log c - \log \delta_{k-1}} = \varliminf_{k\to\infty} \frac{\log(m_1 \cdots m_{k-1})}{-\log \delta_{k-1}}.$$

But E_{k-1} comprises $m_1 \cdots m_{k-1}$ intervals of length δ_{k-1}, so this expression equals the upper bound for $\dim_{\mathrm{H}} F$ given by Proposition 4.1. Thus in the situation where the intervals are well spaced, we get equality in (4.7).

Examples of the following form occur in number theory; see Section 10.3.

Example 4.7

Fix $0 < s < 1$ and let $n_0, n_1, n_2, \ldots$ be a rapidly increasing sequence of integers, say $n_{k+1} \geqslant \max\{n_k^k, 4n_k^{1/s}\}$ for each k. For each k let $H_k \subset \mathbb{R}$ consist of equally spaced equal intervals of lengths $n_k^{-1/s}$ with the midpoints of consecutive intervals distance n_k^{-1} apart. Then writing $F = \bigcap_{k=1}^{\infty} H_k$, we have $\dim_{\mathrm{H}} F = s$.

Calculation. Since $F \subset H_k$ for each k, the set $F \cap [0, 1]$ is contained in at most $n_k + 1$ intervals of length $n_k^{-1/s}$, so Proposition 4.1 gives $\dim_{\mathrm{H}}(F \cap [0, 1]) \leqslant \varliminf_{k\to\infty} \log(n_k + 1)/-\log n_k^{-1/s} = s$. Similarly, $\dim_{\mathrm{H}}(F \cap [n, n+1]) \leqslant s$ for all $n \in \mathbb{Z}$, so $\dim_{\mathrm{H}} F \leqslant s$ as a countable union of such sets.

Now let $E_0 = [0, 1]$ and, for $k \geqslant 1$, let E_k consist of the intervals of H_k that are completely contained in E_{k-1}. Then each interval I of E_{k-1} contains at least $n_k|I| - 2 \geqslant n_k n_{k-1}^{-1/s} - 2 \geqslant 2$ intervals of E_k, which are separated by gaps of at least $n_k^{-1} - n_k^{-1/s} \geqslant \frac{1}{2}n_k^{-1}$ if k is large enough. Using Example 4.6, and noting that setting $m_k = n_k n_{k-1}^{-1/s}$ rather than $m_k = n_k n_{k-1}^{-1/s} - 2$ does not affect the limit,

$$\begin{aligned}\dim_{\mathrm{H}}(F \cap [0,1]) \geqslant \dim_{\mathrm{H}} \bigcap_{k=1}^{\infty} E_k &\geqslant \varliminf_{k\to\infty} \frac{\log((n_1 \cdots n_{k-2})^{1-1/s} n_{k-1})}{-\log(n_k n_{k-1}^{-1/s} \frac{1}{2} n_k^{-1})}\\ &= \varliminf_{k\to\infty} \frac{\log(n_1 \cdots n_{k-2})^{1-1/s} + \log n_{k-1}}{\log 2 + (\log n_{k-1})/s}.\end{aligned}$$

Provided that n_k is sufficiently rapidly increasing, the terms in $\log n_{k-1}$ in the numerator and denominator of this expression are dominant, so that $\dim_{\mathrm{H}} F \geqslant \dim_{\mathrm{H}}(F \cap [0,1]) \geqslant s$, as required. □

Although the Mass distribution principle 4.2 is based on a simple idea, we have seen that it can be very useful in finding Hausdorff and box dimensions. We now develop some important variations of the method.

It is enough for condition (4.1) to hold just for sufficiently small balls centred at each point of F. This is expressed in Proposition 4.9(a). Although mass distribution methods for upper bounds are required far less frequently, we include part (b) because it is, in a sense, dual to (a). Note that density expressions, such as $\lim_{r\to 0} \mu(B(x,r))/r^s$ play a major role in the study of local properties of fractals—see Chapter 5. (Recall that $B(x,r)$ is the closed ball of centre x and radius r.)

