PMA443 Fractals 2009–10

1 Introduction to fractals

Warning: these notes are incomplete! They are intended to ensure that you have, after the lectures, an accurate written record of the details of proofs *et cetera*, but they contain few diagrams and no record of computer demonstrations, both of which are essential to the course. You will need to take notes which include these elements.

1.1 Administration

The course is two lectures per week, Wednesdays at 10.00 in Hicks Lecture Room 9 and Thursdays at 4.10, in Hicks Lecture Theatre C. Set work will be given approximately once a week and solutions distributed one week later; set work will be collected and marked approximately one week in three.

1.2 Books

Prices are correct as of 5 February 2010.

1. K. J. Falconer, 'Fractal geometry: mathematical foundations and applications' 2/e (Wiley, 2003), paperback: ISBN-13: 9780470848623, £25.21, IC 513.84(F) (3 copies), St. George's 513.84(F) and electronic text.

2. D. Gulick,'Encounters with chaos' (McGraw-Hill, 1992) hardback ISBN-13: 9780070252035; paperback ISBN-13: 9780071129275. (The nearest book to the course, it is also recommended for the 'Chaos' module PMA 324.) IC 531.3(G) (5 copies) (Out-of-print, but the paperback is available secondhand on Amazon at £10.00.)

3. Richard M. Crownover, 'Introduction to fractals and chaos' (Jones and Bartlett, 1995) hardback ISBN-13: 9780867204643 £10.00 secondhand on Amazon (A nice introduction, but often less technical than the course.)

4. H. Lauwerier, 'Fractals: Images of Chaos' (Penguin, 1991) ISBN-13: 9780140144116 £2.83 secondhand on Amazon; IC 513.84 (L) (2 copies) (A popular account.)

5. B. B. Mandelbrot, 'The fractal geometry of nature' Freeman 1983 (the original, highly eccentric work on the subject) ISBN-13: 9780716711865 £9.29, at Amazon marketplace Western Bank Library 513.84(M).

6. M. F. Barnsley, 'Fractals everywhere 2/e' (Academic Press,1993) £43.69 (a good systematic treatment of fractals at 3rd year undergraduate level), ISBN-13: 9780120790692, IC 513.84(B).

7. Y. Fisher, (ed.), 'Fractal image compression: theory and application', (Springer, 1995), 341pp., ISBN-13: 9783540942115, Western Bank Library 3B 006.42 (F) (for more technical details of the application of Iterated Function Systems to image compression).

8. H.-O. Peitgen and P. H. Richter,'The Beauty of Fractals: Images of Complex Dynamical Systems' (Springer, 1986) ISBN-13: 9783540158516, £10.54, secondhand on Amazon (for beautiful computer graphics.) Western Bank Library Q 513.84(P)

9. H.-O. Peitgen and D. Saupe 'The science of fractal images' (Springer,1988) ISBN: 0387966080 £30.00 from ABE books (for the computer graphics enthusiast.) Western Bank Library Q 513.84(S).

10. H.-O. Peitgen, H. Jürgens, and D. Saupe book 'Chaos and Fractals' (Springer, 1992) ISBN-13: 9780387202297 £23.87 secondhand at Amazon; Western Bank Library 531.3 (P).

11. Manfred Schroeder 'Fractals, chaos and power laws: minutes from an infinite paradise' paperback ISBN-13: 9780716721369, £4.51 secondhand at Amazon An eccentric introduction to fractals and chaos, served with an unusually large helping of acoustic science.

12. R. L. Devaney and L. Keen 'Chaos and fractals: the mathematics behind the computer graphics' (American Mathematical Society, Proceedings of Symposia in Applied Mathematics vol.39, 1989) ISBN-13: 9780821801376, £1.40 secondhand at Amazon (A collection of articles introducing the subject; technical in places.)

13. Gerald A Edgar 'Classics on Fractals' Westview 2004 ISBN-13: 9780813341538 (An excellent source book for those interested in the history of the subject.) £12.75 at Amazon marketplace; Western Bank Library 514.742 (E)

14. K. J. Falconer, 'Techniques in fractal geometry' (Wiley, 1997), ISBN-13: 9780471957249, £58.41, at Amazon marketplace, Western Bank Library 513.84 (F). (A sequel to the above book by the same author; beyond the scope of this course).

15. W. A. Sutherland, 'Introduction to metric and topological spaces' (OUP, 1975) ISBN-13: 9780198531616, £9.80 secondhand at Amazon (properly includes the metric space theory referred to in the course) SLC 513.83(S)

1.3 Course Description

The concept of fractional dimension has been around for about 90 years, but the term 'fractal' and the interest in them, both popular and scientific, date from the proliferation of microcomputers in the late 1970's. The first aim of this course is to develop an understanding of the classical theory of dimension (there are several competing definitions to consider) and its relation to the recent applications of fractals in science and technology.

However 'fractals' are not just objects with fractional dimension. Most of the well-known fractals also possess 'self-similarity': they are composed of several parts, each of which is a small-scale copy of the whole. Amazingly, specifying this self-similarity is enough to determine the fractal. This has important applications to the compression of image data (and was used by Microsoft in their original encyclopaedia on a CD, 'Encarta'). The proof, which is a beautiful application of the Contraction Mapping Theorem in an unusual setting, is another feature of the course.

The overall structure of the course is this: we begin by studying the construction of fractals as self-similar sets; then we study dimension theory; finally, the two strands are brought together in Hutchinson's Theorem which allows one to compute the dimension from the self-similarity data for a wide class of fractals.

Throughout, the chapters on fractals are interspersed with chapters on some basic analysis prerequisite to the present study and to the understanding of abstract analysis generally. Indeed, whilst the primary aim of this course is the study of dimension and self-similarity, an important secondary aim is to consolidate your knowledge of the basic ideas of abstract analysis. The sections on countability, metric spaces, compactness and Lipschitz maps should be viewed in this light.

You will not be expected to use computers or to have substantial knowledge of computing.

1.4 Dimension

What exactly do we mean by the dimension of a set $A \subseteq \mathbb{R}^2$? If A is a 'filled-in' region, e.g. a disc, we want to say that A has dimension 2. If A is a straight line segment, we want A to have dimension 1. To produce a concept of dimension, we must find a meaningful way of measuring the distinction between these two sets. There are many possibilities — many notions of dimension — but three stand out. We describe one of them briefly here.

Suppose we pick $\varepsilon > 0$ and try to cover A with ε-balls $B(x, \varepsilon)$. Let $N(\varepsilon)$ be the minimum number of such balls needed to cover A. Generally, $N(\varepsilon) \to \infty$ as $\varepsilon \to 0$: but how fast?

If A is a line segment, then $N(\varepsilon) \propto \varepsilon^{-1}$.

If A is a disc, then $N(\varepsilon) \propto \varepsilon^{-2}$.

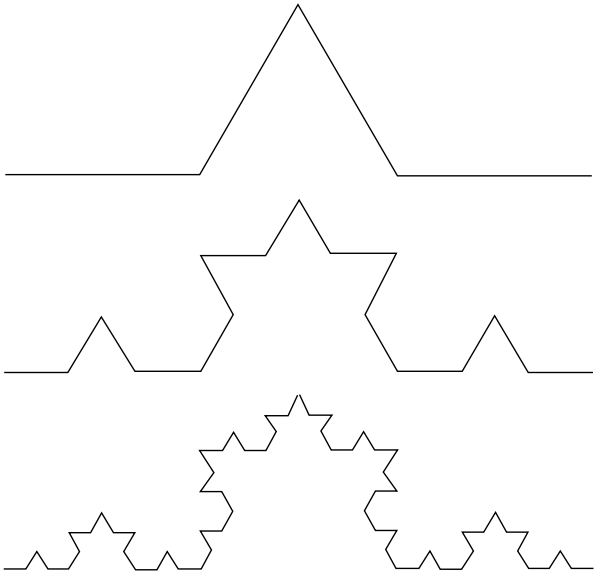
Let us say that if $N(\varepsilon) \propto \varepsilon^{-d}$, then A "has dimension d". More precisely, we define

$$\mathrm{Kdim}A = \lim_{\varepsilon \to 0} \frac{\log N(\varepsilon)}{\log(1/\varepsilon)},$$

if the limit exists.

1.5 Fractals

'Fractals' occur in dynamical systems theory in many ways and they arise from geometrical constructions of self-similar sets such as the *Koch curve*. This curve is most easily described as the limit of a sequence of continuous curves the first three of which are shown below. We shall see later that it has 'dimension' $\log 4/\log 3$.



There is a growing interest in the geometry of objects such as this. The subject originated in the early 20th. century, but only achieved popular recognition through the efforts of B. Mandelbrot. He coined the term 'fractal' — and refused to give a formal definition for it! His book 'The Fractal Geometry of Nature', together with the availability of low-cost microcomputers from the late 1970's onwards, made the mathematical world, and the general public, aware of 'fractal geometry'. His general thesis is that Euclidean geometry is based on shapes which have certain invariance properties, (a circle is invariant under rotations), but not others, (a circle is not invariant under scaling). Self-similar sets, (such as the *Koch curve*: one third of the Koch snowflake), have invariance under scaling, but, typically, not under rotation. Thus, to a disinterested deity, self-similar sets are just as 'natural' as circles. Indeed, these seem much more akin to many of the shapes found in the natural world. A branch of a fern, for example, has side shoots, each of which resembles the whole branch on a smaller scale.

These ideas have led Barnsley to remark that a picture of a fern can be encoded into a very small amount of memory. From this, he has built a technique for encoding visual images which currently looks quite promising as a solution to the problem of compressing visual image data sufficiently to make video-phones viable as a replacement for the telephone.

In order to discuss objects like the Koch snowflake, we need to use a limiting process. With the Koch snowflake, this is fairly easy, since it is a continuous curve in an obvious way. A continuous curve, in $\mathbb{R}^2$, is a pair $(x(t), y(t))$ of continuous, real-valued functions of a real variable and limits of such functions have been well discussed in real analysis courses. However, we shall generally be looking at limits of more complicated sets. We need a theory of convergence of sets in $\mathbb{R}^n$. We need a metric on a space whose 'points' are subsets of $\mathbb{R}^n$.

Since we shall be using a lot of metric space theory, for which we include a brief revision chapter. Assuming this, we shall develop the theory of dimension in fairly general metric spaces rather than just in $\mathbb{R}^n$. We do this, partly for notational convenience, but mainly because it shows how little of the structure of $\mathbb{R}^n$ we need to use, and consequently makes the arguments clearer.

By way of introducing fractals in a more formal way than above, we begin by discussing the simplest fractal of all—the Cantor ternary set. This also enables us to provide a link to last semester's Chaos course, for those students attending both, by showing how it arises in connection with the dynamics of a certain unimodal function. (We shall provide sufficient explanation to make the mathematics clear to those who have not attended the Chaos course, though without the motivation, they may not find the result so exciting!)

2 Some examples of fractals

Pictures of all these are available on links from the course web page.

1. The *Koch curve* has already been mentioned. It was introduced by Helge von Koch in his 1904 paper entitled 'Sur une courbe continue sans tangente obtenue par une construction gé omé trique é lé mentaire' (On a continuous curve without tangents constructible from elementary geometry). Weierstrass had already, in 1872, given a construction of a continuous nowhere-differentiable function. The graph of this was a curve without tangents, but the construction was purely analytical. Weierstrass's function was

$$f(x) = \sum_{n=0}^{\infty} b^n \cos(a^n x\pi),$$

where a is an odd positive integer and b a positive constant less than 1. The point of von Koch's construction is that it is a geometrical solution to a geometrical problem.

The rather pretty *Koch snowflake* is made by putting three Koch curves together.

2. The *quadratic Koch curve* is a another variant; putting four together produces a *quadratic Koch island*.

3. The *Sierpinski gasket* (in French: 'tamis de Sierpinski') (otherwise called the *Sierpinski triangle*) can be viewed as the set resulting from starting with a (filled-in) triangle and removing a half-size triangle from the middle (the convex hull of the mid-points of the sides), then similarly excising half-size triangles from the three remaining triangles, etc.. More formally, if S_n denotes the result of applying this procedure n times, then the Sierpinski gasket is the set $S = \bigcap_{n=1}^{\infty}$. Whether the initial triangle is equilateral of not is not important, the result is essentially the same. (Formally, it is the same up to a 'biLipschitz' equivalence - see later.) This fractal was constructed by Wacław Sierpiń ski in 1915

   Another way of constructing the Sierpinski gasket is to draw the outline of the initial triangle, then join the mid-points of the sides, then do the same with the three smaller triangles, etc, producing at each stage the boundary of the set produced at the corresponding stage of the first construction. The Sierpinski gasket is then the closure of the union of the nth stages of this construction.

   The Sierpinski gasket appears if we draw Pascal's triangle up to the $3.2^n + 1$st line and colour the numbers according to whether they are odd or even. The odd numbers form a pattern which is recognisable as the nth stage in the construction of the Sierpinski gasket.

4. The *Sierpinski carpet* (French: 'tapis de Sierpinski') (Sierpinski 1916) is similar to the gasket, but using a square from which squares are removed. It is subtly different from the gasket (see

comments in the chapter 'Topological Dimension'). The carpet, or rather a suitable stage in its construction, can be used as the basis for a compact antenna for e.g. a mobile phone. Other fractals can also be employed for this.

5. The *Menger sponge*, can be thought of as a three-dimensional version of the Sierpinski carpet.

We can find naturally occurring fractals - at least objects possessing self-similarity over a limited range of scales.

1. The *Romanesco broccoli*.

2. The *Black Spleenwort fern*. The *Barnsely fern* (created by Michael Barnsley, not part of the flora of South Yorkshire!) is a good attempt to simulate the Black Spleenwort.

Other objects, have parts that are self-similar to the whole, though the whole is not the union of these parts. We shall later describe this by means of Iterated Function Systems with Condensation.

1. A line of telegraph posts disappearing into the distance. The line from the second on is just a shrunk down copy of the whole.

2. Certain spirals are likewise composed of the first turn followed by a shrunk down copy of the whole.

3. Spiral shells of various sorts can thus be viewed as fractals of this type.

With these pictures in mind, we must now address the serious mathematics needed to describe the self-similarities and fractional dimensionalities of these images.

3 Metric spaces (revision) and products of metric spaces (new)

Basic idea: much of real analysis can be done in terms of the distance function

$$\text{distance}(x, y) = |x - y| \quad (x, y \in \mathbb{R}).$$

We abstract the properties of this function needed for analysis.

Definition 3.1 A *metric* on a set X is a function $d : X \times X \to \mathbb{R}^+$ such that

1. $d(x, z) \leqslant d(x, y) + d(y, z)$ $(x, y, z \in X)$,
2. $d(x, y) = d(y, x)$ $(x, y \in X)$,
3. $d(x, y) = 0$ iff $x = y$ $(x, y \in X)$.

A *metric space* is a pair (X, d) consisting of a set X and a metric d on X.

Example 3.2
1. The real line $\mathbb{R}$, with metric $d(x, y) = |x - y|$.
2. The real plane $\mathbb{R}^2$, with metric

$$d((x_1, x_2), (y_1, y_2)) = \sqrt{|x_1 - y_1|^2 + |x_2 - y_2|^2}.$$

Many of the notions associated with real analysis have generalizations to metric spaces.

Definition 3.3 A sequence (x_n) in a metric space X (strictly, (X, d)) *converges* to a point $x \in X$ iff for all $\varepsilon > 0$ there exists $N \in \mathbb{Z}^+$ such that for all $n \geqslant N$, $d(x_n, x) < \varepsilon$.

Definition 3.4 A subset A of a metric space X is *open* if, for all $a \in A$, there exists $\varepsilon > 0$ such that

$$B(a, \varepsilon) := \{x \in X : d(x, a) < \varepsilon\} \subseteq A.$$

Example 3.5 Every open ball $B(x, \varepsilon)$ in a metric space is an open set.

Definition 3.6 A subset A of a metric space X is *closed* if whenever a sequence (a_n) in A converges to a point $x \in X$, we have $x \in A$.

Example 3.7 In $\mathbb{R}$ the set $[0, 1]$ is closed, but the set $(0, 1]$ is neither open nor closed. Typically, most subsets of a metric space are neither open nor closed.

A set is closed if and only if its complement is open; consequently, the whole theory of closed sets is a reflection of that of open sets, with unions becoming intersections and vice versa, subsets becoming supersets, *et cetera* .

Definition 3.8 The *closure* of a set A in a metric space X is the set

$$\bar{A} := \{x \in X : \forall \varepsilon > 0 \quad B(x, \varepsilon) \cap A \neq \emptyset\}.$$

Equivalently, $\bar{A}$ is the smallest closed set containing A.

Definition 3.9 A set D is said to be *dense* in a metric space X if $\bar{D} = X$.

Definition 3.10 [new] A metric space is *separable* if it has a countable (see next section) dense subset.

Definition 3.11 If $(X, d_1), (Y, d_2)$, are two metric spaces, then a function $f : X \to Y$ is said to be *continuous* at a point $x_0 \in X$ iff $f(x_n) \to f(x_0)$ in Y whenever $x_n \to x_0$ in X.

If f is continuous at every point, we say that *f is continuous.*

Definition 3.12 [new] A *homeomorphism* is a bijection $f : X \to Y$ such that both f and f^{-1} are continuous. We say that X and Y are *homeomorphic* if there is a homeomorphism $f : X \to Y$.

Homeomorphisms preserve most of the important properties of metric spaces except completeness and total boundedness (see later for definitions). They preserve all properties definable in terms of convergent sequences (but not Cauchy sequences).

We can prove many real analysis theorems just as easily in the context of metric spaces. For example:

Proposition 3.13 *In a metric space X:*

1. *the sets $\emptyset$ and X are open and closed;*
2. *the intersection of two open sets is open: the union of two closed sets is closed;*
3. *arbitrary unions of open sets are open: arbitrary intersections of closed sets are closed.*

Proposition 3.14 *For a function $f : X \to Y$ between metric spaces, the following are equivalent:*

1. *f is continuous;*
2. *for all $x_0 \in X$ and $\varepsilon > 0$ there exists $\delta > 0$ such that $d_2(f(x), f(x_0)) < \varepsilon$ whenever $x \in X$ with $d_1(x_0, x) < \delta$;*
3. *$f^{-1}(G)$ is open in X, for every open set $G \subseteq Y$;*
4. *$f^{-1}(F)$ is closed in X, for every closed set $F \subseteq Y$.*

The proof of this depends on the Axiom of Choice, or at least its weaker version, Sequential Choice[1], to which it is equivalent. Hereafter, we shall assume this axiom without explicit mention. Curiously, many analysts are quite happy to do this but do point out places where the full Axiom of Choice is used.

[1] Given a sequence of non-empty sets (X_n), there is a sequence (x_n) with $x_n \in X_n$ for all n.

Definition 3.15 [new] The *product* of two metric spaces (X_1, d_1) and (X_2, d_2) is the space

$$X_1 \times X_2 = \{(x_1, x_2) : x_1 \in X_1, x_2 \in X_2\}$$

with the metric

$$d((a_1, a_2), (b_1, b_2)) = \max\{d_1(a_1, b_1), d_2(a_2, b_2)\}.$$

Exercise 3.16 Two alternative metrics on $X_1 \times X_2$ are the "taxi-cab metric"

$$d'((a_1, a_2), (b_1, b_2)) = d_1(a_1, b_1) + d_2(a_2, b_2)$$

and

$$d''((a_1, a_2), (b_1, b_2)) = \sqrt{d_1(a_1, b_1)^2 + d_2(a_2, b_2)^2}.$$

Show that, for $a, b \in X_1 \times X_2$,

$$d(a, b) \leqslant d''(a, b) \leqslant d'(a, b) \leqslant 2d(a, b).$$

Definition 3.17 A sequence (x_n) in a metric space (X, d) is said to be *Cauchy* if, for all $\varepsilon > 0$ there exists a positive integer N such that $d(x_p, x_q) < \varepsilon$ for all $p, q \geqslant N$. Informally, one may express this as '$d(x_p, x_q) \to 0$ as $p, q \to \infty$'.

It is easy to show that every convergent sequence is Cauchy; we are particularly fond of metric spaces in which the converse holds.

Definition 3.18 A metric space (X, d) is *complete* if every Cauchy sequence in X converges (in X).

Examples 3.19

1. The set $\mathbb{R}$ in the usual metric is a complete metric space.
2. The set $\mathbb{Q}$ (the rationals) in the usual metric is not complete, because a sequence $x_n \in \mathbb{Q}$ with $x_n \to \sqrt{2}$ in $\mathbb{R}$ is convergent in $\mathbb{R}$, so Cauchy in $\mathbb{R}$, so Cauchy in $\mathbb{Q}$, but is not convergent in $\mathbb{Q}$.

4 Countability

The notions of countable and uncountable sets underlie much of the theory of dimension. For example, we shall prove the following theorem

Theorem. *If A_i $(i = 1, 2, 3, \ldots)$ are subsets of a metric space X, then*

$$\mathrm{Hdim} \bigcup_{i=1}^{\infty} A_i = \sup_i (\mathrm{Hdim} A_i).$$

Here Hdim is a notion of dimension, to be defined later. It will be a sensible notion; for example, we shall have $\mathrm{Hdim}\{\text{point}\} = 0$ and $\mathrm{Hdim}\{\text{line}\} = 1$. However, this would produce a contradiction if we could write

$$\mathbb{R} = \{x_1, x_2, x_3, \ldots\} = \bigcup_{i=1}^{\infty} \{x_i\}.$$

Therefore, the fact, proved below, that this cannot be done is crucial to the attempt to produce a satisfactory definition of dimension.

Definition 4.1 A set C is said to be **countable** if either it is empty or there is a surjection $\theta : \mathbb{N} \to C$, (where $\mathbb{N} := \{1, 2, 3, \ldots\}$). A set is said to be **uncountable** if it is not countable.

In other words, C is countable if it may be written

$$C = \{c_1, c_2, c_3, \ldots\},$$

possibly with repetitions. (This rewriting of the definition results from writing c_i for $\theta(i)$.)

Let us remove the repetitions, i.e. define $\phi : \mathbb{N} \to C$ by letting $\phi(n) = \theta(i)$ with i minimal such that

$$\theta(i) \notin \{\phi(1), \ldots, \phi(n-1)\}.$$

Then either ϕ is a bijection between $\mathbb{N}$ and C or the definition of $\phi(n)$ fails at some point; in which case, ϕ is a bijection between $\{1, 2, \ldots, n-1\}$ and C. In the former case, we say that C is **countably infinite**; in the latter, (which includes the case $C = \emptyset$) that C is **finite**. In both cases, the inverse of ϕ gives us an injection of C into $\mathbb{N}$. (In the case $C = \emptyset$, the mapping ϕ is the empty mapping.) To summarise:

Proposition 4.2 (new) *For a set C, the following are equivalent:*

(i) C is countable; i.e. $C = \emptyset$ or there is a surjection $\mathbb{N} \to C$;

(ii) there is an injection $C \to \mathbb{N}$;

(iii) there is a bijection between C and either $\mathbb{N}$ or the set $\{1, 2, \ldots, n\}$ for some $n \in \mathbb{N} \cup \{0\}$.

Warning: we say that a set C is **countably infinite** if there is a bijection between C and $\mathbb{N}$. In PMA344 the word 'countable' was used for this. Our definition of 'countable' is 'countably infinite or finite'.

Examples 4.3 (i) The integers are countable:

$$\mathbb{Z} = \{0, +1, -1, +2, -2, +3, \ldots\}.$$

(ii) The positive rationals are countable:

$$\mathbb{Q}^+ = \left\{\frac{1}{1}, \frac{2}{1}, \frac{1}{2}, \frac{3}{1}, \frac{2}{2}, \frac{1}{3}, \frac{4}{1}, \frac{3}{2}, \frac{2}{3}, \frac{1}{4}, \frac{5}{1}, \cdots\right\},$$

with repetitions. Note that we have grouped numbers with the sum numerator + denominator equal to 2, then 3, 4, 5, *et cetera*.

(iii) The fact that the set $\mathbb{Q}$ of all rationals is countable is easily proved by combining the two previous ideas.

One of the most useful results for proving sets countable is the following.

Theorem 4.4 *Every countable union of countable sets is countable. That is, if each of the set A_i is countable ($i = 1, 2, 3, \ldots$), then $A = \bigcup_{i=1}^{\infty} A_i$ is countable.*

Proof. Clearly, we may assume that the A_i are all non-empty. Let

$$\begin{aligned} A_1 &= \{a_{11}, a_{12}, a_{13}, a_{14}, \ldots\}; \\ A_2 &= \{a_{21}, a_{22}, a_{23}, a_{24}, \ldots\}; \\ A_3 &= \{a_{31}, a_{32}, a_{33}, a_{34}, \ldots\}; \\ A_4 &= \{a_{41}, a_{42}, a_{43}, a_{44}, \ldots\}; \\ &\ldots \end{aligned}$$

(Note that the case of finitely many A_i is included by the simple expedient of repeating the A_i's in the enumeration.) Then

$$A = \{a_{11}, a_{21}, a_{12}, a_{13}, a_{22}, a_{31}, a_{41}, a_{32}, a_{23}, a_{14}, \ldots\},$$

(with possible repetitions). $\Diamond$

The other basic ways of inferring countability of some sets from the countability of others are contained in the following easy proposition.

Proposition 4.5 (new) *(i) If A is countable and $f : A \to B$ is a surjection, then B is countable.*

(ii) If A is countable and $g : B \to A$ is an injection, then B is countable. In particular, subsets of countable sets are countable.

Proof. For (i): Assume A is countable. If $A = \emptyset$ then $B = \emptyset$. Otherwise, there is a surjection $\theta : \mathbb{N} \to A$ so we have a surjection $f\theta : \mathbb{N} \to B$, showing that B is countable.

For (ii): if A is countable then, by Proposition 4.2, there is an injection $\phi : A \to \mathbb{N}$; hence we have an injection $\phi g : B \to \mathbb{N}$, which proves the countability of B, by Proposition 4.2 again. $\Diamond$

The key fact that makes countability interesting is that the reals are uncountable.

Theorem 4.6 *Each of the sets $[0, 1)$, $\mathbb{R}$ and $\mathbb{R} \setminus \mathbb{Q}$ (the irrationals) is uncountable.*

Proof. We prove first that $[0, 1)$ is uncountable, and the rest will follow easily. Every $x \in [0, 1)$ has a decimal expansion $x = 0.x_1x_2x_3\ldots$ not ending in an infinite string of nines. Suppose $[0, 1)$ is countable. Clearly $[0, 1)$ is infinite; let

$$[0, 1) = \{x^{(1)}, x^{(2)}, x^{(3)}, \ldots\}$$

where

$$\begin{aligned} x^{(1)} &= 0.x_1^{(1)}x_2^{(1)}x_3^{(1)}\ldots, \\ x^{(2)} &= 0.x_1^{(2)}x_2^{(2)}x_3^{(2)}\ldots, \\ x^{(3)} &= 0.x_1^{(3)}x_2^{(3)}x_3^{(3)}\ldots, \\ &\ldots \end{aligned}$$

We now construct a number $y = 0.y_1y_2y_3\ldots$ different from the above by defining

$$y_i = \begin{cases} 0 & \text{if } x_i^{(i)} \neq 0 \\ 1 & \text{if } x_i^{(i)} = 0. \end{cases}$$

Then $0.y_1y_2y_3\ldots$ is the decimal expansion of a number $y \in [0, 1)$ in a form not ending in an infinite string of nines (since it has no nines at all!). Further, there is no N such that $y = x^{(n)}$ because $y_n \neq x_n^{(n)}$, by construction. This contradicts the supposition that $[0, 1) = \{x^{(1)}, x^{(2)}, x^{(3)}, \ldots\}$ and proves the theorem.

The fact that $\mathbb{R}$ is uncountable follows from Proposition 4.5(ii) since if $\mathbb{R}$ were countable, then $[0, 1)$, being a subset of $\mathbb{R}$, would be countable too. Likewise, $[0, 1]$ is uncountable. Finally, we know that $\mathbb{Q}$ is countable, so if $\mathbb{R} \setminus \mathbb{Q}$ were countable, then $\mathbb{R} = \mathbb{Q} \cup (\mathbb{R} \setminus \mathbb{Q})$ would be countable, by Theorem 4.4. This is not so; therefore $\mathbb{R} \setminus \mathbb{Q}$ is uncountable. $\Diamond$

Exercise 4.7 A real number is said to be **algebraic** if it is a root of an equation

$$a_nx^n + a_{n-1}x^{n-1} + \ldots + a_2x^2 + a_1x + a_0 = 0,$$

with integer coefficients $a_n, \ldots, a_0$, (not all zero). By considering the size of the set A_N of all numbers x satisfying such an equation with $|a_n| + \ldots + |a_0| \leqslant N$ and $n \leqslant N$, or otherwise, show that the set of all algebraic numbers is countable.

A real number is **transcendental** if it is not algebraic. Show that the set of all transcendental numbers is uncountable. Deduce that transcendental numbers exist! (Actually, this is the easiest way of proving the existence of transcendental numbers. It is much harder to produce a specific transcendental number and very much harder to prove that interesting numbers such as e and π are transcendental.)

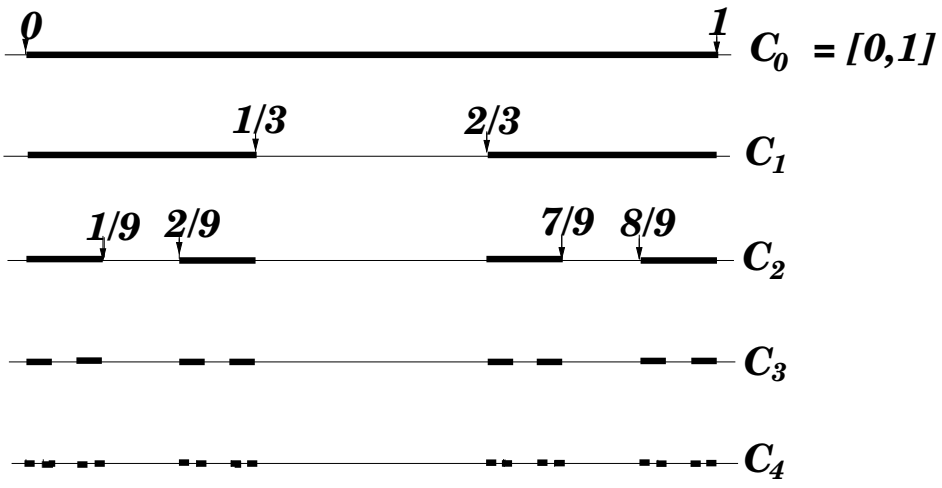
5 The Cantor ternary set

Definition 5.1 [Georg Cantor, 1883] The *Cantor Middle-Thirds Set* or *Cantor Ternary Set* is the set $C \subseteq \mathbb{R}$ defined as follows. Let

$$\begin{aligned} C_0 &= [0,1], \\ C_1 &= [0,\frac{1}{3}] \cup [\frac{2}{3},1], \\ C_2 &= [0,\frac{1}{9}] \cup [\frac{2}{9},\frac{1}{3}] \cup [\frac{2}{3},\frac{7}{9}] \cup [\frac{8}{9},1], \\ &\dots \end{aligned}$$

Then let

$$C = \bigcap_{n=0}^{\infty} C_n.$$



Proposition 5.2 *The Cantor ternary set is closed and contains no non-trivial intervals*

Proof. Each C_n is closed, since it is a finite union of closed intervals, so C is closed, (because every intersection of closed sets is closed.)

To show that C contains no non-trivial intervals, we observe that C_n contains no intervals of length greater than 3^{-n}. If C were to contain an interval of length $\varepsilon > 0$, we should have $\varepsilon > 3^{-N}$ for some N, so the interval could not be contained in C_N, and so not in C. $\Diamond$

We can describe the set C in terms of the possible ternary expansions of its points. Sample ternary expansions:

$$\begin{aligned} 25 \text{ (decimal)} &= 221 \text{ ternary} \\ 5/9 = 0.555\ldots \text{ (decimal)} &= 0.12 \text{ ternary} \\ 1/2 = 0.5 \text{ (decimal)} &= 0.111\ldots \text{ ternary} \end{aligned}$$

The set

$$C_1 = [0, \frac{1}{3}] \cup [\frac{2}{3}, 1],$$

is the set of all numbers which have a ternary expansion starting $0.0\ldots$ or $0.2\ldots$. Note the careful wording, due to the fact that some numbers have two possible ternary expansions: the number $1/3$ is most naturally written in ternary as 0.1, but it also has the ternary expansion $0.02222\ldots$; likewise $1 = 0.2222\ldots$; both of these are in C_1. The set

$$C_2 = [0, \frac{1}{9}] \cup [\frac{2}{9}, \frac{1}{3}] \cup [\frac{2}{3}, \frac{7}{9}] \cup [\frac{8}{9}, 1],$$

is likewise the set of all numbers which have a ternary expansion of the form $0.a_1a_2\ldots$ with $a_1, a_2 \in \{0, 2\}$. Generally, C_n is the set of all numbers which have a ternary expansion of the form $0.a_1a_2\ldots$ with $a_1, \ldots, a_n \in \{0, 2\}$. It follows that C is the set of all numbers which have a ternary expansion of the form $0.a_1a_2\ldots$ with $a_n \in \{0, 2\}$ for all n. This representation leads on to the following proposition.

Proposition 5.3 *The set C is uncountable.*

Proof. We define a map $\theta : C \to [0, 1]$ as follows. Let $x \in C$ have a ternary expansion $x = 0.x_1x_2\ldots$ with all the $x_n \in \{0, 2\}$, (note that such an expansion is unique), then $\theta(x) \in [0, 1]$ is defined by the *binary* expansion

$$0.\frac{x_1}{2}\frac{x_2}{2}\ldots.$$

Now every $y \in [0, 1]$ has at least one binary expansion $y = 0.y_1y_2\ldots$ with the $y_i \in \{0, 1\}$, so $y = \theta(x)$ where $x = 0.a_1a_2\ldots$, with each $a_i = 2y_i$, is in C. Thus θ is surjective. Since [0,1] is uncountable, it follows that C is uncountable. (By a modification of this proof, we can find a bijection between C and [0,1].) $\Diamond$

This result is particularly surprising if you try to get an idea of the 'length' of C. The sets C_n, being finite unions of intervals, have clearly defined lengths. In fact the length of C_n is $(2/3)^n$. Thus C is contained in arbitrarily short sets, and so any reasonable extension of the notion of length to encompass sets such as C (which is what is involved in the subject of "Measure Theory") gives C a length of zero. Nevertheless, C is uncountable.

The map θ constructed in this proof is quite interesting in its own right. It extends to a continuous function $\theta : \mathbb{R} \to I$ by defining θ to be constant in every open interval in the complement of C. Alternatively, one may approach the definition the other way round by defining θ on $\mathbb{R} \setminus C$ first:

$$\begin{aligned} \theta(x) &= 0 \qquad (x < 0) \\ \theta(x) &= 1 \qquad (x > 1) \\ \theta(x) &= \frac{1}{2} \qquad (x \in \left(\frac{1}{3}, \frac{2}{3}\right)) \end{aligned}$$

$$\begin{aligned}\theta(x) &= \frac{1}{4} \qquad (x \in \left(\frac{1}{9}, \frac{2}{9}\right)) \\ \theta(x) &= \frac{3}{4} \qquad (x \in \left(\frac{7}{9}, \frac{8}{9}\right)) \\ &\ldots\end{aligned}$$

This defines a monotonic non-decreasing function on $\mathbb{R} \setminus C$, which may then be extended to a monotonic non-decreasing function on $\mathbb{R}$ by defining

$$\theta(x) = \sup\{\theta(u) : u \in \mathbb{R} \setminus C,\ u < x\} \qquad (x \in C).$$

It is easy to see that θ so defined is continuous on $\mathbb{R}$, since

$$|\theta(x) - \theta(y)| < 2^{-n} \text{ whenever } |x - y| < 3^{-n} \qquad (x, y \in \mathbb{R}).$$

On [0,1], the function θ climbs from 0 to 1, but it is differentiable with derivative 0 on all the intervals of $I \setminus C$. This poses severe problems for any advanced theory of integration. When you differentiate this function, you get a function which is zero except on the set C which, having length zero, is negligible as far as integration goes. Therefore, integrating this, we get the zero function, which is not what we started with! The graph of θ is sometimes called a 'Devil's Staircase'. It was introduced by Cantor in 1884.

The Cantor set arises naturally in the study of the dynamical systems. We require only one definition, from the very beginning of the Chaos module.

Definition 5.4 Let $f : \mathbb{R} \to \mathbb{R}$ be any function and n any positive integer. We shall write

$$f^n(x) = \underbrace{f(f(\ldots(f}_{n}(x))\ldots));$$

that is, $f^0(x) = x$ and $f^{n+1}(x) = f(f^n(x))$ $(n = 0, 1, 2, \ldots)$. The function f^n is called the nth iterate of f.

'Studying the dynamical behaviour of the function f' means studying the sequence of iterates $(f^n(x))$ for various starting points $x \in \mathbb{R}$.

Consider the following function W, which I shall call the 'wigwam function'. It is like the 'tent map' introduced in the Chaos module, but its peak is a bit higher, (and wigwams are a bit taller than ordinary tents, aren't they?) We define $W : \mathbb{R} \to \mathbb{R}$ by

$$W(x) = \begin{cases} 3x & (x \leqslant \frac{1}{2}) \\ 3 - 3x & (x \geqslant \frac{1}{2}) \end{cases}$$

If $x < 0$ then $W^n(x) = 3^n x \to -\infty$ as $n \to \infty$. If $x > 1$ then $W(x) < 0$ and so, again, $W^n(x) \to -\infty$ as $n \to \infty$. If $x \in (\frac{1}{3}, \frac{2}{3})$ then $W(x) > 1$ and so, yet again, $W^n(x) \to -\infty$ as $n \to \infty$. If

$$x \in C_1 \setminus C_2 = (\frac{1}{9}, \frac{2}{9}) \cup (\frac{7}{9}, \frac{8}{9})$$

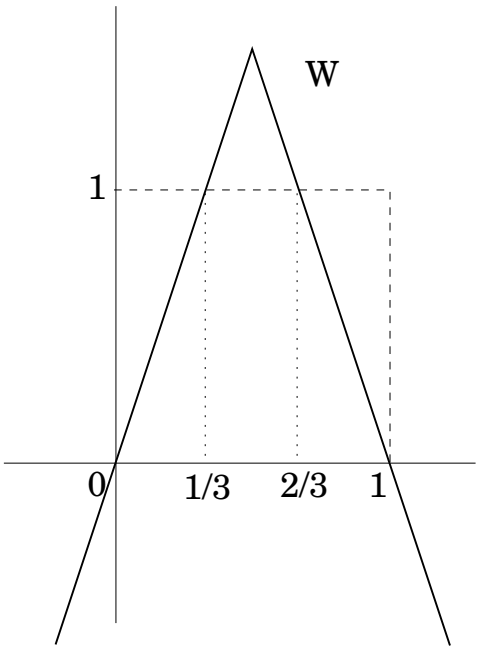
then $W(x) \in (\frac{1}{3}, \frac{2}{3})$ and so $W^n(x) \to -\infty$ as $n \to \infty$. In fact, it is easy to see that if $x \notin C_k$, then $W^{k+1}(x) < 0$, and hence $W^n(x) \to -\infty$ as $n \to \infty$. Thus the dynamical behaviour of W off C is trivial.