We require the following covering lemma in the proof of Proposition 4.9(b).

Covering lemma 4.8

Let $\mathcal{C}$ be a family of balls contained in some bounded region of $\mathbb{R}^n$. Then there is a (finite or countable) disjoint subcollection $\{B_i\}$ such that

$$\bigcup_{B\in\mathcal{C}} B \subset \bigcup_i \tilde{B}_i \tag{4.8}$$

where $\tilde{B}_i$ is the closed ball concentric with B_i and of four times the radius.

Proof. For simplicity, we give the proof when $\mathcal{C}$ is a finite family; the basic idea is the same in the general case. We select the $\{B_i\}$ inductively. Let B_1 be a ball in $\mathcal{C}$ of maximum radius. Suppose that $B_1, \ldots, B_{k-1}$ have been chosen. We take B_k to be the largest ball in $\mathcal{C}$ (or one of the largest) that does not intersect $B_1, \ldots, B_{k-1}$. The process terminates when no such ball remains. Clearly the balls selected are disjoint; we must check that (4.8) holds. If $B \in \mathcal{C}$, then either $B = B_i$ for some i, or B intersects one of the B_i with $|B_i| \geqslant |B|$; if this were not the case, then B would have been chosen instead of the first ball B_k with $|B_k| < |B|$. Either way, $B \subset \tilde{B}_i$, so we have (4.8). (It is easy to see that the result remains true taking $\tilde{B}_i$ as the ball concentric with B_i and of $3+\varepsilon$ times the radius, for any $\varepsilon > 0$; if $\mathcal{C}$ is finite we may take $\varepsilon = 0$.) □

Proposition 4.9

Let μ be a mass distribution on $\mathbb{R}^n$, let $F \subset \mathbb{R}^n$ be a Borel set and let $0 < c < \infty$ be a constant.

(*a*) *If $\overline{\lim}_{r\to 0}\mu(B(x,r))/r^s < c$ for all $x \in F$ then $\mathcal{H}^s(F) \geqslant \mu(F)/c$*
(*b*) *If $\overline{\lim}_{r\to 0}\mu(B(x,r))/r^s > c$ for all $x \in F$ then $\mathcal{H}^s(F) \leqslant 2^s\mu(\mathbb{R}^n)/c$.*

Proof

(*a*) For each $\delta > 0$ let

$$F_\delta = \{x \in F : \mu(B(x,r)) < cr^s \text{ for all } 0 < r \leqslant \delta\}.$$

Let $\{U_i\}$ be a δ-cover of F and thus of F_δ. For each U_i containing a point x of F_δ, the ball B with centre x and radius $|U_i|$ certainly contains U_i. By definition of F_δ,

$$\mu(U_i) \leqslant \mu(B) < c|U_i|^s$$

so that

$$\mu(F_\delta) \leqslant \sum_i\{\mu(U_i) : U_i \text{ intersects } F_\delta\} \leqslant c\sum_i |U_i|^s.$$

Since $\{U_i\}$ is any δ-cover of F, it follows that $\mu(F_\delta)\leqslant c\mathcal{H}^s_\delta(F)\leqslant c\mathcal{H}^s(F)$. But F_δ increases to F as δ decreases to 0, so $\mu(F) \leqslant c\mathcal{H}^s(F)$ by (1.7).

(*b*) For simplicity, we prove a weaker version of (*b*) with 2^s replaced by 8^s, but the basic idea is similar. Suppose first that F is bounded. Fix $\delta > 0$ and let $\mathcal{C}$ be the collection of balls

$$\{B(x,r) : x \in F, 0 < r \leqslant \delta \text{ and } \mu(B(x,r)) > cr^s\}.$$