The interesting dynamics of W are confined to C. Observe that $W : C_k \to C_{k-1}$ $(k = 0, 1, 2, \ldots)$ so $W : C \to C$.

The Cantor ternary set is one of the simplest of fractals. What we have shown here is that it arises naturally from the elementary function W as soon as we ask about dynamical behaviour. The final remarks in this section (which are not examinable) are addressed to those who attended the Chaos module.

The dynamics of W on C may be described precisely by showing that the dynamical system (W, C) is topologically conjugate to a dynamical system (σ, Σ_2) consisting of the shift map on a space of infinite sequences. (You will see references to this in past papers, but it is not now part of either module.)

This topological conjugacy provides an easy proof that the dynamical system (W, C) has the following properties:



1. $|\mathrm{Per}_n(W)| = 2^n$;
2. the set $\mathrm{Per}(W)$ is dense in C;
3. there is a dense orbit for W in C.

The picture we have described here is very similar to that which obtains in the case of the quadratic maps F_μ ($\mu > 4$): a set which is homeomorphic to the Cantor set away from which the dynamical behaviour is trivial and on which the dynamics are those of the shift map. A similar picture occurs for the quadratic maps Q_c ($c < -2$), the Cantor-like set being the Julia set. Indeed, all the Q_c with c outside the Mandelbrot set show the same pattern, with a Julia set homeomorphic to the Cantor set, though spread out across the plane rather than confined to a line. This is a picture which occurs frequently in dynamical systems theory.

6 Compactness

We recall the results of the final chapter of PMA307 Metric Spaces.

Definition 8.1. Let $A \subseteq X$ be a subset of a metric space. We say that A is compact if every sequence in A has a subsequence that converges to a point of A.

Example 8.2. All closed intervals of the form $[a, b]$ are compact, by the Bolzano-Weierstrass Theorem. However, the real numbers $\mathbb{R}$, or any interval which is not bounded such as $[a, \infty)$ or $(-\infty, b]$, are not compact. For example, in $\mathbb{R}$ there is the sequence 0,1,2,3,... which has no convergent subsequence.

Lemma 8.3. *Let $A \subseteq X$ be a closed subset of a compact space X. Then A is compact.*

Proposition 8.4. *Let A be a compact subset of a metric space X. Then A is complete and so A is closed in X .*

Definition 8.5. A subset A of a metric space (X, d) is **bounded** if there is a $D > 0$ such that $d(a, b) \leqslant D$ for all $a, b \in A$. Equivalently, A is bounded if $A \subseteq B(x, R)$ for some $x \in X$ and $R > 0$.

Proposition 8.6. *Let A be a compact subset of a metric space (X, d). Then A is bounded.*

Theorem 8.7 (Heine-Borel). *A subset of $\mathbb{R}^N$ with the Euclidean metric is compact if and only if it is closed and bounded.*

Example 6.1 Let $X = \mathbb{R}$ and define a metric d on X by

$$d(x,y) = \begin{cases} |x-y| & \text{if } |x-y| < 1 \\ 1 & \text{if } |x-y| \geqslant 1. \end{cases}$$

Then a sequence (x_n) converges to a point x in (X,d) iff $x_n \to x$ in the usual metric on $\mathbb{R}$. Likewise, (x_n) is Cauchy in X iff it is Cauchy in $\mathbb{R}$. Consequently, anything defined in terms of these notions is the same in (X,d) as in $\mathbb{R}$. Thus (X,d) is complete, but not compact. Clearly, from its definition, (X,d) is bounded. Thus complete and bounded does not imply compact.

Theorem 8.9. *Let $f : X \to Y$ be a continuous map between metric spaces, and let $K \subseteq X$ be compact. Then $f(K)$ is compact.*

Corollary 8.10. *A function f which is real-valued and continuous on a compact set K is bounded on K and attains its bounds.*

Definition 8.14. Let X be a metric space. A collection $\{U_i : i \in I\}$ of subsets of X is **a cover** of $E \subseteq X$, or **covers** E, if

$$E \subseteq \bigcup_{i \in I} U_i.$$

If the indexing set I is a finite set then $\{U_i : i \in I\}$ is a **finite cover**. If each of the U_i is an open set then the collection is an **open cover**. A finite collection $U_{i_1}, \ldots, U_{i_n}$ with $i_1, \ldots i_n \in I$ is called a **finite subcover** of E if $E \subseteq U_{i_1} \cup \ldots \cup U_{i_n}$.

Definition 6.2 We use the term ε-ball to mean an open ball of radius ε, i.e. $B(x,\varepsilon)$ for some point x.

Definition 8.15. A subset K of metric space (X,d) is **totally bounded** if for every $\varepsilon > 0$ there is a finite collection of ε-balls covering K; i.e. for every $\varepsilon > 0$, there is a finite set $\{x_1, x_2, \ldots, x_n\} \subseteq K$ such that

$$K \subseteq B(x_1,\varepsilon) \cup B(x_2,\varepsilon) \cup \ldots \cup B(x_n,\varepsilon).$$

Exercise 6.3 Show that if K is totally bounded, the x_i in the above definition may be taken to lie anywhere in X. (Hint: use the x_i in X for $\varepsilon/2$ to get the desired x_i in K for ε.)

Proposition 8.16. *Every compact metric space is totally bounded.*

Definition 8.17. A metric space X is said to have **the Heine–Borel property** if every open cover of X has a finite subcover.

Theorem 8.19. Let (X,d) be a metric space. The following are equivalent:

(a) X is compact;

(b) X is totally bounded and complete;

(c) X has the Heine–Borel property.

A minor variation on this is the following.

Theorem 6.4 *A subset K of a complete metric space X is compact if and only if it is closed and totally bounded.*

Exercise 6.5 Show that:

(a) finite unions of compact sets are compact;

(b) the intersection of a closed set and a compact set is compact.

Exercise 6.6 Show that every decreasing sequence $K_1 \supseteq K_2 \supseteq K_3 \supseteq \ldots$ of compact non-empty sets in a metric space has a non-empty intersection.

Exercise 6.7 Show that if $K_1 \subseteq X_1$ and $K_2 \subseteq X_2$ are compact subsets of metric spaces X_1, X_2, then $K_1 \times K_2$ is a compact subset of the metric space $X_1 \times X_2$.

Theorem 6.8 *If X,Y are metric spaces with X compact and if $f : X \to Y$ is a continuous injective map, then $f : X \to f(X)$ is a homeomorphism.*

Proof. We need to show that the map $f^{-1} : f(X) \to X$ is continuous. Let F be a closed subset of X. Then F is compact, since closed subsets of compact sets are compact. By Theorem 8.9 above, $f(F)$ is compact. Therefore the set $(f^{-1})^{-1}(F) = f(F)$ is closed in $f(X)$, since compact sets are closed. $\Diamond$

7 Lipschitz maps and contractions

Definition 7.1 If $f : X \to Y$ is a mapping between metric spaces, then we say f is *Lipschitz* if there is a constant λ such that

$$d(f(x_1), f(x_2)) \leqslant \lambda d(x_1, x_2) \qquad (x_1, x_2 \in X).$$

The least such λ is called the *Lipschitz constant* $\operatorname{Lip} f$ of f. If f is not Lipschitz, we write $\operatorname{Lip} f = \infty$. If $\operatorname{Lip} f < 1$, (note, *strictly* less than 1) we say f is a *contraction*.

If $f : X \to Y$ is a bijection such that both f and f^{-1} are Lipschitz, we say f is *biLipschitz*.

If $d(f(x_1), f(x_2)) = \lambda d(x_1, x_2) \qquad (x_1, x_2 \in X)$, we say f is a *similitude*.

Proposition 7.2 *All Lipschitz maps are continuous.*

Proof. obvious. $\Diamond$

However, not all continuous maps are Lipschitz, and it will turn out that Lipschitz maps rather than continuous maps are the key to the theory of fractals.

Example 7.3 Let $f : [0,1] \to [0,1]$ (with the usual metric on [0,1]) be the positive square root function. Then if

$$d(f(0), f(x)) \leqslant \lambda d(0, x) \qquad (x \in [0,1])$$

we have

$$\sqrt{x} \leqslant \lambda x \qquad (x \in [0,1])$$

so

$$\lambda \geqslant x^{-1/2} \qquad (x \in [0,1])$$

and there is no finite λ satisfying this.

Proposition 7.4 *If $f : \mathbb{R} \to \mathbb{R}$ is differentiable then f is Lipschitz with $\operatorname{Lip} f \leqslant \lambda$ if and only if $|f'(x)| \leqslant \lambda$ for all $x \in \mathbb{R}$.*

Proof. (Exercise.) $\Diamond$

Example 7.5 The modulus function $f : \mathbb{R} \to \mathbb{R}$, $f(x) = |x|$ is not differentiable at 0, but is Lipschitz with Lipschitz constant 1.

Example 7.6 Let $f : \mathbb{R}^n \to \mathbb{R}^n$ be an *affine map*, i.e.

$$f((x_1, x_2, \ldots, x_n)) = (y_1, y_2, \ldots, y_n) + (b_1, b_2, \ldots, b_n),$$

where

$$y_i = \sum_{j=1}^{n} a_{ij} x_j,$$

for some fixed matrix $\boldsymbol{A} = (a_{ij})$ and vector $\boldsymbol{b} = (b_1, b_2, \ldots, b_n)$. In matrix notation,

$$f(\boldsymbol{x}) = \boldsymbol{A}\boldsymbol{x} + \boldsymbol{b}.$$

Then

$$\begin{aligned} d(f(\boldsymbol{x}), f(\boldsymbol{y})) &= d(\boldsymbol{Ax}, \boldsymbol{Ay}) \\ &= \sqrt{\sum_{i=1}^{n} \left(\sum_{j=1}^{n} a_{ij}(x_j - y_j) \right)^2} \\ &\leqslant \sqrt{\sum_{i=1}^{n} \left(\sum_{j=1}^{n} a_{ij}^2 \right) \left(\sum_{j=1}^{n} (x_j - y_j)^2 \right)} \qquad (1) \\ &= \sqrt{\sum_{i,j} a_{ij}^2}\, d(\boldsymbol{x}, \boldsymbol{y}), \end{aligned}$$

where (1) uses the Cauchy-Schwarz inequality. Actually, one can do better than this: the precise Lipschitz constant of f is called the *operator norm* $\|\boldsymbol{A}\|$ of the matrix $\boldsymbol{A}$, and it can be shown that it is the square root of the largest eigenvalue of $\boldsymbol{A}^T\boldsymbol{A}$. Here, $\boldsymbol{A}^T$ denotes the transpose of $\boldsymbol{A}$. All the eigenvalues of $\boldsymbol{A}^T\boldsymbol{A}$ are real and non-negative.

If $\boldsymbol{A}$ is invertible, then so is f and

$$f^{-1}(\boldsymbol{x}) = \boldsymbol{A}^{-1}(\boldsymbol{x} - \boldsymbol{b}) = \boldsymbol{A}^{-1}\boldsymbol{x} - \boldsymbol{A}^{-1}\boldsymbol{b},$$

so f is biLipschitz.

It may be shown that, when $\mathbb{R}^n$ has the usual Euclidean metric, the map f is a similitude if and only if A is a scalar multiple of an orthogonal matrix.

Theorem 7.7 (The Contraction Mapping Principle) *Let (X, d) be a complete metric space. Let $f : X \to X$ be a contraction with* $\operatorname{Lip} f = \lambda \in [0, 1)$. *Then f has a unique fixed point $x \in X$ and*

$$d(x_0, x) \leqslant (1 - \lambda)^{-1} d(x_0, f(x_0)) \qquad (x_0 \in X).$$

This has, essentially, been shown in the the PMA307 Metric Spaces course. What was proved was the following.

Proposition 7.8. *Let $f : X \to X$ be a contraction of the complete metric space (X, d), so that $d(f(x), f(y)) \leqslant kd(x, y)$ for some $0 \leqslant k < 1$, and let x_0 be any point of X. Then the sequence (x_n) defined by $x_{n+1} = f(x_n)$ converges to the unique fixed point x. Furthermore, for any n we have*

$$d(x_n, x) \leqslant \frac{k^n}{1 - k} d(x_0, f(x_0)).$$

Thus we get a bound for the distance of x_n from the limit x in terms of x_0.

Writing λ in place of k and specializing to the case $n = 0$, we get Theorem 7.7.

Exercise 7.8 Show that total boundedness and completeness are each preserved by biLipschitz maps. By considering the map $x \mapsto \tan x : (-\pi/2, \pi/2) \to \mathbb{R}$, show that neither is preserved by homeomorphisms. (You may assume that tan and arctan are continuous on the domains in question.)

8 The Hausdorff metric

In this chapter we shall look at the set of all compact subsets of $\mathbb{R}^N$ and define a metric on this set, so that we shall be able to talk of a sequence (K_n) of compact sets converging to a compact set K. Our construction would work equally well for the set of all compact subsets of a general metric space, but all our applications will be in $\mathbb{R}^N$, and it is more pleasant to begin our expedition into the new territories from the familiar ground of $\mathbb{R}^N$.

Let $\mathcal{H}_N$ denote the set of all non-empty compact subsets of $\mathbb{R}^N$. For $x \in \mathbb{R}^N$ and $K \in \mathcal{H}_N$ we define

$$d(x, K) = \inf\{d(x, y) : y \in K\}.$$

Note that the function $y \mapsto d(x,y)$ is a continuous function from K into $\mathbb{R}$ (easy exercise), so, by an exercise on compact sets, if K is compact, this function is bounded below and attains its bound; i.e. there exists $y_0 \in K$ with

$$d(x,K) = d(x,y_0)$$

so we may write

$$d(x,K) = \min\{d(x,y) : y \in K\}.$$

If, further, $x \notin K$, then $x \neq y_0$, so $d(x,K) = d(x,y_0) > 0$. Thus $d(x,K) = 0$ if and only if $x \in K$.

Proposition 8.1 *If $K \in \mathcal{H}_N$ then $x \mapsto d(x,K)$ is a continuous function on $\mathbb{R}^N$.*

Proof. For $x, y \in \mathbb{R}^N$ and $a_0 \in K$ such that $d(y,K) = d(y,a_0)$,

$$d(x,K) \leqslant d(x,a_0) \leqslant d(x,y) + d(y,a_0) = d(x,y) + d(y,K). \tag{2}$$

Likewise

$$d(y,K) \leqslant d(x,y) + d(x,K).$$

Combining these gives

$$|d(x,K) - d(y,K)| \leqslant d(x,y).$$

So, $x \mapsto d(x,K)$ is Lipschitz, with Lipschitz constant 1, and is therefore continuous. $\Diamond$

Again, for $A, B \in \mathcal{H}_N$ we deduce

$$\sup\{d(x,B) : x \in A\} = \max\{d(x,B) : x \in A\},$$

because the function $x \mapsto d(x,B)$ is continuous on the compact set A. We shall call this quantity $\rho(A,B)$. Then

1.

$$\begin{aligned}
\rho(A,C) &= \max\{d(a,C) : a \in A\} \\
&= d(a_0,C), \text{ for some } a_0 \in A, \\
&\leqslant d(a_0,b) + d(b,C), \text{ for all } b \in B, \text{ by (2)}, \\
&\leqslant d(a_0,b) + \rho(B,C), \text{ for all } b \in B.
\end{aligned}$$

So

$$\begin{aligned}
\rho(A,C) &\leqslant \inf\{d(a_0,b) : b \in B\} + \rho(B,C) \\
&= d(a_0,B) + \rho(B,C) \\
&\leqslant \rho(A,B) + \rho(B,C).
\end{aligned}$$

2. $\rho(A,B) \neq \rho(B,A)$, in general. [Draw a picture of typical sets $A, B \subseteq \mathbb{R}^N$.]

3. We have $\rho(A,A) = 0$ for all $A \in \mathcal{H}_N$, since $d(x,A) = 0$ when $x \in A$. Conversely, if $\rho(A,B) = 0$ for some $A, B \in \mathcal{H}_N$, then $d(a,B) = 0$ for all $a \in A$, so $a \in B$ for all $a \in A$, i.e. $A \subseteq B$.

The improvement we need to ρ is now clear. Let

$$d_H(A,B) = \max\{\rho(A,B), \rho(B,A)\} \qquad (A, B \in \mathcal{H}_N).$$

(I have adopted a different notation from that in Barnsley's book because I like things called "d" to be metrics. His d is my ρ; his h is my d_H.)

For $A, B, C \in \mathcal{H}_N$,

1.

$$\begin{aligned} d_H(A,C) &= \max\{\rho(A,C),\rho(C,A)\} \\ &\leqslant \max\{\rho(A,B)+\rho(B,C),\rho(C,B)+\rho(B,A)\} \\ &\leqslant \max\{\rho(A,B),\rho(B,A)\}+\max\{\rho(B,C),\rho(C,B)\} \\ &= d_H(A,B)+d_H(B,C). \end{aligned}$$

2. $d_H(A,B) = d_H(B,A)$, as a result of our "improvement".

3. $d_H(A,A) = \rho(A,A) = 0$ and if $d_H(A,B) = 0$, then $\rho(A,B) = 0$ and $\rho(B,A) = 0$, so $A \subseteq B$ and $B \subseteq A$, so $A = B$.

Thus d_H is a metric on $\mathcal{H}_N$.

Definition 8.2 The metric d_H is called the *Hausdorff metric* on $\mathcal{H}_N$.

Barnsley calls $\mathcal{H}_N$ "the space where fractals live" or (less accurately) "the space of fractals".

Exercise 8.3 Show that it is NOT generally true that

$$d(x,A) = d_H(\{x\},A) \qquad (x \in \mathbb{R}^N; A \in \mathcal{H}_N).$$

Exercise 8.4 Show that $d(x,A) \leqslant d(x,B) + d_H(B,A)$ $(x \in \mathbb{R}^N; A,B \in \mathcal{H}_N)$.

Exercise 8.5 (which is needed in the next chapter). Show that, for $A,B,C \in \mathcal{H}_N$,

$$\rho(A \cup B, C) = \max\{\rho(A,C),\rho(B,C)\}$$

and

$$\rho(A, B \cup C) \leqslant \min\{\rho(A,B),\rho(A,C)\}.$$

Deduce that, for $A,B,C,D \in \mathcal{H}_N$,

$$d_H(A \cup B, C \cup D) \leqslant \max\{d_H(A,C), d_H(B,D)\}.$$

Theorem 8.6 *The metric space $\mathcal{H}_N$ is complete.*

Actually, it is true for all complete metric spaces X that the corresponding space of all non-empty compact sets with the Hausdorff metric is complete, but the easy characterization of compact sets in $\mathbb{R}^N$ greatly simplifies our proof. The general proof may be found in Barnsley's book, pages 35-39, (to which one must add the fact that completeness plus total boundedness implies compactness.)

Lemma 8.7 *For a metric space (X,d), the following are equivalent:*
(i) X is complete;
(ii) if (x_n) is a sequence in X such that $d(x_n,x_{n+1}) \leqslant 2^{-(n+1)}$ $(n \geqslant 1)$, then (x_n) is convergent.

Proof of Lemma.

1. **(i) $\Rightarrow$ (ii)**
   If (x_n) is a sequence in X such that $d(x_n,x_{n+1}) \leqslant 2^{-(n+1)}$ $(n \geqslant 1)$, then for $m \geqslant n$,

$$\begin{aligned} d(x_n,x_m) &\leqslant d(x_n,x_{n+1}) + d(x_{n+1},x_{n+2}) + \ldots + d(x_{m-1},x_m) \\ &\leqslant 2^{-(n+1)} + 2^{-(n+2)} + \ldots + 2^{-m} \\ &\leqslant 2^{-n}, \end{aligned}$$

   so (x_n) is Cauchy. The implication (i) $\Rightarrow$ (ii) follows.

2. **(ii) ⇒ (i)**

Suppose (ii) and let (x_n) be a Cauchy sequence in X. Then there exist $n_1 < n_2 < \dots$ such that for each r,

$$d(x_p, x_q) < 2^{-(r+1)} \qquad (p, q \geqslant n_r).$$

Therefore the subsequence $(x_{n_r})_{r=1}^{\infty}$ satisfies the hypothesis of (ii) and so converges to some $x \in X$. But a Cauchy sequence with a convergent subsequence is necessarily convergent (by PMA307 Proposition 8.4). Therefore the whole Cauchy sequence (x_n) is convergent.$\diamond$

Proof of Theorem. Let (A_n) be a sequence in $\mathcal{H}_N$ such that

$$d_H(A_n, A_{n+1}) < 2^{-(n+1)}$$

for all n. By the lemma, it suffices to show that such sequences (A_n) are convergent in $\mathcal{H}_N$. Let

$$A = \{x : d(x, A_n) \leqslant 2^{-n} \text{ for all } n \geqslant 1\}.$$

Then

$$A = \bigcap_{n=1}^{\infty} B_n$$

where

$$B_n = \{x : d(x, A_n) \leqslant 2^{-n}\}.$$

Now B_n is closed as it is the inverse image of the closed set $[0, 2^{-n}] \subseteq \mathbb{R}$ under the continuous map $x \mapsto d(x, A_n)$. Further, B_n is bounded: A_n is compact, so bounded, say $A_n \subseteq B(x, r)$; so $B_n \subseteq B(x, r + 2^{-n})$. Since all closed, bounded subsets of $\mathbb{R}^N$ are compact, (this is where our great simplification occurs), we deduce that B_n is compact, for each n. Since the A_n are non-empty, so are the B_n.

Moreover

$$\begin{aligned} x \in B_{n+1} &\Rightarrow d(x, A_{n+1}) \leqslant 2^{-(n+1)} \\ &\Rightarrow d(x, A_n) \leqslant d(x, A_{n+1}) + d_H(A_{n+1}, A_n) \qquad \text{by Exercise 8.5,} \\ &\Rightarrow d(x, A_n) \leqslant 2^{-(n+1)} + 2^{-(n+1)} = 2^{-n} \\ &\Rightarrow x \in B_n \end{aligned}$$

Thus the B_n form a decreasing sequence of compact non-empty sets, so their intersection A is compact and non-empty, by Exercise 6.6.

We show that A is the limit of the sequence (A_n) in $\mathcal{H}_N$. If $a \in A$ then $d(a, A_n) \leqslant 2^{-n}$ for all n, by the definition of A, so $\rho(A, A_n) \leqslant 2^{-n}$. Conversely, if $a_n \in A_n$, there exists $a_{n+1} \in A_{n+1}$ with $d(a_n, a_{n+1}) \leqslant 2^{-(n+1)}$, then $a_{n+2} \in A_{n+2}$ with $d(a_{n+1}, a_{n+2}) \leqslant 2^{-(n+2)}$, *et cetera* . It follows that the sequence (a_n) is Cauchy and so convergent in $\mathbb{R}^N$; $a_n \to a$, say. Now, for all $r \geqslant 0$,

$$\begin{aligned} d(a, A_{n+r}) &\leqslant d(a, a_{n+r}) \\ &= \lim_{m \to \infty} d(a_m, a_{n+r}) \\ &\leqslant \lim_{m \to \infty} \left(2^{-(n+r+1)} + 2^{-(n+r+2)} + \dots + 2^{-m}\right) \\ &\leqslant 2^{-(n+r)}. \end{aligned}$$

Thus $a \in B_i$ $(i \geqslant n)$. Since the sequence (B_i) is decreasing, it follows that

$$a \in \bigcap_{i=1}^{\infty} B_i = A.$$

Thus $a \in A$ and $d(a_n, a) \leqslant 2^{-n}$; so $\rho(A_n, A) \leqslant 2^{-n}$. We have shown $d(A_n, A) \leqslant 2^{-n}$, for all n, so A is the limit of the sequence (A_n) in $\mathcal{H}_N$. $\diamondsuit$

9 Iterated Function Systems

Consider the following examples of "self-similar fractals": the *Sierpinski triangle* (or *Sierpinski gasket*), the *Koch curve* and the *Barnsley fern*.

All these are examples of sets A such that

$$A = \bigcup_{i=1}^{M} w_i(A)$$

for some set $\{w_i : 1 \leqslant i \leqslant M\}$ of contractions on $\mathbb{R}^2$. In fact, $M = 3$ for the Sierpinski triangle and $M = 4$ for the Koch curve. The fern is trickier: here, $M = 4$. The map w_2 takes the fern onto that part of the fern beyond the first two branches; w_3 and w_4 take the fern onto the first two branches and w_1, (in the notation of Table 3.8.3 in Barnsley's book), takes the fern, squashes it into a straight line interval and fits this in as the bottom part of the stem. Close examination reveals that the stems are composed of straight line segments, but this imperfection is easily ignored.

Definition 9.1 An *iterated function system (IFS)* on $\mathbb{R}^N$ is a finite set

$$\mathcal{W} = \{w_1, w_2, \ldots, w_M\}$$

of contractions on $\mathbb{R}^N$.

We say that a set A is *self-similar for* $\mathcal{W}$ if

$$A = \bigcup_{i=1}^{M} w_i(A). \tag{3}$$

We are particularly interested in non-empty *compact* self-similar sets: we are not interested in the fact that, in the above examples of IFS's $\mathcal{W}$ in $\mathbb{R}^2$, the sets $A = \mathbb{R}^2$ and $A = \emptyset$ satisfy (3).

Given an IFS $\mathcal{W}$, we define a map $W : \mathcal{H}_N \to \mathcal{H}_N$ by

$$W(K) = \bigcup_{i=1}^{M} w_i(K) \qquad (K \in \mathcal{H}_N).$$

(Since each w_i is continuous, the compactness of A implies the compactness of each $w_i(A)$; hence $W(A)$, being a finite union of compact sets, is compact.) We are looking for fixed points of W.

Let $s_i = \operatorname{Lip} w_i$ $(1 \leqslant i \leqslant M)$ and let $s = \max s_i$.

Theorem 9.2 *The map $W : \mathcal{H}_N \to \mathcal{H}_N$ is Lipschitz with $\operatorname{Lip} W \leqslant s$.*

Proof Consider first the case of just one mapping w_1. If $A, B \in \mathcal{H}_N$, then

$$\begin{aligned} \rho(w_1(A), w_1(B)) &= \max\{\min\{d(w_1(a), w_1(b)) : b \in B\} : a \in A\} \\ &\leqslant \max\{\min\{s_1 d(a, b) : b \in B\} : a \in A\} \\ &= s_1 \rho(A, B). \end{aligned}$$

Hence $d_H(w_1(A), w_1(B)) \leqslant s_1 d_H(A, B)$. (In fact, in this case, $\operatorname{Lip} W = s_1$, as may be seen by considering the action of W on singletons.)

The general case is then immediate from the following general lemma.

Lemma 9.3 *If $\phi_i : \mathcal{H}_N \to \mathcal{H}_N$ are Lipschitz with $\operatorname{Lip} \phi_i = s_i$ $(1 \leqslant i \leqslant M)$, then $\phi : \mathcal{H}_N \to \mathcal{H}_N$ defined by*

$$\phi(A) = \bigcup_{i=1}^{M} \phi_i(A)$$

is Lipschitz with $\operatorname{Lip} \phi \leqslant \max_i s_i$.

Proof. It suffices to prove the $M = 2$ case, as the general case is then proved by an easy induction based on the $M = 2$ case.

If $A, B \in \mathcal{H}_N$ then

$$\begin{aligned} d_H(\phi(A), \phi(B)) &= d_H(\phi_1(A) \cup \phi_2(A), \phi_1(B) \cup \phi_2(B)) \\ &\leqslant \max\{d_H(\phi_1(A), \phi_1(B)), d_H(\phi_2(A), \phi_2(B))\}, \text{ by an exercise in Ch. 7,} \\ &\leqslant \max\{s_1 d_H(A, B), s_2 d_H(A, B)\} \\ &= \max\{s_1, s_2\} d_H(A, B). \end{aligned}$$

This completes the proof of the lemma and hence the proof of the theorem. $\Diamond$

The main result of this theorem is that W is a *contraction mapping* on $\mathcal{H}_N$. We can therefore apply the Contraction Mapping Principle to obtain a fixed point for W.

Theorem 9.4 *Let $\mathcal{W}$ be an IFS on $\mathbb{R}^N$. Then there is a unique non-empty compact set $A \in \mathcal{H}_N$ which is self-similar for $\mathcal{W}$.*

Definition 9.5 We call this set A the *attractor* of $\mathcal{W}$.

Examples 9.6 1. The Cantor ternary set is the attractor of an IFS $\{w_0, w_1\}$ on $\mathbb{R}$ given by:

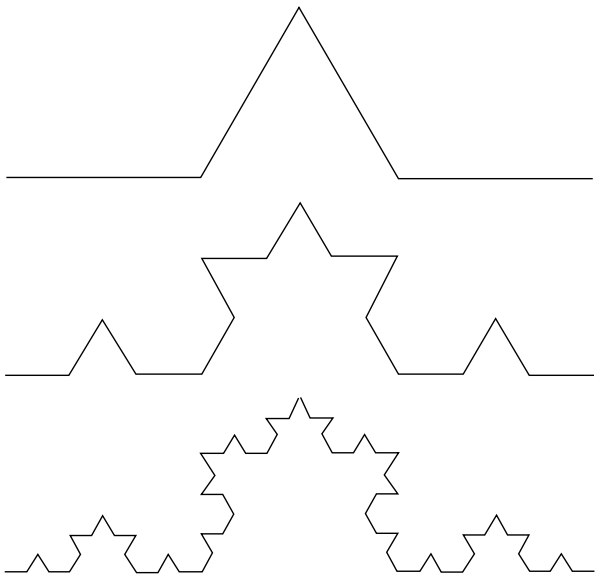
$$\begin{aligned} w_0(x) &= x/3, \\ w_1(x) &= (2 + x)/3. \end{aligned}$$

So

$$\begin{aligned} w_0(0) &= 0, & w_0(1) &= 1/3, \\ w_1(0) &= 2/3, & w_1(1) &= 1. \end{aligned}$$

2. The Koch curve may be defined by the IFS $\{w_1, w_2, w_3, w_4\}$ on $\mathbb{R}^2$ such that each w_i is an orientation-preserving similitude with $\operatorname{Lip} w_i = 1/3$ and

$$\begin{aligned} w_1(0,0) &= (0,0), & w_1(1,0) &= (1/3,0), \\ w_2(0,0) &= (1/3,0), & w_2(1,0) &= (1/2, 1/2\sqrt{3}), \\ w_3(0,0) &= (1/2, 1/2\sqrt{3}), & w_3(1,0) &= (2/3,0), \\ w_4(0,0) &= (2/3,0), & w_4(1,0) &= (1,0). \end{aligned}$$



Alternatively, the Koch curve is the attractor of an IFS $\{w_1, w_2\}$ consisting of orientation-reversing similitudes with

$$\begin{aligned} w_1(0,0) &= (0,0), & w_1(1,0) &= (1/2, 1/2\sqrt{3}), \\ w_2(0,0) &= (1/2, 1/2\sqrt{3}), & w_2(1,0) &= (1,0). \end{aligned}$$

The Contraction Mapping Principle, as we stated it, yields further information as to the location of the attractor. This translates into the following result.

Theorem 9.7 Barnsley's Collage Theorem. *Let $K \in \mathcal{H}_N$ and $\varepsilon > 0$ be given. Let $\mathcal{W}$ be an IFS such that*

$$d_H\left(K, \bigcup_{i=1}^{M} w_i(K)\right) < \varepsilon. \tag{4}$$

Let A be the attractor of $\mathcal{W}$. Then

$$d_H(A, K) < \frac{\varepsilon}{1-s},$$

where, as before, $s = \max_i \operatorname{Lip} w_i$.

The proof is immediate from the Contraction Mapping Principle (Theorem 6.7), applied to the mapping $W : \mathcal{H}_N \to \mathcal{H}_N$, since (4) is the statement $d_H(L, W(L)) < \varepsilon$.

Exercise 9.8 (hard). Show that if an IFS $\mathcal{W}$ in $\mathbb{R}^N$ has attractor A and K is a non-empty compact set such that $W(K) \subseteq K$, then

$$\bigcap_{n=1}^{\infty} W^n(K) = A.$$

Let us now consider the practical business of drawing fractals. Typically, our contractions w_n are affine maps and we are probably working in $\mathbb{R}^2$. One algorithm for producing the attractor is to follow the proof of the Contraction Mapping Principle: start with a set $A_0 \in \mathcal{H}_N$ and construct the sequence $W^n(A_0)$.

Reference

`http://links.uwaterloo.ca/`

Another approach is, in a sense, to replace the sets $W^n(A_0)$ by probability distributions. Select a point x at random, then plot $x_1 = w_{i_1}(x), x_2 = w_{i_2}(x), \ldots$, where $i_1, i_2, \ldots$ are selected randomly from $\{1, 2, \ldots, M\}$.

The simplest example of this is the construction of the Sierpinski gasket by the "Chaos Game". This "game" is played as follows. Select a point x_0 in (or near) a triangle ABC, preferably one of the vertices; select a vertex, A, B or C at random; let x_1 be the mid-point between x_0 and the selected vertex; continue. With a little thought, it will be seen that this corresponds to the random selection of one of three affine maps with Lipschitz constants 1/2.

References

Article: `http://math.bu.edu/DYSYS/chaos-game/chaos-game.html`
Software: `http://math.bu.edu/DYSYS/applets/chaos-game.html`

To construct (an approximation to) the attractor by this "Random Iteration Algorithm", we *either* let the algorithm run for a while before we start plotting, *or* we start with a point x_0 known to be in the attractor; for example a fixed point of one of the w_i (one of the vertices A, B, C, in the above example).

So much for the reproduction of images from an IFS. How do we produce an IFS to fit a given image? The Collage Theorem is the key to this. It says that if we take a set L (a leaf in some of the pictures shown), and represent it approximately as a collage of reduced copies of itself, i.e. if we write

$$L \approx \bigcup_{i=1}^{M} w_i(L), \tag{5}$$

for some contractions $w_1, w_2, \ldots, w_M$, then this IFS represents L fairly well. To be precise, if the error in (6) is ε, then the Hausdorff distance between the attractor of the IFS and L will be at most $(1-s)^{-1}\varepsilon$. Note that it helps to use w_i's with small Lipschitz constants.

The usefulness of this is that the attractor is close to L but, rather than being a blurred version of L as a classically engineered approximation might be, it is an image with a lot of detail, hopefully having a similar "texture" to L. Perhaps, in this way, it is more likely to fool the brain than is a blurred image containing as much information about L?

Barnsley and his company "Iterated Systems Inc." have developed these ideas into a working system for turning video pictures into IFS codes and back into pictures again. This "fractal compression" of images was used in an early CD encyclopaedia and you might come across fractal compressed files (with file extension .FIF) elsewhere.

IFS with condensation

Definition 9.9 If $C \in \mathcal{H}_N$ and $\{w_1, \ldots, w_M\}$ is an IFS on $\mathbb{R}^N$, then we call the $M+1$-tuple $\mathcal{W} = \{C, w_1, \ldots, w_M\}$ an *IFS with condensation* and we say that a set $A \subseteq \mathbb{R}^N$ is *self-similar for* $\mathcal{W}$ if

$$A = C \cup \bigcup_{i=1}^{M} w_i(A). \tag{6}$$

Theorem 9.10 *Let $\mathcal{W}$ be an IFS with condensation on $\mathbb{R}^N$. Then there is a unique non-empty compact set $A \in \mathcal{H}_N$ which is self-similar for $\mathcal{W}$.*

The proof of this follows in the same way as for Theorem 9.4. If we define $w_0 : \mathcal{H}_N \to \mathcal{H}_N$ by $w_0(K) = C$ for all $K \in \mathcal{H}_N$ then w_0 is a constant function on $\mathcal{H}_N$ and so a contraction and we only need to show that the function $W : \mathcal{H}_N \to \mathcal{H}_N$ defined by

$$W(K) = \bigcup_{i=0}^{M} w_i(K)$$

is a contraction. This follows as in the proof of Theorem 9.4, using Lemma 9.3.

As before, we call this set A the *attractor* of $\mathcal{W}$.

10 Topological dimension

The notion of 'topological dimension' is not our main concern in this course, but a quick discussion of it (without proofs) will set the scene for the more refined notions of dimension which follow. We begin with a few remarks about the history of dimension theory.

Before the advent of modern set theory and topology, the word 'dimension' was used only in a vague sense. A set or 'configuration' was said to be n-dimensional if n was the least number of real parameters needed to describe its points. Two problems arose in the late 19th century.

1. Cantor produced a bijection between $\mathbb{R}$ and $\mathbb{R}^2$. This bijection was highly discontinuous, so it showed that one needed to think of *continuous* parameterizations.

2. Peano produced a continuous surjection $f : [0,1] \to [0,1] \times [0,1]$.

Peano's example means that the 'continuous parameterizations' will have to be homeomorphisms. But is there another weird example which will kill that idea? Can $\mathbb{R}^n$ and $\mathbb{R}^m$ be homeomorphic with $m \neq n$? This is the key question, and it is surprisingly hard. It was solved by Brouwer in 1911. The answer was no, so mathematicians could breathe again! There was hope of a sensible topological dimension theory.

Brouwer's proof did not produce an explicit, workable definition of dimension. The foundations of the present theory were laid by Poincaré in 1912, and the formal definition is due to Brouwer in 1913. The theory was developed by Urysohn, Menger, Hurewicz and Tumarkin in the 1920's and the definitive account (for "separable" metric spaces) is Hurewicz and Wallman's classic book

W. Hurewicz and H. Wallman, *Dimension theory,* (Princeton University Press, Princeton, 1941).

(Incidentally, Menger's paper containing a recursive definition of dimension in a separable metric space was submitted to Monatshefte fur Mathematik und Physik in 1922, when he was 20; according to Kass's article about Menger in Notices of the American Mathematical Society, May 1996.)