Then by the hypothesis of (*b*) $F \subset \bigcup_{B\in\mathcal{C}} B$. Applying the Covering lemma 4.8 to the collection $\mathcal{C}$, there is a sequence of disjoint balls $B_i \in \mathcal{C}$ such that $\bigcup_{B\in\mathcal{C}} B \subset \bigcup_i \tilde{B}_i$ where $\tilde{B}_i$ is the ball concentric with B_i but of four times the radius. Thus $\{\tilde{B}_i\}$ is an 8δ-cover of F, so

$$\begin{aligned}\mathcal{H}^s_{8\delta}(F) &\leqslant \sum_i |\tilde{B}_i|^s \leqslant 4^s\sum_i |B_i|^s\\ &\leqslant 8^s c^{-1}\sum_i \mu(B_i) \leqslant 8^s c^{-1}\mu(\mathbb{R}^n).\end{aligned}$$

Letting $\delta \to 0$, we get $\mathcal{H}^s(F) \leqslant 8^s c^{-1}\mu(\mathbb{R}^n) < \infty$. Finally, if F is unbounded and $\mathcal{H}^s(F) > 8^s c^{-1}\mu(\mathbb{R}^n)$, the $\mathcal{H}^s$-measure of some bounded subset of F will also exceed this value, contrary to the above. □

Note that it is immediate from Proposition 4.9 that if $\lim_{r\to 0} \log \mu(B(x,r))/\log r = s$ for all $x \in F$ then $\dim_{\mathrm{H}} F = s$.

Applications of Proposition 4.9 will occur throughout the book.

We conclude this section by a reminder that these calculations can be used in conjunction with the basic properties of dimensions discussed in Chapters 2 and 3. For example, since $f(x) = x^2$ is Lipschitz on [0, 1] and bi-Lipschitz on $[\frac{2}{3}, 1]$, it follows that $\dim_{\mathrm{H}}\{x^2 : x \in C\} = \dim_{\mathrm{H}} f(C) = \log 2/\log 3$, where C is the middle third Cantor set.

4.2 Subsets of finite measure

This section may seem out of place in a chapter about finding dimensions. However, Theorem 4.10 is required for the important potential theoretic methods developed in the following section. Sets of infinite measure can be awkward to work with, and reducing them to sets of positive finite measure can be a very useful simplification.

Theorem 4.10 guarantees that any (Borel) set F with $\mathcal{H}^s(F) = \infty$ contains a subset E with $0 < \mathcal{H}^s(E) < \infty$ (i.e. with E an s-set). At first, this might seem obvious—just shave pieces off F until what remains has positive finite measure. Unfortunately it is not quite this simple—it is possible to jump from infinite measure to zero measure without passing through any intermediate value. Stating this in mathematical terms, it is possible to have a decreasing sequence of sets $E_1 \supset E_2 \supset \dots$ with $\mathcal{H}^s(E_k) = \infty$ for all k, but with $\mathcal{H}^s(\bigcap_{k=1}^{\infty} E_k) = 0$. (For a simple example, take $E_k = [0, 1/k] \subset \mathbb{R}$ and $0 < s < 1$.) To prove the theorem we need to look rather more closely at the structure of Hausdorff measures. Readers mainly concerned with applications may prefer to omit the proof!

Theorem 4.10

Let F be a Borel subset of $\mathbb{R}^n$ with $0 < \mathcal{H}^s(F) \leqslant \infty$. Then there is a compact set $E \subset F$ such that $0 < \mathcal{H}^s(E) < \infty$.

**Sketch of proof.* The complete proof of this is complicated. We indicate the ideas involved in the case where F is a compact subset of $[0, 1) \subset \mathbb{R}$ and $0 < s < 1$.

We work with the net measures $\mathcal{M}^s$ which are defined in (2.17)–(2.18) using the binary intervals $[r2^{-k}, (r+1)2^{-k})$ and are related to Hausdorff measure by (2.19). We define inductively a decreasing sequence $E_0 \supset E_1 \supset E_2 \supset \dots$ of compact subsets of F. Let $E_0 = F$. For $k \geqslant 0$ we define E_{k+1} by specifying its intersection with each binary interval I of length 2^{-k}. If $\mathcal{M}^s_{2^{-(k+1)}}(E_k \cap I) \leqslant 2^{-sk}$ we let $E_{k+1} \cap I = E_k \cap I$. Then

$$\mathcal{M}^s_{2^{-(k+1)}}(E_{k+1} \cap I) = \mathcal{M}^s_{2^{-k}}(E_k \cap I) \tag{4.9}$$

since using I itself as a covering interval in calculating $\mathcal{M}^s_{2^{-k}}$ gives an estimate at least as large as using shorter binary intervals. On the other hand, if $\mathcal{M}^s_{2^{-(k+1)}}(E_{k+1} \cap I) > 2^{-sk}$ we take $E_{k+1} \cap I$ to be a compact subset of $E_k \cap I$ with $\mathcal{M}^s_{2^{-(k+1)}}(E_{k+1} \cap I) = 2^{-sk}$. Such a subset exists since $\mathcal{M}^s_{2^{-(k+1)}}(E_k \cap I \cap [0, u])$ is finite and continuous in u. (This is why we need to work with the $\mathcal{M}^s_\delta$ rather than $\mathcal{M}^s$.) Since $\mathcal{M}^s_{2^{-k}}(E_k \cap I) = 2^{-sk}$, (4.9) again holds. Summing (4.9) over all binary intervals of length 2^{-k} we get

$$\mathcal{M}^s_{2^{-(k+1)}}(E_{k+1}) = \mathcal{M}^s_{2^{-k}}(E_k). \tag{4.10}$$

Repeated application of (4.10) gives $\mathcal{M}^s_{2^{-k}}(E_k) = \mathcal{M}^s_1(E_0)$ for all k. Let E be the compact set $\bigcap_{k=0}^{\infty} E_k$. Taking the limit as $k \to \infty$ gives $\mathcal{M}^s(E) = \mathcal{M}^s_1(E_0)$ (this step needs some justification). The covering of $E_0 = F$ by the single interval $[0, 1)$ gives $\mathcal{M}^s(E) = \mathcal{M}^s_1(E_0) \leqslant 1$. Since $\mathcal{M}^s(E_0) \geqslant \mathcal{H}^s(E_0) > 0$ we have $\mathcal{M}^s_{2^{-k}}(E_0) > 0$ if k is large enough. Thus either $\mathcal{M}^s(E) = \mathcal{M}^s_1(E_0) \geqslant 2^{-ks}$, or $\mathcal{M}^s_1(E_0) < 2^{-ks}$ in which case $\mathcal{M}^s(E) = \mathcal{M}^s_1(E_0) = \mathcal{M}^s_{2^{-k}}(E_0) > 0$. Thus $0 < \mathcal{M}^s(E) < \infty$, and the theorem follows from (2.19). □

A number of results, for example those in Chapter 5, apply only to s-sets, i.e. sets with $0 < \mathcal{H}^s(F) < \infty$. One way of approaching s-dimensional sets with $\mathcal{H}^s(F) = \infty$ is to use Theorem 4.10 to extract a subset of positive finite measure, to study its properties as an s-set, and then to interpret these properties in the context of the larger set F. Similarly, if $0 < s < t$, any set F of Hausdorff dimension t has $\mathcal{H}^s(F) = \infty$ and so contains an s-set.

The following proposition which follows from Proposition 4.9, leads to an extension of Theorem 4.10.

Proposition 4.11

Let F be a Borel set satisfying $0 < \mathcal{H}^s(F) < \infty$. There is a constant b and a compact set $E \subset F$ with $\mathcal{H}^s(E) > 0$ such that

$$\mathcal{H}^s(E \cap B(x, r)) \leqslant br^s \tag{4.11}$$

for all $x \in \mathbb{R}^n$ and $r > 0$.