More recently, research has been concentrated on extension of the theory to general topological spaces and the definition we give below is one of these more modern developments. A good modern account is Pears' book.

A. R. Pears, *Dimension theory of general spaces,* (Cambridge University Press, Cambridge, 1975)

(*The concept we are about to define is called "covering dimension". There are two other competing concepts: "small inductive dimension" and "large inductive dimension". Covering dimension and large inductive dimension are equal in all metric spaces. In separable metric spaces, all three are equal. An example of Prabir Roy (1962) shows that they do not all coincide in some non-separable metric space.*)

Definition 10.1 A **covering** $\{A_\lambda\}_{\lambda\in\Lambda}$ of a metric space X is a family of subsets of X such that

$$X = \bigcup_{\lambda\in\Lambda} A_\lambda.$$

An **open covering** is a covering each of whose sets A_λ is open. A **finite covering** is one with Λ finite. A covering $\{B_\mu\}_{\mu\in M}$ is said to be a **refinement** of $\{A_\lambda\}_{\lambda\in\Lambda}$ if for each $\mu \in M$ there is some $\lambda \in \Lambda$ with $B_\mu \subseteq A_\lambda$. The **order** of a family $\{A_\lambda\}_{\lambda\in\Lambda}$ of subsets, not all empty, is the largest integer n for which there exist $\lambda_1, \lambda_2, \ldots, \lambda_{n+1} \in \Lambda$ such that

$$A_{\lambda_1} \cap \ldots \cap A_{\lambda_{n+1}} \neq \emptyset.$$

(If there is no such integer n, we say that $\{A_\lambda\}_{\lambda\in\Lambda}$ has order ∞. A family of empty subsets has order -1.)

Definition 10.2 The **topological dimension** $\mathrm{topdim}X$ of a metric space X is the least integer n such that every finite open covering of X has an open refinement of order $\leqslant n$. If there is no such n, we write $\mathrm{topdim}X = \infty$.

The key idea here is that a space of dimension $\leqslant n$ should have an open covering with no more than $n+1$ sets overlapping at any point. However, we might have a space which is mainly one-dimensional but with a tiny two-dimensional piece. Such a space should be reckoned as two-dimensional, but if the two-dimensional piece were covered by one set of such a covering, it would be ignored. Hence the need to insist not just on the existence of one such covering, but on the existence of such a covering refining any given covering.

Theorem 10.3 *Topological dimension has the following properties:*

1. *it is integer-valued;*
2. *it is preserved under homeomorphisms: if $f : X \to Y$ is a homeomorphism between the metric spaces X and Y, then* $\mathrm{topdim}X = \mathrm{topdim}Y$*;*
3. $\mathrm{topdim}\mathbb{R}^n = n$ $(n \geqslant 1)$*;*
4. *if $Y \subseteq X$, then* $\mathrm{topdim}Y \leqslant \mathrm{topdim}X$*;*
5. *if $Y_1, Y_2 \subseteq X$, then*
$$\mathrm{topdim}(Y_1 \cup Y_2) \leqslant \mathrm{topdim}Y_1 + \mathrm{topdim}Y_2 + 1;$$

6. *if $Y_1, Y_2, \ldots$ is a sequence of closed subsets of X, then*

$$\text{topdim}\left(\bigcup_{i=1}^{\infty} Y_i\right) = \sup_i(\text{topdim}Y_i);$$

7. *if X_1, X_2 are metric spaces, then*

$$\text{topdim}(X_1 \times X_2) \leqslant \text{topdim}X_1 + \text{topdim}X_2.$$

The proofs are mainly non-trivial, so, as we are really interested in more refined notions of dimension, we omit them.

Examples 10.4 (without proofs)

1. Finite sets have dimension zero. Note that, in each of these examples, we are looking at the set as a metric space in its own right. Thus, singletons are open sets.
2. All countable sets have dimension zero. In particular:
   (a) (needed later) the set
   $$X = \{1/n : n = 1, 2, 3, \ldots\} \cup \{0\}$$
   with its usual metric as a subset of $\mathbb{R}$ has dimension zero;
   (b) the set of all rationals have dimension zero.
3. The set of all irrationals has dimension zero.
4. The Cantor ternary set C has dimension zero. Furthermore, every separable metric space of dimension zero is homeomorphic to a subset of C. We say that C is a *universal space* for the class of separable zero-dimensional metric spaces
5. The Sierpinski carpet is a universal space for the class of one-dimensional compact subsets of the plane. This was why Sierpinski introduced it in his 1916 paper (W. Sierpinski, 'Sur une courbe cantorienne qui contient une image biunivoque et continue de toute courbe donné e' *C.R. Acad. Sci. Paris*, **162** (1916) 629632). The Sierpinski gasket, however, does not have this property; no subspace of the gasket is homeomorphic to any plane figure that has five or more line segments meeting at a common point.
6. The Menger sponge is one-dimensional and is a universal space for the class of separable one-dimensional metric spaces. In fact, this is why Menger constructed it. Note that it follows that every one-dimensional separable metric space can be embedded in $\mathbb{R}^3$. Can every one-dimensional separable metric space can be embedded in $\mathbb{R}^2$? For the answer, see the Graph Theory module. In the same paper (K. Menger, 'Allgemeine Räume und Cartesische Räume', *Proc.Akad.Wetensch.Amst.* **29** (1926)476–482) Menger described similar universal spaces for higher dimensions.

11 Kolmogorov dimension

Throughout this chapter, ε will be less than 1. This means that the quantity $\log(1/\varepsilon)$, which appears frequently, is positive.

We now introduce the simplest concept of dimension capable of taking fractional values. The motivating idea is the following question: if S is a d-dimensional set, how much information is required to specify the position of a point in S to within ε? Rather than discuss the concept of "information" in detail, let us rephrase the question. Imagine S covered by balls of radius ε so that specifying a point to within ε means specifying a ball to which the point belongs. What is the least number $N(\varepsilon)$ of balls needed to cover S? It is easy to see that if S is a d-dimensional cube in the usual sense, then $N(\varepsilon) \asymp \varepsilon^{-d}$ as $\varepsilon \to 0$. (This notation means that there exist numbers $0 < a \leqslant b$ such that $a \leqslant N(\varepsilon)/\varepsilon^{-d} \leqslant b$ for all sufficiently small ε.)

Definition 11.1 Let K be a non-empty compact subset of a metric space X. For each $\varepsilon > 0$, let $N(\varepsilon)$ be the minimum number of open balls of radius ε (ε-balls) centred on points of K needed to cover K. (Since K is compact, it is totally bounded by Theorem 6.4: i.e. $N(\varepsilon)$ is finite. The fact that K is non-empty implies that $N(\varepsilon) > 0$.) Then we define the *Kolmogorov dimension* of K by

$$\mathrm{Kdim} K = \lim_{\varepsilon \to 0} \frac{\log N(\varepsilon)}{\log(1/\varepsilon)},$$

if this limit exists. Otherwise we say that K does not have Kolmogorov dimension.

There is another way to define Kolmogorov dimension, using 'limsup' in place of 'lim'; the resulting quantity is always defined, so it is not necessary to qualify all results by requiring that the sets concerned have Kolmogorov dimension. This approach was used in this course up to June 2000. It has been superseded by the conceptually simpler approach using limits. The result is that, in general, the theorems are more untidy, but there is a benefit in the theorem on dimensions of products.

Kolmogorov dimension is properly called "capacity", but the latter term is so frequently used in a different way in potential theory that I prefer to avoid it. It is also known as "Minkowski dimension".

Example 11.2 To illustrate this definition in action, we compute the Kolmogorov dimension of the Cantor ternary set C. We recall the definition of C. Let

$$\begin{aligned} C_0 &= [0,1], \\ C_1 &= [0,\frac{1}{3}] \cup [\frac{2}{3},1], \\ C_2 &= [0,\frac{1}{9}] \cup [\frac{2}{9},\frac{1}{3}] \cup [\frac{2}{3},\frac{7}{9}] \cup [\frac{8}{9},1], \\ &\dots \end{aligned}$$

Then

$$C = \bigcap_{n=0}^{\infty} C_n.$$

Given $\varepsilon > 0$, let n be such that $3^{-(n+1)} < 2\varepsilon \leqslant 3^{-n}$, i.e. $n = [\log_3(1/2\varepsilon)]$, then no ε-ball centred on a point of C can intersect more than one interval of C_n. This is because the gaps between the intervals of C_n are all at least 3^{-n}. Now every interval of C_n contains points of C. Therefore, at least 2^n such ε-balls are needed to cover C: i.e. $N(\varepsilon) \geqslant 2^n$.

On the other hand, we can cover C_{n+1} and so C by 2^{n+2} ε-balls centred on the end points of the closed intervals of which C_{n+1} is composed. Therefore $N(\varepsilon) \leqslant 2^{n+2}$.

Thus

$$\frac{n \log 2}{\log 2 + (n+1)\log 3} \leqslant \frac{\log N(\varepsilon)}{\log(1/\varepsilon)} \leqslant \frac{(n+2)\log 2}{\log 2 + n \log 3}.$$

As $\varepsilon \to 0$, we have $n \to \infty$, and so the outer terms tend to $(\log 2)/(\log 3)$. The Sandwich Rule implies that (the limit exists and)

$$\lim_{\varepsilon \to 0} \frac{\log N(\varepsilon)}{\log(1/\varepsilon)} = \frac{\log 2}{\log 3},$$

so the Cantor Ternary Set has Kolmogorov dimension, equal to

$$(\log 2)/(\log 3) = 0.6309297536\dots.$$

Remark 11.3 There is an interesting approximation

$$\frac{\log 2}{\log 3} \approx \frac{12}{19} = 0.6315789\dots$$

which is the basis for the equal temperament system in music. The octave, which represents a frequency ratio of 2 is divided into 12 semitones, each representing a frequency ratio of $2^{1/12}$. A pure interval of one 'twelfth' is a frequency ratio of 3 and this is approximated by 19 semitones:

$$3 \approx 2^{19/12}.$$

Taking logs:

$$\log 3 \approx \frac{19}{12} \log 2.$$

For more information on scales and temperament see:

> Manfred Schroeder, "Fractals, chaos, power laws" (Freeman, 1991) 99–101;
> Alexander Wood, "The physics of music" (Methuen, 1962) Chapter 11.

We can draw two important morals from this, both of which differentiate Kolmogorov dimension sharply from topological dimension.

Remark 11.4 The Kolmogorov dimension is not necessarily an integer.

Remark 11.5 The Kolmogorov dimension is not generally invariant under homeomorphisms. This is not so immediate to prove, but it is strongly suggested by the way that the number $(\log 2)/(\log 3)$ arises from the 2 and 3 involved in the *geometry* of C.

Let us set up a slightly different Cantor set D by removing the middle halves of intervals: let

$$\begin{aligned} D_0 &= [0,1], \\ D_1 &= [0,\frac{1}{4}] \cup [\frac{3}{4},1], \\ D_2 &= [0,\frac{1}{16}] \cup [\frac{3}{16},\frac{1}{4}] \cup [\frac{3}{4},\frac{13}{16}] \cup [\frac{15}{16},1], \\ &\cdots \end{aligned}$$

and

$$D = \bigcap_{n=0}^{\infty} D_n.$$

The calculation of dimension goes over with 3 replaced by 4 to yield

$$\mathrm{Kdim} D = \frac{\log 2}{\log 4} = \frac{1}{2}.$$

However, we can easily produce a homeomorphism $f : D \to C$ by noting that D is the set of all points having an expansion in the quaternary scale involving only 0's and 3's, in the same way that C consists of those numbers having an expansion in the ternary scale involving only 0's and 2's. The mapping f consists of replacing 3's in quaternary expansions of points in D by 2's and calling the resulting strings ternary expansions of points in C. Alternatively, a function $f : [0,1] \to [0,1]$ whose restriction maps $D \to C$ can be defined as the limit of a sequence of monotonic increasing functions $f_n : [0,1] \to [0,1]$ which map the intervals of D_n onto the intervals of C_n and which are linear between the end-points of these intervals. The sequences (f_n) and (f_n^{-1}) are both uniformly convergent, so the limit f of (f_n) is a homeomorphism.

Now let us consider a notion of dimension for non-empty compact sets $K \subseteq \mathbb{R}^n$ which is equivalent to Kolmogorov dimension, but is better suited to practical estimation.

Definition 11.6 We define the *grid dimension* of a non-empty compact set $K \subseteq \mathbb{R}^n$ as follows. For each $\varepsilon > 0$, we choose a grid of n orthogonal sets of parallel hyperplanes with separation ε.

We let $N_g(\varepsilon)$ denote the number of (closed) grid cubes containing points of K and then define

$$\mathrm{griddim} K = \lim_{\varepsilon \to 0} \frac{\log N_g(\varepsilon)}{\log(1/\varepsilon)},$$

if the limit exists.

As it stands, this definition depends on the choice of grid for each $\varepsilon > 0$. However, the next theorem shows that this does not matter.

Theorem 11.7 *If K is a non-empty compact subset of $\mathbb{R}^n$ then, for any choices of grids, K has grid dimension if and only if it has Kolmogorov dimension, in which case* $\mathrm{griddim}K = \mathrm{Kdim}K$. *(Hence* $\mathrm{griddim}K$ *is independent of the choices of grids.)*

Proof. Let C be a closed ε-grid cube containing at least one point $x \in K$. Then the diameter of C, the distance from one vertex to the opposite vertex, is $\sqrt{n}\,\varepsilon$, so $C \subseteq B(x, 2\sqrt{n}\,\varepsilon)$ (we allow a spare factor of 2 here to allow for the case when x is a vertex, the cube being closed and the ball open — the overkill is irrelevant). Thus every closed ε-grid cube which meets K is contained in a $2\sqrt{n}\,\varepsilon$-ball centred on a point of K. Now K is covered by $N_g(\varepsilon)$ such cubes, and therefore K is covered by $N_g(\varepsilon)$ of these $2\sqrt{n}\,\varepsilon$-balls centred on points of K. Therefore the least number of such balls needed to cover K is no more than $N_g(\varepsilon)$. That is, $N(2\sqrt{n}\,\varepsilon) \leqslant N_g(\varepsilon)$, for all $\varepsilon > 0$: equivalently

$$N(\varepsilon) \leqslant N_g\left(\frac{1}{2\sqrt{n}}\varepsilon\right),$$

for all $\varepsilon > 0$ (by replacing ε by $\varepsilon/(2\sqrt{n})$.

Conversely, every open ball of radius ε meets no more than 3^n ε-grid cubes, so $N_g(\varepsilon) \leqslant 3^n N(\varepsilon)$.

We complete the proof with an argument we shall need repeatedly, and which we therefore package as a technical lemma.

Lemma 11.8 (Comparison Lemma) *Let $A(\varepsilon)$, $B(\varepsilon)$ be two positive-real-valued functions on $\mathbb{R}^+$ and suppose that there exist positive constants $\lambda_1, \lambda_2, \mu_1, \mu_2$ such that for all $\varepsilon > 0$*

(a) $A(\varepsilon) \leqslant \lambda_1 B(\mu_1 \varepsilon)$ and

(b) $B(\varepsilon) \leqslant \lambda_2 A(\mu_2 \varepsilon)$,

then the limit

$$\lim_{\varepsilon \to 0} \frac{\log A(\varepsilon)}{\log(1/\varepsilon)}$$

exists if and only if the limit

$$\lim_{\varepsilon \to 0} \frac{\log B(\varepsilon)}{\log(1/\varepsilon)}$$

exists, in which case the two limits are equal.

Before proving the lemma, we observe that the lemma will complete the proof of our theorem by putting $A(\varepsilon) = N(\varepsilon)$, $B(\varepsilon) = N_g(\varepsilon)$, $\lambda_1 = 1$, $\mu_1 = 1/(2\sqrt{n})$, $\lambda_2 = 3^n$, and $\mu_2 = 1$. (Remember that n is fixed, so it can happily form part of the expressions for the constants μ_1 and λ_2.) $\Diamond$

Proof of Lemma. From (b) we have, on replacing ε by $\mu_2\varepsilon$,

$$\lambda_3 B(\mu_3 \varepsilon) \leqslant A(\varepsilon),$$

where $\lambda_3 = \lambda_2^{-1}$ and $\mu_3 = \mu_2^{-1}$.

Then

$$\begin{aligned}
\left(\frac{\log \lambda_3 + \log B(\mu_3\varepsilon)}{\log(1/\mu_3\varepsilon)}\right)\left(\frac{\log(1/\varepsilon) - \log \mu_3}{\log(1/\varepsilon)}\right) &= \frac{\log\left(\lambda_3 B(\mu_3\varepsilon)\right)}{\log(1/\varepsilon)} \\
&\leqslant \frac{\log A(\varepsilon)}{\log(1/\varepsilon)} \\
&\leqslant \frac{\log\left(\lambda_1 B(\mu_1\varepsilon)\right)}{\log(1/\varepsilon)} \\
&= \left(\frac{\log \lambda_1 + \log B(\mu_1\varepsilon)}{\log(1/\mu_1\varepsilon)}\right)\left(\frac{\log(1/\varepsilon) - \log \mu_1}{\log(1/\varepsilon)}\right).
\end{aligned}$$

Now as $\varepsilon \to 0$,we have $\log(1/\varepsilon) \to \infty$ and so

$$\frac{\log(1/\varepsilon) - \log \mu_i}{\log(1/\varepsilon)} \to 1 \qquad (i = 1, 3),$$

and

$$\frac{\log \lambda_i}{\log(1/\mu_i \varepsilon)} \to 0 \qquad (i = 1, 3).$$

Suppose

$$L_B := \lim_{\varepsilon \to 0} \frac{\log B(\varepsilon)}{\log(1/\varepsilon)}$$

exists; then, as $\varepsilon \to 0$, we have $\mu_i \varepsilon \to 0$, so

$$\frac{\log B(\mu_i \varepsilon)}{\log(1/\mu_i \varepsilon)} \to L_B.$$

Hence, in the above chain of inequalities, the first and last expressions both tend to L_B. Therefore, by the Sandwich Rule,

$$\frac{\log A(\varepsilon)}{\log(1/\varepsilon)} \to L_B,$$

as desired. This proves half of the lemma, but the other half is similar, with the rôles of A and B being reversed. $\Diamond$

To make an "experimental determination of dimension", we put ε-grids for various ε over the set K and compute $N_g(\varepsilon)$. We then plot $\log N_g(\varepsilon)$ against $\log(1/\varepsilon)$. Typically, these points might lie approximately on a straight line; that is, there might be a relation of the form

$$\log N_g(\varepsilon) = c + d \log(1/\varepsilon), \tag{7}$$

coming from a relation

$$N_g(\varepsilon) = k\varepsilon^{-d}$$

where $c = \log k$. In this case

$$\mathrm{Kdim} K = \lim \frac{\log N_g(\varepsilon)}{\log(1/\varepsilon)} = d,$$

which is the slope of the graph (7). Notice that the slope of (7) gives a better approximation to the Kolmogorov dimension than taking the value of

$$\frac{\log N_g(\varepsilon)}{\log(1/\varepsilon)}$$

for the smallest value of ε considered, (always assuming the linear relation (7)).