Proof. In Proposition 4.9(*b*) take μ as the restriction of $\mathcal{H}^s$ to F, i.e. $\mu(A) = \mathcal{H}^s(F \cap A)$. Then, if

$$F_1 = \left\{x \in \mathbb{R}^n : \overline{\lim_{r\to 0}}\, \mathcal{H}^s(F \cap B(x, r))/r^s > 2^{1+s}\right\}$$

it follows that $\mathcal{H}^s(F_1) \leqslant 2^s 2^{-(1+s)} \mu(F) = \frac{1}{2}\mathcal{H}^s(F)$. Thus $\mathcal{H}^s(F \backslash F_1) \geqslant \frac{1}{2}\mathcal{H}^s(F) > 0$, so if $E_1 = F \backslash F_1$ then $\mathcal{H}^s(E_1) > 0$ and $\overline{\lim}_{r\to 0} \mathcal{H}^s(F \cap B(x, r))/r^s \leqslant 2^{1+s}$ for $x \in E_1$. By Egoroff's theorem (see also Section 1.3) it follows that there is a compact set $E \subset E_1$ with $\mathcal{H}^s(E) > 0$ and a number $r_0 > 0$ such that

$\mathcal{H}^s(F \cap B(x,r))/r^s \leqslant 2^{2+s}$ for all $x \in E$ and all $0 < r \leqslant r_0$. However, we have that $\mathcal{H}^s(F \cap B(x,r))/r^s \leqslant \mathcal{H}^s(F)/r_0^s$ if $r \geqslant r_0$ so (4.11) holds for all $r > 0$. □

Corollary 4.12

Let F be a Borel subset of $\mathbb{R}^n$ with $0 < \mathcal{H}^s(F) \leqslant \infty$. Then there is a compact set $E \subset F$ such that $0 < \mathcal{H}^s(E) < \infty$ and a constant b such that

$$\mathcal{H}^s(E \cap B(x,r)) \leqslant br^s$$

for all $x \in \mathbb{R}^n$ and $r > 0$.

Proof. Theorem 4.10 provides us with a compact subset F_1 of F of positive finite measure, and applying Proposition 4.11 to F_1 gives the result. □

Corollary 4.12, which may be regarded as a converse of the Mass distribution principle 4.2, is often called ‘Frostman’s lemma’.

4.3 Potential theoretic methods

In this section we introduce a technique for calculating Hausdorff dimensions that is widely used both in theory and in practice. This replaces the need for estimating the mass of a large number of small sets, as in the Mass distribution principle, by a single check for the convergence of a certain integral.

The ideas of potential and energy will be familiar to readers with a knowledge of gravitation or electrostatics. For $s \geqslant 0$ the *s-potential* at a point x of $\mathbb{R}^n$ due to the mass distribution μ on $\mathbb{R}^n$ is defined as

$$\phi_s(x) = \int \frac{\mathrm{d}\mu(y)}{|x-y|^s}. \tag{4.12}$$

(If we are working in $\mathbb{R}^3$ and $s = 1$ then this is essentially the familiar Newtonian gravitational potential.) The *s-energy* of μ is

$$I_s(\mu) = \int \phi_s(x)\mathrm{d}\mu(x) = \iint \frac{\mathrm{d}\mu(x)\mathrm{d}\mu(y)}{|x-y|^s}. \tag{4.13}$$

The following theorem relates Hausdorff dimension to seemingly unconnected potential theoretic ideas. Particularly useful is part (a): if there is a mass distribution on a set F which has finite s-energy, then F has dimension at least s.

Theorem 4.13

Let F be a subset of $\mathbb{R}^n$.

(a) *If there is a mass distribution μ on F with $I_s(\mu) < \infty$ then $\mathcal{H}^s(F) = \infty$ and* $\dim_{\mathrm{H}} F \geqslant s$.

(*b*) *If F is a Borel set with* $\mathcal{H}^s(F) > 0$ *then there exists a mass distribution* μ *on F with* $I_t(\mu) < \infty$ *for all* $0 < t < s$.

Proof

(*a*) Suppose that $I_s(\mu) < \infty$ for some mass distribution μ with support contained in F. Define

$$F_1 = \left\{ x \in F : \overline{\lim_{r \to 0}}\, \mu(B(x,r))/r^s > 0 \right\}.$$

If $x \in F_1$ we may find $\varepsilon > 0$ and a sequence of numbers $\{r_i\}$ decreasing to 0 such that $\mu(B(x,r_i)) \geqslant \varepsilon r_i^s$. Since $\mu(\{x\}) = 0$ (otherwise $I_s(\mu) = \infty$) it follows from the continuity of μ that, by taking q_i $(0 < q_i < r_i)$ small enough, we get $\mu(A_i) \geqslant \frac{1}{4}\varepsilon r_i^s$ $(i = 1, 2, \ldots)$, where A_i is the annulus $B(x,r_i)\backslash B(x,q_i)$. Taking subsequences if necessary, we may assume that $r_{i+1} < q_i$ for all i, so that the A_i are disjoint annuli centred on x. Hence for all $x \in F_1$

$$\phi_s(x) = \int \frac{\mathrm{d}\mu(y)}{|x-y|^s} \geqslant \sum_{i=1}^{\infty} \int_{A_i} \frac{\mathrm{d}\mu(y)}{|x-y|^s}$$
$$\geqslant \sum_{i=1}^{\infty} \tfrac{1}{4}\varepsilon r_i^s r_i^{-s} = \infty$$

since $|x-y|^{-s} \geqslant r_i^{-s}$ on A_i. But $I_s(\mu) = \int \phi_s(x)\mathrm{d}\mu(x) < \infty$, so $\phi_s(x) < \infty$ for μ-almost all x. We conclude that $\mu(F_1) = 0$. Since $\overline{\lim}_{r\to 0}\mu(B(x,r))/r^s = 0$ if $x \in F\backslash F_1$, Proposition 4.9(*a*) tells us that, for all $c > 0$, we have

$$\mathcal{H}^s(F) \geqslant \mathcal{H}^s(F\backslash F_1) \geqslant \mu(F\backslash F_1)/c \geqslant (\mu(F) - \mu(F_1))/c = \mu(F)/c.$$

Hence $\mathcal{H}^s(F) = \infty$.

(*b*) Suppose that $\mathcal{H}^s(F) > 0$. We use $\mathcal{H}^s$ to construct a mass distribution μ on F with $I_t(\mu) < \infty$ for every $t < s$.

By Corollary 4.12 there exist a compact set $E \subset F$ with $0 < \mathcal{H}^s(E) < \infty$ and a constant b such that

$$\mathcal{H}^s(E \cap B(x,r)) \leqslant br^s$$

for all $x \in \mathbb{R}^n$ and $r > 0$. Let μ be the restriction of $\mathcal{H}^s$ to E, so that $\mu(A) = \mathcal{H}^s(E \cap A)$; then μ is a mass distribution on F. Fix $x \in \mathbb{R}^n$ and write

$$m(r) = \mu(B(x,r)) = \mathcal{H}^s(E \cap B(x,r)) \leqslant br^s. \qquad (4.14)$$

Then, if $0 < t < s$

$$\begin{aligned}\phi_t(x) &= \int_{|x-y|\leqslant 1} \frac{\mathrm{d}\mu(y)}{|x-y|^t} + \int_{|x-y|>1} \frac{\mathrm{d}\mu(y)}{|x-y|^t}\\ &\leqslant \int_0^1 r^{-t}\mathrm{d}m(r) + \mu(\mathbb{R}^n)\\ &= [r^{-t}m(r)]_{0+}^1 + t\int_0^1 r^{-(t+1)}m(r)\mathrm{d}r + \mu(\mathbb{R}^n)\\ &\leqslant b + bt\int_0^1 r^{s-t-1}\mathrm{d}r + \mu(\mathbb{R}^n)\\ &= b\left(1+\frac{t}{s-t}\right) + \mathcal{H}^s(F) = c,\end{aligned}$$

say, after integrating by parts and using (4.14). Thus $\phi_t(x) \leqslant c$ for all $x \in \mathbb{R}^n$, so that $I_t(\mu) = \int \phi_t(x)\mathrm{d}\mu(x) \leqslant c\mu(\mathbb{R}^n) < \infty$. □