This is, of course, an experimental approximation to an abstract mathematical notion. We are measuring the "texture" of the set K only *over a certain range of scales*, the range of ε's used, whereas the mathematical idea refers to the limit as ε tends to zero. Nevertheless, it is a useful way of measuring "texture".

Examples 11.9 1. Barnsley, in his book, gives some examples of woodcuts and invites the reader to estimate their dimensions. He conjectures that individual artists produce work of characteristic dimension.

2. Over a wide range of scales, the dimension of the surface of the human lung is 2.17. (There is in article in the February 1990 issue of "Scientific American" on chaos and fractals in physiology. Although it doesn't go into details, it does include some pretty pictures of a latex cast of the lung and of a computer model of a similar fractal structure.) The surface of the grey matter of the brain is even more convoluted with a dimension estimated at 2.65 ± 0.05.

3. Georgia D. Tourassi et al. [Phys. Med. Biol. **51** (2006) 1299–1312] investigated to the use of fractal dimension to identify architectural distortion of breasts in mammograms. Lower fractal dimension is associated with abnormal structure. This is important as architectural distortion is a sign of malignant breast tumours often missed in mammographic interpretation. The bibliography of that paper gives several references to a variety of other medical application of fractal analysis.

4. S. S. Cross et al. in the Sheffield pathology department studied the application of fractal dimension to the renal arterial tree. Again, lower fractal dimension is associated with a diseased kidney.

5. Jianhua Wu, et al. ['Medical Image Retrieval Based on Fractal Dimension' 2008 The 9th International Conference for Young Computer Scientists] report as follows.

> 'Abstract: Content-based medical image retrieval becomes a hot research topic due to the rapid increase of image database. It is useful that a doctor consults analogical cases to diagnose for a patient. So it is very important for doctors to quickly and exactly search out the similar pathological images from large numbers of images in clinic. Fractal texture feature is introduced to medical images, according to experiments, it is discovered that the normal lung and several kinds of common lung diseases CT images have different fractal dimensions, which indicates that fractal dimensions of images can distinguish most lung diseases. Fractal feature is applied in medical images retrieval, and compared with general approaches, experiments show that high precision and recall of retrieval are achieved, and our method also can achieve a comparatively lower computation cost, and the retrieval time is short. The method is applied well and gives much better performance in medical images retrieval.'

Here is another way of characterizing Kolmogorov dimension, which is, in a sense, 'dual' to the original definition.

Theorem 11.10 *For a non-empty compact set K in a metric space and $\varepsilon > 0$, let $M(\varepsilon)$ be the largest number m for which there is set $\{x_1, x_2, \ldots, x_m\} \subseteq K$ with $d(x_i, x_j) \geqslant \varepsilon$ for all $i \neq j$. Then*

$$\lim_{\varepsilon \to 0} \frac{\log M(\varepsilon)}{\log(1/\varepsilon)}$$

exists if and only if $\mathrm{Kdim}K$ *exists, in which case they are equal.*

Proof. If $\{x_1, x_2, \ldots, x_m\} \subseteq K$ with $d(x_i, x_j) \geqslant \varepsilon$ for all $i \neq j$, and m maximal, then

$$K \subseteq B(x_1, \varepsilon) \cup B(x_2, \varepsilon) \cup \ldots \cup B(x_m, \varepsilon),$$

for if $x \in K$ were a point outside all the $B(x_i, \varepsilon)$, then $\{x_1, x_2, \ldots, x_m, x\}$ would be a strictly larger set with the distance between any pair of distinct elements greater than or equal to ε. It follows that $N(\varepsilon) \leqslant M(\varepsilon)$.

Conversely, if $\{x_1, x_2, \ldots, x_m\} \subseteq K$ with $d(x_i, x_j) \geqslant \varepsilon$ for all $i \neq j$, and if

$$K \subseteq B(y_1, \varepsilon/2) \cup B(y_2, \varepsilon/2) \cup \ldots \cup B(y_N, \varepsilon/2),$$

then no two distinct x_i can belong to the same $B(y_j, \varepsilon/2)$. Therefore, $m \leqslant N$. Hence $M(\varepsilon) \leqslant N(\varepsilon/2)$.

The theorem then follows from the Comparison Lemma by putting $A(\varepsilon) = N(\varepsilon)$, $B(\varepsilon) = M(\varepsilon)$, $\lambda_1 = 1$, $\mu_1 = 1$, $\lambda_2 = 1$, and $\mu_2 = 1/2$. $\Diamond$

Exercise 11.11 Show that an equivalent definition of Kolmogorov dimension is obtained if $N(\varepsilon)$ is replaced by the number of open balls of *diameter* ε needed to cover K.

We turn now to a quick discussion of the basic properties of Kolmogorov dimension.

Theorem 11.12 *(i) Every nonempty finite set has Kolmogorov dimension zero.*

(ii) Let $f : X \to Y$ be a Lipschitz mapping between metric spaces and K is a non-empty compact subset of X which has Kolmogorov dimension. Show that **if** *$f(K)$ (which is necessarily a non-empty compact subset of Y) also has Kolmogorov dimension, then* $\mathrm{Kdim}f(K) \leqslant \mathrm{Kdim}K$.

(iii) If $f : X \to Y$ is a bijective biLipschitz mapping between metric spaces and K is a non-empty compact subset of X which has Kolmogorov dimension, then $f(K)$ is a non-empty compact subset of Y which has Kolmogorov dimension, and $\mathrm{Kdim}f(K) = \mathrm{Kdim}K$.

(iv) Let $K \subseteq X, L \subseteq Y$ be non-empty compact subsets of metric spaces and suppose that K and L both have Kolmogorov dimension. Then the (non-empty compact) subset $K \times L \subseteq X \times Y$ has Kolmogorov dimension, and

$$\mathrm{Kdim}(K \times L) = \mathrm{Kdim}K + \mathrm{Kdim}L.$$

(v) Let $K_1 \subseteq K_2$ be non-empty compact subsets of a metric space X both of which have Kolmogorov dimension. Then

$$\mathrm{Kdim}K_1 \leqslant \mathrm{Kdim}K_2.$$

(vi) Let $A \subseteq K \subseteq B$ be non-empty compact subsets of a metric space X such that both A and B have Kolmogorov dimension and $\mathrm{Kdim}A = \mathrm{Kdim}B$. Then K has Kolmogorov dimension (and $\mathrm{Kdim}K = \mathrm{Kdim}A = \mathrm{Kdim}B$).

(vii) Let K_1, K_2 be non-empty compact subsets of a metric space X both of which have Kolmogorov dimension. Then $K_1 \cup K_2$ has Kolmogorov dimension and

$$\mathrm{Kdim}(K_1 \cup K_2) = \max\{\mathrm{Kdim}K_1, \mathrm{Kdim}K_2\}.$$

We omit the proofs.

Theorem 11.13 *For $n = 1, 2, 3, \ldots$, the n-dimensional unit cube $[0,1]^n$ has Kolmogorov dimension equal to n.*

Here, $[0,1]^n$ may denote $[0,1]^n$ as a subspace of $\mathbb{R}^n$ either in its usual metric or in the product metric. Since the two metrics are biLipschitz equivalent, the Kolmogorov dimension is the same. Further, $[0,1]^n$ could denote a subset of $\mathbb{R}^m$ for some $m \geqslant n$, in an obvious way, and the same result would hold, since there are Lipschitz maps $\mathbb{R}^m \to \mathbb{R}^n$ and $\mathbb{R}^n \to \mathbb{R}^m$ taking $[0,1]^n$ onto $[0,1]^n$.

Proof. We think of $[0,1]^n$ in the Euclidean metric on $\mathbb{R}^n$ and calculate its Kolmogorov dimension by the grid dimension method. If we form an ε-grid parallel to the sides of the cube $[0,1]^n$ with one corner at a grid point, then

$$N_g(\varepsilon) = ([1/\varepsilon] + 2)^n,$$

where $[.]$ denotes the integer part. Hence

$$\frac{\log N_g(\varepsilon)}{\log(1/\varepsilon)} \to n$$

as $\varepsilon \to 0$, and the result follows. $\Diamond$

This is where we reach the end of our eulogy of Kolmogorov dimension. It goes out of control when we take countably infinite unions with results which, we suggest, run counter to any intuition we may have about fractional dimensional sets. Consider the following simple example.

Example 11.14 Consider the compact set

$$K = \{\frac{1}{n} : n = 1, 2, 3, \ldots\} \cup \{0\} \subseteq \mathbb{R}.$$

Given $\varepsilon \in (0, 1/2)$, let $n = n(\varepsilon)$ be the positive integer such that

$$\frac{1}{2(n+1)^2} < \varepsilon \leqslant \frac{1}{2n^2},$$

then K can be covered by ε-balls centred on the $2n + 1$ points

$$\frac{k + \frac{1}{2}}{(n+1)^2} \qquad (0 \leqslant k \leqslant n)$$

and

$$\frac{1}{\ell} \qquad (1 \leqslant \ell \leqslant n).$$

Thus $N(\varepsilon) \leqslant 2n+1$.

Conversely, if $p \neq q$ and $p, q \leqslant n$, then

$$d\left(\frac{1}{p}, \frac{1}{q}\right) = \frac{|p-q|}{pq} > \frac{1}{n^2} \geqslant 2\varepsilon.$$

Hence, at least n ε-balls are required to cover K: so $N(\varepsilon) \geqslant n$. Therefore,

$$\begin{aligned} \frac{\log n}{\log 2 + 2\log(n+1)} &= \frac{\log n}{\log\left(2(n+1)^2\right)} \\ &\leqslant \frac{\log(N(\varepsilon)}{\log(1/\varepsilon)} \\ &\leqslant \frac{\log(2n+1)}{\log(2n^2)} \\ &= \frac{\log 2 + \log\left(n+\frac{1}{2}\right)}{\log 2 + 2\log n}. \end{aligned}$$

As $\varepsilon \to 0$, we have $n \to \infty$ and the first and last terms both tend to $\frac{1}{2}$, so the Kolmogorov dimension of K exists and is equal to $\frac{1}{2}$.

This example is deeply wounding to Kolmogorov dimension as a definition of dimension. It is a particularly simple set, accumulating only at one point. One expects its dimension, in any reasonable sense, to be that of a single point, namely zero. If Kolmogorov dimension behaved even as well as topological dimension under countable union, this would be so. We recall that if $Y_1, Y_2, \ldots$ is a sequence of closed subsets of a separable metric space X, then

$$\text{topdim}\left(\bigcup_{i=1}^{\infty} Y_i\right) = \max_i(\text{topdim} Y_i).$$

In particular, the topological dimension of any countable set is zero, since the topological dimension of a point is zero. However, the Kolmogorov dimension of a point is also zero, but we have here an example of a countable set of non-zero Kolmogorov dimension, so the analogous theorem for Kolmogorov dimension fails.

There is a contrary view however. Because Kolmogorov dimension does not take the same value for all countable sets, it is a more delicate invariant; one which is capable of expressing something of the geometry of different countable sets. It is noteworthy that recent work of Lapidus

> M. L. Lapidus,"Fractal drum, inverse spectral problems for elliptic operators and a partial resolution of the Weyl-Berry conjecture", *Trans. Amer. Math. Soc.,* **325** (1991), 465–528,
>
> Christina Q. He; Michel L. Lapidus, *Generalized Minkowski Content, Spectrum of Fractal Drums, Fractal Strings and the Riemann Zeta-Function*, (AMS, MEMO, **608**, 1997) 97pp., ISBN 0-8218-0597-5,

on the detailed asymptotics of eigenvalues of the Laplacian features the Minkowski dimension (= Kolmogorov dimension) of the boundary rather than the Hausdorff dimension to provide the solution.

Exercise 11.15 For a real number $s > 1$, consider the compact set

$$X = \{\frac{1}{n^s} : n = 1, 2, 3, \ldots\} \cup \{0\} \subseteq \mathbb{R}$$

as a subset of $\mathbb{R}$. Prove that $\text{Kdim} X = \frac{1}{s+1}$.

Exercise 11.16 Find the Kolmogorov dimension of the set

$$S = \{(x, y) \in \mathbb{R}^2 : x^2 + y^2 \in \{1/n^2 : n = 1, 2, 3, \ldots\} \cup \{0\}\}.$$

Exercise 11.17 Find a sequence of finite subsets $F_n \subseteq \mathbb{R}$ such that $F_n \to C$ as $n \to \infty$ in $\mathcal{H}_1$, where C is the Cantor ternary set. Deduce that it is not generally true that $A_n \to A$ implies $\mathrm{Kdim} A_n \to \mathrm{Kdim} A$.

Actually, this is far from being the easiest example of the discontinuity of dimension. Consider the sets $K_n = [-1/n, 1/n] \subseteq \mathbb{R}$. We have $\bigcap_{n=1}^{\infty} K_n = \{0\}$ and $K_n \to \{0\}$ in the sense of the Hausdorff metric (because $d_H(K_n, \{0\}) = 1/n \to 0$), yet $\mathrm{Kdim} K_n = 1$ for all n and $\mathrm{Kdim}\{0\} = 0$. This is equally true of topological dimension and of Hausdorff dimension (below).

In the next chapter, we turn to Hausdorff dimension: a more sophisticated notion than Kolmogorov dimension, giving the same results for most of the self-similar sets we wish to study, but capable of overcoming the obstacle of countably infinite unions.

12 Hausdorff dimension

In this chapter, we present an approach to dimension theory which avoids the problems we found associated with Kolmogorov dimension. This new idea, 'Hausdorff dimension' otherwise known as 'Hausdorff-Besicovitch dimension', is rooted in 'measure theory' — the theory of length, area and volume. For this presentation, we have modified the definitions to avoid formal measure theory. Our definitions are equivalent to the standard ones, based on Hausdorff measure, but it is hoped that they will be much easier to understand. The aura of measure theory still pervades our work, though, and the reader who wishes to pursue Hausdorff dimension theory much further will need to devote some time to learning that subject.

Definition 12.1 A set $A \subseteq \mathbb{R}$ is a *1-null set* (or a *set of linear measure zero* if, for every $\varepsilon > 0$, there is a covering of A by open intervals

$$A \subseteq \bigcup_{i=1}^{\infty} (a_i, b_i)$$

with

$$\sum_{i=1}^{\infty} (b_i - a_i) < \varepsilon.$$

Examples 12.2

1. The set $\mathbb{Q}$ of all rationals is 1-null. This follows from the fact that $\mathbb{Q}$ is countable, $\mathbb{Q} = \{q_1, q_2, q_2, \ldots\}$, for then

$$\mathbb{Q} \subseteq \bigcup_{i=1}^{\infty} (q_i - 2^{-i-1}\varepsilon, q_i + 2^{-i-1}\varepsilon).$$

   Indeed, every countable set is 1-null.

2. The Cantor ternary set C is 1-null. To see this, we note that the total length of the closed intervals comprising C_n is $(2/3)^n$, and $(2/3)^n \to 0$ as $n \to \infty$. Given ε, choose n so that $(2/3)^n < \varepsilon$ and then expand each of the intervals of C_n to a slightly larger open interval to obtain a covering of C by 2^n open intervals of total length less than ε.

The above definition for subsets of $\mathbb{R}$ may easily be modified to apply to subsets of metric spaces.

Definition 12.3 A subset A of a metric space X is *1-null* if, for every $\varepsilon > 0$, there is a covering of A by open balls

$$A \subseteq \bigcup_{i=1}^{\infty} B(x_i, \delta_i)$$

with the $\delta_i \geqslant 0$ and

$$\sum_{i=1}^{\infty} \delta_i < \varepsilon.$$

Here, and henceforth, we allow the δ_i to be possibly zero so that coverings by finitely many balls are legitimate. This is not really significant, but it makes things intuitively easier.

This enables us to talk about, for example, subsets of $\mathbb{R}^2$ being 1-null: intuitively, being "of zero length". To describe sets "of zero area" we have the notion of "2-null".

Definition 12.4 A subset A of a metric space X is *2-null* if, for every $\varepsilon > 0$, there is a covering of A by open balls

$$A \subseteq \bigcup_{i=1}^{\infty} B(x_i, \delta_i)$$

with

$$\sum_{i=1}^{\infty} \pi \delta_i^2 < \varepsilon.$$

Thus a circle, for example, is 2-null, but not 1-null, and this corresponds to the idea that its dimension is at least 1 but less than 2. To fix the dimension more accurately, we introduce the notion of "d-null" sets for all $d > 0$. We note first that the factor of π in the definition of 2-null can be omitted without altering the effect of the definition; likewise, there is no need for a factor $(4/3)\pi$ in the definition of 3-null and no need to consider the analogues for fractional d.

Definition 12.5 Let d be a positive real number. A subset A of a metric space X is *d-null* if, for every $\varepsilon > 0$, there is a covering of A by open balls

$$A \subseteq \bigcup_{i=1}^{\infty} B(x_i, \delta_i)$$

with

$$\sum_{i=1}^{\infty} \delta_i^d < \varepsilon.$$

Remark 12.6 We may, if we like, require $x_i \in A$ in the above definition without affecting the result. In fact, if we have a covering

$$A \subseteq \bigcup_{i=1}^{\infty} B(x_i, \delta_i)$$

with

$$\sum_{i=1}^{\infty} \delta_i^d < \varepsilon,$$

we may assume that each of the balls $B(x_i, \delta_i)$ contains a point $a_i \in A$, since otherwise we could drop that ball from the covering. Then $d(a_i, x_i) < \delta_i$, so $B(x_i, \delta_i) \subseteq B(a_i, 2\delta_i)$, so

$$A \subseteq \bigcup_{i=1}^{\infty} B(a_i, 2\delta_i)$$

with

$$\sum_{i=1}^{\infty} (2\delta_i)^d < 2^d \varepsilon,$$

and the number $2^d\varepsilon$ can be made arbitrarily small, by choosing ε sufficiently small.

Lemma 12.7 *If $d_1 \leqslant d_2$ and A is d_1-null, then A is d_2-null.*

Proof. For every $\varepsilon \in (0,1)$, there is a covering of A,

$$A \subseteq \bigcup_{i=1}^{\infty} B(x_i, \delta_i)$$

with

$$\sum_{i=1}^{\infty} \delta_i^{d_1} < \varepsilon.$$

Now for each i, $\delta_i^{d_1} < \varepsilon < 1$, so $\delta_i < 1$, so $\delta_i^{d_2} \leqslant \delta_i^{d_1}$. Therefore

$$\sum_{i=1}^{\infty} \delta_i^{d_2} < \varepsilon.$$

The fact that we have proved this for $0 < \varepsilon < 1$ is enough to show that A is d_2-null, since if it holds for some ε it certainly holds for any larger ε. The force of the definition is that it should hold however *small* ε might be. $\Diamond$

We can now define Hausdorff dimension.

Definition 12.8 Let A be a non-empty subset of a metric space, which need not be compact. The *Hausdorff dimension* of A is the number

$$\mathrm{Hdim}A = \inf\{d : A \text{ is } d\text{-null}\}.$$

If there is no d such that A is d-null, we write $\mathrm{Hdim}A = \infty$. If A is d-null for every $d > 0$, the definition gives $\mathrm{Hdim}A = 0$.

Proposition 12.9 *For every non-empty compact A having Kolmogorov dimension,* $\mathrm{Hdim}A \leqslant \mathrm{Kdim}A$.