Important applications of Theorem 4.13 will be given later in the book, for example, in the proof of the projection theorems in Chapter 6 and in the determination of the dimension of Brownian paths in Chapter 16. The theorem is often used to find the dimension of fractals F_θ which depend on a parameter θ. There may be a natural way to define a mass distribution μ_θ on F_θ for each θ. If we can show, that for some s,

$$\int I_s(\mu_\theta)\mathrm{d}\theta = \iiint \frac{\mathrm{d}\mu_\theta(x)\mathrm{d}\mu_\theta(y)\mathrm{d}\theta}{|x-y|^s} < \infty,$$

then $I_s(\mu_\theta) < \infty$ for almost all θ, so that $\dim_{\mathrm{H}} F_\theta \geqslant s$ for almost all θ.

Readers familiar with potential theory will have encountered the definition of the *s-capacity* of a set F:

$$C_s(F) = \sup_\mu\{1/I_s(\mu) : \mu \text{ is a mass distribution on } F \text{ with } \mu(F) = 1\}$$

(with the convention that $1/\infty = 0$). Thus another way of expressing Theorem 4.13 is

$$\dim_{\mathrm{H}} F = \inf\{s \geqslant 0 : C_s(F) = 0\} = \sup\{s \geqslant 0 : C_s(F) > 0\}.$$

Whilst this is reminiscent of the definition (2.11) of Hausdorff dimension in terms of Hausdorff measures, it should be noted that capacities behave very differently from measures. In particular, they are not generally additive.

*4.4 Fourier transform methods

In this section, we do no more than indicate that Fourier transforms can be a powerful tool for analysing dimensions.

The n-dimensional Fourier transforms of an integrable function f and a mass distribution μ on $\mathbb{R}^n$ are defined by

$$\hat{f}(u) = \int_{\mathbb{R}^n} f(x)\exp(\mathrm{i}x\cdot u)\mathrm{d}x \qquad (u \in \mathbb{R}^n) \tag{4.15}$$

$$\hat{\mu}(u) = \int_{\mathbb{R}^n} \exp(\mathrm{i}x\cdot u)\mathrm{d}\mu(x) \qquad (u \in \mathbb{R}^n) \tag{4.16}$$

where $x \cdot u$ represents the usual scalar product. (Fourier transformation extends to a much wider class of function using the theory of distributions.)

The s-potential (4.12) of a mass distribution μ is just the convolution

$$\phi_s(x) = (|\cdot|^{-s} * \mu)(x) \equiv \int |x-y|^{-s}\mathrm{d}\mu(y).$$

Formally, the transform of $|x|^{-s}$ may be shown to be $c|u|^{s-n}$, where c depends on n and s, so the convolution theorem, which states that the transform of the convolution of two functions equals the product of the transforms of the functions, gives

$$\hat{\phi}_s(u) = c|u|^{s-n}\hat{\mu}(u).$$

Parseval's theorem tells us that

$$\int \phi_s(x)\mathrm{d}\mu(x) = (2\pi)^n \int \hat{\phi}_s(u)\overline{\hat{\mu}(u)}\mathrm{d}u$$

where the bar denotes complex conjugation, so

$$I_s(\mu) = (2\pi)^n c \int |u|^{s-n}|\hat{\mu}(u)|^2\mathrm{d}u. \tag{4.17}$$

This expression for $I_s(\mu)$, which may be established rather more rigorously, is sometimes a convenient way of expressing the energy (4.13) required in Theorem 4.13. Thus if there is a mass distribution μ on a set F for which the integral (4.17) is finite, then $\dim_{\mathrm{H}}F \geqslant s$. In particular, if

$$|\hat{\mu}(u)| \leqslant b|u|^{-t/2} \tag{4.18}$$

for some constant b, then, noting that, by (4.16), $|\hat{\mu}(u)| \leqslant \mu(\mathbb{R}^n)$ for all u, we have from (4.17) that

$$I_s(\mu) \leqslant c_1 \int_{|u|\leqslant 1} |u|^{s-n}\mathrm{d}u + c_2 \int_{|u|>1} |u|^{s-n}|u|^{-t}\mathrm{d}u$$

which is finite if $0 < s < t$. Thus if (4.18) holds, any set F which supports μ has Hausdorff dimension at least t. The greatest value of t for which there is a mass distribution μ on F satisfying (4.18) is sometimes called the *Fourier dimension* of F, which never exceeds the Hausdorff dimension.

4.5 Notes and references

Many papers are devoted to calculating dimensions of various classes of fractal, for example the papers of Eggleston (1952), Beardon (1965) and Peyrière (1977) discuss fairly general constructions.

The potential theoretic approach was, essentially, due to Frostman (1935); see Taylor (1961), Hayman and Kennedy (1976), Carleson (1967) or Kahane (1985) for more recent accounts. For an introduction to Fourier transforms see Papoulis (1962).

The work on subsets of finite measure originates from Besicovitch (1952) and a very general treatment is given in Rogers (1998). Complete proofs of Theorem 4.10 may be found in Falconer (1985a) and Mattila (1995).

Subsets of finite positive packing measure are investigated by Joyce and Preiss (1995).

Exercises

4.1 What is the Hausdorff dimension of the 'Cantor tartan' given by $\{(x, y) \in \mathbb{R}^2 :$ either $x \in F$ or $y \in F\}$ where F is the middle third Cantor set?

4.2 Use the mass distribution principle and a natural upper bound to show that the set of numbers in $[0, 1]$ containing only even digits has Hausdorff dimension $\log 5/\log 10$.

4.3 Use the mass distribution method to show that the 'Cantor dust' depicted in figure 0.4 has Hausdorff dimension 1. (Hint: note that, taking the square E_0 to have side 1, any two squares in the set E_k of the construction are separated by a distance of at least 4^{-k}.)

4.4 Fix $0 < \lambda \leqslant \frac{1}{2}$, and let F be the set of real numbers

$$F = \left\{ \sum_{k=1}^{\infty} a_k \lambda^k : a_k = 0 \text{ or } 1 \text{ for } k = 1, 2, \dots \right\}.$$

Find the Hausdorff and box dimensions of F.

4.5 What is the Hausdorff dimension of $F \times F \subset \mathbb{R}^2$, where F is the middle third Cantor set?

4.6 Let F be the middle third Cantor set. What is the Hausdorff dimension of the plane set given by $\{(x, y) \in \mathbb{R}^2 : x \in F \text{ and } 0 \leqslant y \leqslant x^2\}$?

4.7 Use a mass distribution method to obtain the result of Example 4.5 directly rather than via Example 4.4.

4.8 Show that every number $x \geqslant 0$ may be expressed in the form

$$x = m + \frac{a_2}{2!} + \frac{a_3}{3!} + \cdots$$

where $m \geqslant 0$ is an integer and a_k is an integer with $0 \leqslant a_k \leqslant k - 1$ for each k. Let $F = \{x \geqslant 0 : m = 0 \text{ and } a_k \text{ is even for } k = 2, 3, \ldots\}$. Find $\dim_{\mathrm{H}} F$.

4.9 Show that there is a compact subset F of $[0, 1]$ of Hausdorff dimension 1 but with $\mathcal{H}^1(F) = 0$. (Hint: try a 'Cantor set' construction, but reduce the proportion of intervals removed at each stage.)

4.10 Deduce from Theorem 4.10 that if F is a Borel subset of $\mathbb{R}^n$ with $\mathcal{H}^s(F) = \infty$ and c is a positive number, then there is a Borel subset E of F with $\mathcal{H}^s(E) = c$.

4.11 Let μ be the natural mass distribution on the middle third Cantor set F (see after Principle 4.2). Estimate the s-energy of μ for $s < \log 2/\log 3$ and deduce from Theorem 4.13 that $\dim_{\mathrm{H}} F \geqslant \log 2/\log 3$.