Proof. Suppose, to the contrary, that $\mathrm{Hdim}A > \mathrm{Kdim}A$. Let $\eta = (\mathrm{Hdim}A - \mathrm{Kdim}A)/3$, and let $d = \mathrm{Kdim}A + 2\eta$. Then $d - \eta > \mathrm{Kdim}A$, so, for all sufficiently small $\varepsilon > 0$,

$$\frac{\log N(\varepsilon)}{\log(1/\varepsilon)} < d - \eta.$$

Therefore

$$N(\varepsilon) < \varepsilon^{-d+\eta}.$$

Thus there is a covering

$$A \subseteq \bigcup_{i=1}^{\infty} B(x_i, \delta_i)$$

with

$$\delta_i = \begin{cases} \varepsilon & (1 \leqslant i \leqslant N(\varepsilon)) \\ 0 & (N(\varepsilon) < i < \infty) \end{cases}$$

and so

$$\sum_{i=1}^{\infty} \delta_i^d \leqslant \varepsilon^d N(\varepsilon) < \varepsilon^{\eta}.$$

Now ε^{η} can be made arbitrarily small by making ε sufficiently small, so we have shown that A is d-null. However, $d < \mathrm{Hdim}A$, so this contradicts the definition of $\mathrm{Hdim}A$ and completes the proof. $\Diamond$

Corollary 12.10 *Every non-empty finite set has Hausdorff dimension zero.*

Notice the contrast between Kdim and Hdim. The latter is defined using *countable* coverings by balls which are *not of uniform radius*. These features account for the superiority of Hdim over Kdim.

Proposition 12.11 *Every non-empty countable set has Hausdorff dimension zero.*

Proof. Let $A = \{a_1, a_2, a_3, \ldots\}$ be countably infinite and let $d > 0$. For every $\varepsilon > 0$, we cover A,

$$A \subseteq \bigcup_{i=1}^{\infty} B(a_i, \delta_i),$$

where $\delta_i = (2^{-i}\varepsilon)^{1/d}$. Then

$$\sum_{i=1}^{\infty} \delta_i^d = \sum_{i=1}^{\infty} 2^{-i}\varepsilon = \varepsilon,$$

can be made arbitrarily small. Therefore A is d-null. Since this holds for all $d > 0$, $\mathrm{Hdim} A = 0$. $\Diamond$

In particular, the set $\{1/n : n = 1, 2, 3, \ldots\} \cup \{0\}$ has Hausdorff dimension zero. We saw in the last chapter that its Kolmogorov dimension was, rather unsatisfactorily, $1/2$. This makes us feel a little happier about Hausdorff dimension, and it shows that the inequality in Proposition 12.9 can be strict.

Theorem 12.12 *Let $f : X \to Y$ be a Lipschitz mapping between metric spaces and let $A \subseteq X$. Then* $\mathrm{Hdim} f(A) \leqslant \mathrm{Hdim} A$.

Proof. We must show that if $d > 0$ and A is d-null, then $f(A)$ is d-null.

Suppose $\mathrm{Lip}\, f = \lambda$; then, given $\varepsilon > 0$, if A is d-null there is a covering

$$A \subseteq \bigcup_{i=1}^{\infty} B(x_i, \delta_i),$$

with

$$\sum_{i=1}^{\infty} \delta_i^d < \varepsilon/\lambda^d.$$

Then

$$f(A) \subseteq \bigcup_{i=1}^{\infty} f(B(x_i, \delta_i)) \subseteq \bigcup_{i=1}^{\infty} B(f(x_i), \lambda\delta_i),$$

so we have a covering of $f(A)$. Moreover

$$\sum_{i=1}^{\infty} (\lambda\delta_i)^d < \varepsilon,$$

so we have proved that $f(A)$ is d-null. Therefore

$$\{d : A \text{ is } d\text{-null}\} \subseteq \{d : f(A) \text{ is } d\text{-null}\},$$

so

$$\begin{aligned} \mathrm{Hdim} A &= \inf\{d : A \text{ is } d\text{-null}\} \\ &\geqslant \inf\{d : f(A) \text{ is } d\text{-null}\} \\ &= \mathrm{Hdim} f(A). \end{aligned}$$

$\Diamond$

Corollary 12.13 *If f is biLipschitz, then* $\mathrm{Hdim} f(A) = \mathrm{Hdim} A$.

Theorem 12.14 *1. If $A \subseteq B$ are subsets of a metric space X, then* $\mathrm{Hdim} A \leqslant \mathrm{Hdim} B$.

2. If A_i $(i = 1, 2, 3, \ldots)$ are subsets of a metric space X, then

$$\mathrm{Hdim} \bigcup_{i=1}^{\infty} A_i = \sup_i (\mathrm{Hdim} A_i).$$

Proof.

1. Left as an exercise.

2. The first part of this theorem implies that, for each i,

$$\mathrm{Hdim}A_i \leqslant \mathrm{Hdim}\bigcup_{i=1}^{\infty} A_i.$$

It follows that

$$\sup_i(\mathrm{Hdim}A_i) \leqslant \mathrm{Hdim}\bigcup_{i=1}^{\infty} A_i.$$

To establish the converse inequality, we show that, for any $d > 0$,

$$d > \sup_i(\mathrm{Hdim}A_i) \;\;\Rightarrow\;\; d \geqslant \mathrm{Hdim}\bigcup_{i=1}^{\infty} A_i;$$

this follows from

$$d > \mathrm{Hdim}A_i \text{ for all } i \;\;\Rightarrow\;\; d \geqslant \mathrm{Hdim}\bigcup_{i=1}^{\infty} A_i;$$

which, in turn, follows from

$$A_i \text{ is } d\text{-null for all } i \;\;\Rightarrow\;\; \bigcup_i A_i \text{ is } d\text{-null.}$$

It is in this proof that we use, crucially, the fact that our definition of d-null allows coverings by infinitely many balls.

Let $\varepsilon > 0$. If A_i is d-null, we can cover it:

$$A_i \subseteq \bigcup_{j=1}^{\infty} B(x_{ij}, \delta_{ij})$$

with

$$\sum_{j=1}^{\infty} \delta_{ij}^d < 2^{-i}\varepsilon.$$

Then

$$\bigcup_{i=1}^{\infty} A_i \subseteq \bigcup_{i=1}^{\infty}\bigcup_{j=1}^{\infty} B(x_{ij}, \delta_{ij}),$$

with

$$\sum_{i,j=1}^{\infty} \delta_{ij}^d < \sum_{i=1}^{\infty} 2^{-i}\varepsilon = \varepsilon.$$

Thus $\bigcup_i A_i$ is d-null if all of the A_i are d-null, and the theorem is proved.

$\Diamond$

Theorem 12.15 *Let X, Y be metric spaces and $A \subseteq X$, $B \subseteq Y$ two non-empty subsets. Then*

$$\mathrm{Hdim}A \times B \geqslant \mathrm{Hdim}A + \mathrm{Hdim}B.$$

If either A has Kolmogorov dimension and $\mathrm{Kdim}A = \mathrm{Hdim}A$*, then*

$$\mathrm{Hdim}A \times B = \mathrm{Hdim}A + \mathrm{Hdim}B.$$

Proof omitted.

Exercise 12.16 Show that the unit interval $[0,1]$ has Hausdorff dimension 1: you can split this into two parts.

1. Show that, for $d = 1$,
$$[0,1] \subseteq \bigcup_{i\in I} B(a_i;\delta_i) \implies \sum_{i\in I} \delta_i^d \geqslant \frac{1}{2},$$
by using the compactness of $[0,1]$ to restrict attention to the case I finite.

2. Show that, for all $d > 1$ and $\varepsilon > 0$, there exists a cover
$$[0,1] \subseteq \bigcup_{i\in I} B(a_i;\delta_i)$$
with $\sum_{i\in I} \delta_i^d < \varepsilon$.

It follows from Theorem 1 that any subset of $\mathbb{R}$ which contains a non-trivial interval has Hausdorff dimension 1.

It may be shown, in similar fashion, that any subset of $\mathbb{R}^n$ which contains a non-trivial n-cube has Hausdorff dimension n.

13 Dimensions of self-similar sets

In this chapter, we seek to evaluate the Hausdorff dimensions of attractors of IFSs. First, let us give a name to the answer we hope to obtain.

Definition 13.1 Let $\mathcal{W} = \{w_1, w_2, \ldots, w_M\}$ be an IFS in $\mathbb{R}^N$ with $\operatorname{Lip} w_i = s_i \in (0,1)$ $(1 \leqslant i \leqslant M)$. The *similarity dimension* of $\mathcal{W}$ is the unique solution D of the equation
$$\sum_{i=1}^{M} s_i^D = 1 \tag{8}$$

Notice that the left hand side of (8) is a continuous, strictly decreasing function of D which is M at $D = 0$ and tends to zero as D tends to infinity; hence there is a unique solution.

If $s_1 = s_2 = \ldots = s_M = s$, then (8) reduces to $Ms^D = 1$, i.e. $\log M + D \log s = 0$,
$$D = \frac{\log M}{\log 1/s}.$$

Examples 13.2 1. The Cantor Ternary Set is the attractor of a set of two contractions, one contracting by a factor of $1/3$ towards 0, the other by the same factor towards 1. Thus $s = 1/3$, $M = 2$ and $D = (\log 2)/(\log 3)$.

2. The Koch curve: $s = 1/3$, $M = 4$ and $D = (\log 4)/(\log 3)$.

3. The quadric Koch curve: $s = 1/4$, $M = 8$ and $D = (\log 8)/(\log 4) = 3/2$.

Theorem 13.3 *Let $\mathcal{W} = \{w_1, w_2, \ldots, w_M\}$ be an IFS on $\mathbb{R}^N$ with attractor A and similarity dimension D. Then* $\operatorname{Hdim} A \leqslant D$.

Proof. The set A is compact and so bounded; i.e. $A \subseteq B(x,\rho)$ for some $x \in \mathbb{R}^N$ and $\rho > 0$. If $\operatorname{Lip} w_i = s_i$ $(1 \leqslant i \leqslant M)$, then $w_i(B(x,\rho)) \subseteq B(w_i(x), s_i\rho)$. Therefore
$$A = \bigcup_{i=1}^{M} w_i(A) \subseteq \bigcup_{i=1}^{M} w_i(B(x,\rho)) \subseteq \bigcup_{i=1}^{M} B(w_i(x), s_i\rho).$$

Repeating the argument:

$$\begin{aligned} A &= \bigcup_{j=1}^{M} w_j(A) \\ &\subseteq \bigcup_{j=1}^{M}\bigcup_{i=1}^{M} w_j(B(w_i(x), s_i\rho)) \\ &\subseteq \bigcup_{i,j=1}^{M} B(w_j w_i(x), s_j s_i \rho). \end{aligned}$$

Generally,

$$A \subseteq \bigcup_{i_1,\ldots,i_n=1}^{M} B(w_{i_1}\ldots w_{i_n}(x), s_{i_1}\ldots s_{i_n}\rho). \tag{9}$$

Now, for every $d > D$, $\sum_{i=1}^{M} s_i^d < 1$, so

$$\sum_{i_1,\ldots,i_n=1}^{M} (s_{i_1}\ldots s_{i_n}\rho)^d = \rho^d \left(\sum_{i=1}^{M} s_i^d\right)^n$$

tends to zero as $n \to \infty$. Thus A is d-null for all $d > D$; so $\mathrm{Hdim} A \leqslant D$. $\Diamond$

Corollary 13.4 *If* $s_1 = s_2 = \ldots = s_M = s$ *and if* $\mathrm{Kdim} A$ *exists, then* $\mathrm{Kdim} A \leqslant D$.

Proof. In this case, (9) becomes

$$A \subseteq \bigcup_{i_1,\ldots,i_n=1}^{M} B(w_{i_1}\ldots w_{i_n}(x), s^n\rho).$$

Thus $N(\varepsilon) \leqslant M^n$ when $\varepsilon \geqslant s^n\rho$. Using the estimate $N(\varepsilon) \leqslant M^n$ when $s^n\rho \leqslant \varepsilon < s^{n-1}\rho$, (i.e. letting $n = \log(\varepsilon/\rho)/\log s$ if this is an integer and $[\log(\varepsilon/\rho)/\log s] + 1$ otherwise), we have

$$\log(1/\varepsilon) \geqslant (n-1)\log(1/s) + \log(1/\rho),$$

so

$$\begin{aligned} \frac{\log N(\varepsilon)}{\log(1/\varepsilon)} &\leqslant \frac{n \log M}{(n-1)\log(1/s) + \log(1/\rho)} \\ &\to \frac{\log M}{\log(1/s)} = D, \end{aligned}$$

as $\varepsilon \to 0$. Hence $\mathrm{Kdim} A \leqslant D$. $\Diamond$

Proving that the Hausdorff dimension is greater than or equal to the similarity dimension is much more difficult. In fact, it is not true, in general. Let us consider some examples.

Example 13.5 Let w_1, w_2, w_3, w_4 be the similitudes defining the Koch curve K and let π_1 be the orthogonal projection onto the straight line joining the end-points of K. Then the IFS

$$\{w_1\pi_1, w_2\pi_1, w_3\pi_1, w_4\pi_1\}$$

has as its attractor stage 1 in the construction of K: a curve consisting of four straight line segments. The Hausdorff dimension of this attractor is clearly 1, but the similarity dimension of the IFS is the same as that of $\{w_1, w_2, w_3, w_4\}$, namely, $(\log 4)/(\log 3)$. The problem here is that though $\mathrm{Lip}\, w_i\pi_1 = 1/3$, the map $w_i\pi_1$ compresses by $1/3$ in one direction and to zero in the other. The moral is that we should restrict our attention to *similitudes*.

Convention. Henceforth,

$$\mathcal{W} = \{w_1, w_2, \ldots, w_M\}$$

will be an IFS on $\mathbb{R}^N$, with $w_1, w_2, \ldots, w_M$ similitudes.

Example 13.6 Let $\{w_1, w_2, \ldots, w_{15}\}$ be an IFS in $\mathbb{R}^2$ consisting of 15 similitudes, each shrinking by a factor of $1/3$: $M = 15, 1/s = 3$, so $D = (\log 15)/(\log 3) > 2$.

This cannot be the Hausdorff dimension, since $\mathrm{Hdim} A \leqslant \mathrm{Hdim}\mathbb{R}^2 = 2$. Clearly what has gone wrong here is that the images $w_i(A)$ of the attractor overlap.

The overlap problem is the most difficult one. The easiest way round it involves restricting attention to systems with so little overlap that their attractors have topological dimension zero. However, we certainly need to do better than that, because we want to find the Hausdorff dimensions of topologically one-dimensional objects such as the Koch curve. The key definition turns out to be the following.

Definition 13.7 The IFS $\mathcal{W} = \{w_1, w_2, \ldots, w_M\}$ on $\mathbb{R}^N$ satisfies the *open set condition* if there exists a non-empty bounded open set $O \subseteq \mathbb{R}^N$ such that

1.

$$\bigcup_{i=1}^{M} w_i(O) \subseteq O,$$

2.

$$w_i(O) \cap w_j(O) = \emptyset \quad (i \neq j).$$

Examples 13.8

1. The Cantor ternary set is the attractor of an IFS $\{w_0, w_1\}$ (as given in chapter 8) which satisfies the open set condition with $O = (0, 1)$
2. To make the IFS $\{w_1, w_2, w_3, w_4\}$ generating the Koch curve satisfy the open set condition, we take O to be the open triangle with vertices $(0, 0)$, $(1/2, 1/2\sqrt{3})$, $(1, 0)$. Notice how this condition depends on the fact that the Koch IFS does not crumple the unit interval too much: the angles produced are not too acute. A little consideration of this and similar examples will convince the reader that the open set condition is, indeed, a "non-overlap condition".

Theorem 13.9 (Hutchinson's Theorem) *Let $\mathcal{W}$ be an IFS on $\mathbb{R}^N$ whose contractions are similitudes and which satisfies the open set condition, and let D be its similarity dimension and A its attractor. Then* $\mathrm{Hdim} A = D$.

Hutchinson's proof may be found in

> J. E. Hutchinson "Fractals and self-similarity" *Indiana University Mathematics Journal,* **30** (1981), 713 – 747.

It uses a fair amount of measure theory and whilst measure-theoretic ideas are essential to the proof, we shall try to give an account which minimizes this aspect. *The remainder of this chapter is not examinable as bookwork.*

Proof of Theorem. We begin with a technical lemma.

Lemma 13.10 *Suppose $0 < c < C < \infty$, $x \in \mathbb{R}^N$ and $0 < \rho < \infty$. Let $\{U_i\}$ be a family of disjoint open sets in $\mathbb{R}^N$. Suppose each U_i contains a ball of radius $c\rho$ and is contained in a ball of radius $C\rho$. Then at most $(1 + 2C)^N c^{-N}$ of the $\overline{U_i}$ meet the ball $B(x, \rho)$.*

Proof of Lemma. Without loss of generality, let the given ball be $B(x, \rho)$. Suppose $\overline{U_1}, \ldots, \overline{U_k}$ meet $B(x, \rho)$. Since each of the $\overline{U_i}$ has diameter less than $2C\rho$, it follows that each of $\overline{U_1}, \ldots, \overline{U_k}$ is contained in $B(x, (1 + 2C)\rho)$.

Let $\alpha_N r^N$ denote the N-dimensional volume of a ball of radius r: thus $\alpha_1 = 2$, $\alpha_2 = \pi$, $\alpha_3 = \frac{4}{3}\pi$, and generally,

$$\alpha_N = \frac{\Gamma(1/2)^N}{\Gamma((N/2)+1)},$$

where $\Gamma(.)$ is the Γ-function. However, the precise value of α_N is quite irrelevant: all that matters is that it is a number depending on N.

The U_i are disjoint and each contains a ball of radius $c\rho$. These balls are therefore disjoint and, since they are all contained in the ball $B(x,(1+2C)\rho)$, their total volume is at most the volume of the large ball:

$$k\alpha_N c^N \rho^N \leqslant \alpha_N (1+2C)^N \rho^N.$$

Hence $k \leqslant (1+2C)^N c^{-N}$. $\Diamond$

Let $\mathcal{W} = (w_1, w_2, \ldots, w_M)$, with $\operatorname{Lip} w_i = s_i$, ordered so that $s_1 = \min_i s_i$. For every finite sequence $(i_1, i_2, \ldots, i_p)$, we define

$$A_{i_1 \ldots i_p} = w_{i_1} \circ \ldots \circ w_{i_p}(A).$$

Let O be a non-empty bounded open set such that

1. $W(O) = \bigcup_{i=1}^M w_i(O) \subseteq O$ and
2. $w_i(O) \cap w_j(O) = \emptyset$ $(i \neq j)$.

For every finite sequence $(i_1, i_2, \ldots, i_p)$, we define

$$O_{i_1 \ldots i_p} = w_{i_1} \circ \ldots \circ w_{i_p}(O).$$

Now $w_{i_1} \circ \ldots \circ w_{i_p}$ is a homeomorphism, so

$$\overline{O_{i_1 \ldots i_p}} = w_{i_1} \circ \ldots \circ w_{i_p}(\overline{O}),$$

and the notation

$$\overline{O}_{i_1 \ldots i_p}$$

for this set is unambiguous. Since O is non-empty and bounded, its closure $\overline{O}$ is non-empty and compact. Moreover,

$$W(\overline{O}) = \overline{W(O)} \subseteq \overline{O}.$$

Therefore, $W^n(\overline{O}) \subseteq \overline{O}$; i.e. $\rho(W^n(\overline{O}), \overline{O}) = 0$, for all n. But $W^n(\overline{O}) \to A$ in the Hausdorff metric; so

$$\rho(A, \overline{O}) \leqslant \rho(A, W^n(\overline{O})) \leqslant d(A, W^n(\overline{O})) \to 0.$$

Therefore $\rho(A, \overline{O}) = 0$; i.e.

$$A \subseteq \overline{O}.$$

Applying $w_{i_1} \circ \ldots \circ w_{i_p}$ to this inclusion shows that

$$A_{i_1 \ldots i_p} \subseteq \overline{O}_{i_1 \ldots i_p}.$$

We now form a set I of finite sequences $(j_1, \ldots, j_q)$ of integers in $[1, M]$, as follows. We take each *infinite* sequence $(j_1, \ldots, j_q, \ldots)$ of integers in $[1, M]$ and truncate it to $(j_1, \ldots, j_q)$ using the least q such that

$$s_{j_1} \ldots s_{j_q} \leqslant \rho.$$

Since $s_1 = \min_i s_i$, this implies that

$$s_1 \rho < s_{j_1} \ldots s_{j_q} \leqslant \rho.$$

This set I has the property that no member of it is an initial segment of another member: i.e. if $(i_1, \ldots, i_p), (j_1, \ldots, j_q) \in I$ then **neither**

1. $p \leqslant q$ and $(i_1, \ldots, i_p) = (j_1, \ldots, j_p)$ **nor**

2. $q \leqslant p$ and $(i_1, \ldots, i_q) = (j_1, \ldots, j_q)$

holds.

Consider now the implications of the second hypothesis on the set O:

$$w_i(O) \cap w_j(O) = \emptyset \qquad (i \neq j).$$

Applying the bijection $w_{i_1} \circ \ldots \circ w_{i_r}$ to both sides,

$$w_{i_1} \circ \ldots \circ w_{i_r} \circ w_i(O) \cap w_{i_1} \circ \ldots \circ w_{i_r} \circ w_j(O) = \emptyset \qquad (i \neq j).$$

Since $w_k(O) \subseteq O$ for all k, it follows that

$$O_{i_1 \ldots i_p} \cap O_{j_1 \ldots j_q} = \emptyset$$

unless one of $(i_1, \ldots, i_p)$ and $(j_1, \ldots, j_q)$ is an initial segment of the other.

Thus the sets

$$O_{j_1 \ldots j_q} \qquad ((j_1, \ldots, j_q) \in I)$$

are pairwise disjoint.

Suppose O contains a ball of radius c_1 and is contained in a ball of radius c_2. Then, since the w_j are similitudes, each $O_{j_1 \ldots j_q}$ contains a ball of radius $s_{j_1} \ldots s_{j_q} c_1$ and hence a ball of radius $s_1 \rho c_1$, and is contained in a ball of radius $s_{j_1} \ldots s_{j_q} c_2$ and hence in a ball of radius ρc_2.

Given $x \in \mathbb{R}^N$, it follows from Lemma 13.10, with $c = s_1 c_1$, $C = c_2$, that at most $(1+2c_2)^N (s_1 c_1)^{-N}$ of the sets $\overline{O}_{j_1 \ldots j_q}$, $((j_1, \ldots, j_q) \in I)$ meet the ball $B(x, \rho)$. Hence, at most $(1 + 2c_2)^N (s_1 c_1)^{-N}$ of the sets $A_{j_1 \ldots j_q}$, $((j_1, \ldots, j_q) \in I)$ meet the ball $B(x, \rho)$.

Imagine choosing a sequence $i_1, i_2, \ldots$ of numbers between 1 and M, each i_j being chosen randomly with the probability of $i_j = r$ being s_r^D for $1 \leqslant r \leqslant M$. (In probabilistic language we have a sequence of independent random variables X_i with $\mathcal{P}(X_i = r) = s_r^D$.) Notice that we are using the definition of D as the number such that

$$\sum_{r=1}^{M} s_r^D = 1$$

to tell us that this is a probability distribution.

Given a set $E \subseteq \mathbb{R}^N$, we define $\mu(E)$ to be the probability that the sequence chosen is such that

$$A_{i_1 \ldots i_p} \cap E \neq \emptyset$$

for some p. We claim (without proof) the following facts about μ.

(A) $\mu(E)$ is defined for (at least) all open sets and all closed sets.

(B) $\mu(\mathbb{R}^N) = \mu(A) = 1$.

(C) If $E \subseteq \bigcup_{i=1}^{\infty} E_i$ and $\mu(E)$, $\mu(E_i)$ $(1 \leqslant i < \infty)$ are defined, then

$$\mu(E) \leqslant \sum_{i=1}^{\infty} \mu(E_i).$$

We now estimate $\mu(B(x, \rho))$. The probability of selecting a sequence with initial segment $(j_1, \ldots, j_q)$ is

$$s_{j_1}^D \ldots s_{j_q}^D \leqslant \rho^D$$

when $(j_1, \ldots, j_q) \in I$. Therefore

$$\mu(B(x, \rho)) \leqslant \frac{(1 + 2c_2)^N}{(s_1 c_1)^N} \rho^D. \tag{10}$$

Now suppose

$$A \subseteq \bigcup_{i=1}^{\infty} B(x_i, \rho_i).$$

Then

$$\begin{aligned} 1 &= \mu(A) \\ &\leqslant \sum_{i=1}^{\infty} \mu(B(x_i, \rho_i)), \qquad \text{by (C)}, \\ &\leqslant \frac{(1+2c_2)^N}{(s_1 c_1)^N} \sum_{i=1}^{\infty} \rho_i^D, \qquad \text{by (10)}. \end{aligned}$$

This shows that A is not D-null, so, in particular, $\mathrm{Hdim} A \geqslant D$. This, combined with Theorem 13.3, gives the desired result. $\Diamond$

For a discussion of a fractal—the "Barnsley wreath"—which falls outside the scope of Hutchinson's Theorem, see

G. A. Edgar "A fractal puzzle" *The Mathematical Intelligencer,* **13** (1991), 44 – 50.

14 Fractal curves

Some famous examples of fractals — the Koch curve and the quadric Koch curve — are parametrised curves; that is, they are images of continuous mappings of the unit interval into $\mathbb{R}^n$, in these cases, $\mathbb{R}^2$. We now present a sufficient condition on an IFS for its attractor to be such a curve.

Theorem 14.1 *Let $\mathcal{W} = \{w_1, w_2, \ldots, w_M\}$ be an IFS on $\mathbb{R}^n$. Let a be the fixed point of w_1 and b the fixed point of w_M. Suppose that*

$$w_i(b) = w_{i+1}(a) \qquad (1 \leqslant i \leqslant M-1).$$

Then there is a continuous map $f : [0,1] \to \mathbb{R}^n$ whose image is the attractor of $\mathcal{W}$.

Proof. Let $C([0,1], \mathbb{R}^n)$ be the space of all continuous functions $f : [0,1] \to \mathbb{R}^n$ with the metric

$$d(f,g) = \sup_{0 \leqslant t \leqslant 1} d(f(t), g(t)),$$

where the d on the right-hand side is, of course, the usual metric in $\mathbb{R}^n$.

Exercise 14.2 Show that $C([0,1], \mathbb{R}^n)$ is complete (using the fact that $\mathbb{R}^n$ is complete).

Now let

$$\mathcal{F} = \mathcal{F}(a,b) = \{f \in C([0,1], \mathbb{R}^n) : f(0) = a \text{ and } f(1) = b\}.$$

Then $\mathcal{F}$ is a closed subset of $C([0,1], \mathbb{R}^n)$. (Proof: if $f_i \in \mathcal{F}$ $(i = 1, 2, 3, \ldots)$ and $f_i \to f$ in $C([0,1], \mathbb{R}^n)$, then $f_i(t) \to f(t)$ for each $t \in [0,1]$ and, in particular, $a = f_i(0) \to f(0)$ and $b = f_i(1) \to f(1)$, so $f(0) = a$ and $f(1) = b$.) It follows that $\mathcal{F}$ is a complete metric space.

For $i = 1, 2, \ldots, M$, define $g_i : [(i-1)/M, i/M] \to [0,1]$ by $g_i(t) = Mt - i + 1$. Define $V(f)$ for $f \in \mathcal{F}$ by

$$V(f)(t) = w_i(f(g_i(t))) \qquad (t \in [(i-1)/M, i/M], 1 \leqslant i \leqslant M).$$

We need to show that $V(f)(t)$ is well-defined, since we have defined it twice at the points i/M $(1 \leqslant i \leqslant M-1)$. At these points we have

$$\begin{aligned} w_i(f(g_i(i/M))) &= w_i(f(1)) \\ &= w_i(b), \text{ since } f \in \mathcal{F}, \\ &= w_{i+1}(a), \text{ by the hypothesis of the theorem}, \\ &= w_{i+1}(f(0)), \text{ since } f \in \mathcal{F}, \\ &= w_{i+1}(f(g_{i+1}(i/M)). \end{aligned}$$

We show that $V(f) \in \mathcal{F}$. Certainly $V(f)$ is continuous because it is a composition of continuous functions in each of the closed intervals $[(i-1)/M, i/M]$, $(1 \leqslant i \leqslant M)$. To show that $V(f)(0) = a$ and $V(f)(1) = b$ we have:

$$\begin{aligned} V(f)(0) &= w_1(f(g_1(0))) \\ &= w_1(f(0)) \\ &= w_1(a), \text{ since } f \in \mathcal{F}, \\ &= a, \end{aligned}$$

since a is, by definition, the fixed point of w_1; the proof that $V(f)(1) = b$ is similar.

Lemma 14.3 *The map V is a contraction on $\mathcal{F}$.*

Proof. Let $s_i = \operatorname{Lip} w_i$ $(1 \leqslant i \leqslant M)$ and let $s = \max s_i$. Let $f_1, f_2 \in \mathcal{F}$ and $t \in [0,1]$. Let i be such that $t \in [(i-1)/M, i/M]$. Then

$$\begin{aligned} d(V(f_1)(t), V(f_2)(t)) &= d(w_i(f_1(g_i(t))), w_i(f_2(g_i(t)))) \\ &\leqslant s_i d(f_1(g_i(t)), f_2(g_i(t))) \\ &\leqslant s_i d(f_1, f_2). \end{aligned}$$

Taking the supremum over all $t \in [0,1]$ gives

$$d(V(f_1), V(f_2)) \leqslant s d(f_1, f_2).$$

Thus V is Lipschitz with $\operatorname{Lip} V \leqslant s$. Since $\mathcal{W}$ is an IFS, $s < 1$ and V is a contraction, which proves the lemma. $\Diamond$

The next step is inevitable: we apply the Contraction Mapping Principle. The conclusion is that there is a unique $f \in \mathcal{F}$ such that $V(f) = f$. The image $f([0,1])$ of f is a non-empty compact set (since [0,1] is compact and f is continuous). Let us consider the action of the map $W : \mathcal{H}_N \to \mathcal{H}_N$ on $f([0,1]) \in \mathcal{H}_N$.

$$\begin{aligned} W(f([0,1])) &= \bigcup_{i=1}^{M} w_i(f([0,1])) \\ &= \bigcup_{i=1}^{M} w_i(f(g_i([(i-1)/M, i/M]))) \\ &= \bigcup_{i=1}^{M} V(f)([(i-1)/M, i/M]) \\ &= \bigcup_{i=1}^{M} f([(i-1)/M, i/M]) \\ &= f([0,1]). \end{aligned}$$

However, there is only one non-empty compact set A with $W(A) = A$, namely, the attractor of $\mathcal{W}$, so we have shown that $f([0,1])$ is the attractor, and the theorem is proved. $\Diamond$

Remark 14.4 In fact a stronger result than Theorem 14.1 is true: if the attractor of an IFS is connected, then it is a fractal curve. However, Theorem 14.1 has the advantage of providing a means to see the attractor as a curve by finding successive approximations to it, as the next example shows.

Examples 14.5 1. We recall that the Koch curve was defined earlier as the attractor of the IFS $\{w_1, w_2, w_3, w_4\}$ on $\mathbb{R}^2$ such that each w_i is an orientation-preserving similitude with $\operatorname{Lip} w_i = 1/3$ and

$$\begin{aligned} w_1(0,0) &= (0,0), & w_1(1,0) &= (1/3, 0), \\ w_2(0,0) &= (1/3, 0), & w_2(1,0) &= (1/2, 1/2\sqrt{3}), \\ w_3(0,0) &= (1/2, 1/2\sqrt{3}), & w_3(1,0) &= (2/3, 0), \\ w_4(0,0) &= (2/3, 0), & w_4(1,0) &= (1,0). \end{aligned}$$

Clearly, this IFS satisfies the hypothesis of Theorem 14.1, with $a = (0,0)$ and $b = (1,0)$. We can even see the construction of a sequence of approximating curves through the proof of the theorem. Let $f_0 \in \mathcal{F}$ be the straight line

$$f_0(t) = (t, 0) \qquad (t \in [0,1]).$$

If we substitute the usual proof of the Contraction Mapping Principle into the proof of Theorem 14.1, we see that the fixed point f of V is $\lim_{i\to\infty} f_i$ where $f_i = V^i(f_0)$ $(i = 1, 2, 3, \ldots)$. These f_i form the usual sequence of approximations to f.

The alternative definition of the Koch curve as the attractor of an IFS consisting of two orientation-reversing similitudes provides another example and another sequence of approximating curves. Notice, however, that the Koch curve is also the attractor of IFS's which do not satisfy the hypothesis of Theorem 14.1: for example, if we replace w_3 by $w_3'(x, y) = w_3(1 - x, y)$ in the first IFS, we still have an IFS for which the Koch curve is self-similar, but the hypothesis of Theorem 14.1 no longer holds.

2. The Cantor ternary set C is the attractor of an IFS $\{w_0, w_1\}$ on $\mathbb{R}$ given by:

$$\begin{aligned} w_0(x) &= x/3, \\ w_1(x) &= (2 + x)/3. \end{aligned}$$

This IFS clearly does not satisfy the hypothesis of Theorem 14.1. Indeed, in this case, no IFS can do so, since there is no continuous map $f : [0,1] \to C$ (by the Intermediate Value Theorem).

3. The *quadric Koch curve* has not yet received a mention, so let us introduce it here. It is the attractor of an IFS $\{w_1, w_2, \ldots, w_8\}$ on $\mathbb{R}^2$ such that each w_i is an orientation-preserving similitude with $\operatorname{Lip} w_i = 1/4$ and

$$\begin{array}{lcllcl} w_1(0,0) &=& (0,0), & w_1(1,0) &=& (1/4,0), \\ w_2(0,0) &=& (1/4,0), & w_2(1,0) &=& (1/4,1/4), \\ w_3(0,0) &=& (1/4,1/4), & w_3(1,0) &=& (1/2,1/4), \\ w_4(0,0) &=& (1/2,1/4), & w_4(1,0) &=& (1/2,0), \\ w_5(0,0) &=& (1/2,0), & w_5(1,0) &=& (1/2,-1/4), \\ w_6(0,0) &=& (1/2,-1/4), & w_6(1,0) &=& (3/4,-1/4), \\ w_7(0,0) &=& (3/4,-1/4), & w_7(1,0) &=& (3/4,0), \\ w_8(0,0) &=& (3/4,0), & w_8(1,0) &=& (1,0). \end{array}$$

Its dimension is $(\log 8)/(\log 4) = 3/2$ and this IFS clearly satisfies the hypothesis of Theorem 14.1, so it is a continuously parameterized curve.

The *quadric Koch island* is formed by joining together four quadric Koch curves.

Exercise 14.6 Show that the Sierpinski triangle (with outer boundary an equilateral triangle) is a continuously parameterized curve.

Example 14.7 Space-filling curves The closed square $[0,1]^2$ can be represented as the attractor of an IFS in various ways and, by careful choice of the IFS we can satisfy the hypothesis of Theorem 14.1. For example, consider the IFS $\mathcal{W} = \{w_1, w_2, w_3, w_4\}$ on $\mathbb{R}^2$ where w_1, w_2, w_3, w_4 are similitudes with Lipschitz constant $1/2$; w_1 and w_4 are orientation-reversing, w_2 and w_3 are orientation-preserving and

$$\begin{array}{lcllcl} w_1(0,0) &=& (0,0), & w_1(1,0) &=& (0,1/2), \\ w_2(0,0) &=& (0,1/2), & w_2(1,0) &=& (1/2,1/2), \\ w_3(0,0) &=& (1/2,1/2), & w_3(1,0) &=& (1,1/2), \\ w_4(0,0) &=& (1,1/2), & w_4(1,0) &=& (1,0). \end{array}$$

We conclude that there is a continuous surjection $f : [0,1] \to [0,1]^2$; a *space-filling curve*. Notice also that the open set condition is satisfied with O being the open unit square $(0,1)^2$. We can therefore

deduce that the Hausdorff dimension of the square is the similarity dimension of this IFS, which is $(\log 4)/(\log 2) = 2$. Good!

Another example is the IFS $\mathcal{W} = \{w_1, w_2, \ldots, w_9\}$ on $\mathbb{R}^2$ where all the w_i are orientation-preserving similitudes with Lipschitz constant $1/3$ and

$$\begin{array}{lcllcl}
w_1(0,0) & = & (0,0), & w_1(1,1) & = & (1/3,1/3),\\
w_2(0,0) & = & (1/3,1/3), & w_2(1,1) & = & (0,2/3),\\
w_3(0,0) & = & (0,2/3), & w_3(1,1) & = & (1/3,1),\\
w_4(0,0) & = & (1/3,1), & w_4(1,1) & = & (2/3,2/3),\\
w_5(0,0) & = & (2/3,2/3), & w_5(1,1) & = & (1/3,1/3),\\
w_6(0,0) & = & (1/3,1/3), & w_6(1,1) & = & (2/3,0),\\
w_7(0,0) & = & (2/3,0), & w_7(1,1) & = & (1,1/3),\\
w_8(0,0) & = & (1,1/3), & w_8(1,1) & = & (2/3,2/3),\\
w_9(0,0) & = & (2/3,2/3), & w_9(1,1) & = & (1,1).
\end{array}$$

This gives another space-filling curve. Again, the open set condition is satisfied with $O = (0,1)^2$ and so the Hausdorff dimension of the square is the similarity dimension of this IFS, which is $(\log 9)/(\log 3) = 2$.

Perhaps we should call these "area-filling" curves. If you want to fill a cube, or higher dimensional hypercube, it is easy to find suitable IFS's.

Non-examinable proofs

The following pieces of text will not be examined as bookwork.

1. The proof of Theorem (8.6), the completeness of $\mathcal{H}_N$, including Lemma (8.7).
2. All of the chapter on Topological Dimension.
3. The proof of Theorem (13.9), Hutchinson's Theorem, as indicated in the duplicated notes.
4. The proof of Theorem (14.1).

Note that you will be expected to know the statements of Theorems (8.6), (13.9) and (14.1) and be able to apply these results.

In general you can expect the examination paper to test your knowledge of basic **definitions**, without which you would not know what you were talking about, then the statements of **key theorems**, and then the ability to solve problems and reconstruct proofs encountered in lectures. Recent past papers should be a good guide to the type of questions to be set, except for small changes in the syllabus: Question 4 parts (ii) and (iv) of the 2008 paper are no longer in the syllabus; the 2009 paper is a good guide. Note that there will be a compulsory question 1, as in previous years.