Fractal Geometry

From Self-Similarity to Brownian Motion

Based on a course taught in the winter term 2000/01
by Peter Mörters

Universität Kaiserslautern

Contents

Preface

The present volume represents my notes for the lecture *Fractal Geometry* held at Universität Kaiserlautern in the winter term 2000/2001. The aim of the lecture was

- to show that fractal geometry is a very diverse and colourful mathematical subject, comprising much more than just calendar pictures.
- to present rigorous and interesting mathematics while retaining the joy of the beautiful pictures and surprising statements fractal geometry can offer.
- to present many general mathematical principles, which can be rediscovered in other contexts: like the ergodic theorem, contraction principle, the max-flow min-cut theorem, branching processes, Brownian motion,....
- to introduce the audience into the world of random phenomena.

Of course the material chosen is my very personal selection and it is easy to find serious omissions (complex dynamics, for example). In brief, I chose to make a journey starting from Hutchinson's notion of self-similarity and ending with the study of fractals derived from Borwnian motion. The first chapter is devoted to *self-similarity*, self-similar sets are constructed using first the *contraction principle* (i.e. Banach's fixed point theorem) and then the *chaos game* based on the ergodic theorem. I use the opportunity to give a modern proof of *Birkhoff's ergodic theorem*, which is nowadays rarely presented in probability lectures. Both proofs can be used as algorithms to bring self-similar sets to the computer screen.

Next, I introduce and compare notions of dimension, highlighting the prominent role of Hausdorff dimension (a chapter on a class of fractals where Hausdorff and packing dimension naturally differ had to be omitted due to lack of time, unfortunately). Techniques for calculating the Hausdorff dimension are introduced, most notably the *potential theoretic method*. As a first application we derive the formula for the dimension of self-similar sets with open-set condition.

The third chapter is devoted to a brief discussion of some aspects of the geometry of self-simlilar sets. Highlights are the recent proof of Peres, Simon and Solomyak for the invisibility of certain self-similar 1-sets, which is presented in full detail. We use this result to construct *Besicovitch sets*, one of the most striking examples of fractal geometry.

In the fourth chapter I discuss a rich and beautiful class of random fractals, the so-called *Galton-Watson fractals*. The necessary tools from graph theory and probability on graphs are derived and the chapter culminates in an applicable dimension formula.

In the fifth and final chapter Brownian motion is briefly introduced and the dimension of the three basic sets: the graph, range and zeroset are calculated using the methods derived before. The calculation of the dimension of the zeroset is using the approach of Graf, Mauldin and Williams and is another highlight of the lecture. We finish the lecture with a brief glimpse of Peres' recent idea of intersection equivalence, which establishes a surprising relation of Brownian motion and Galton-Watson fractals.

For most of the lecture I assumed the audience to be familiar with a tiny bit of measure theory (as included in most analysis courses) and some basic probability (as taught to second year students in Kaiserslautern). Only in the last chapter I used some more involved probability theory (mainly the strong Markov property), which my audience was willing to endure.

As usual I remind the reader who did not attend the lecture that the present text cannot replace the lecture, in particular as I did not include any pictures here (and they are a truly important ingredient of the lecture).

Finally, I would like to thank the audience of the lecture for their enthusiasm and patience, and Jochen Blath for the great work he did in the tutorials and many discussions, which featured the lecture.

Kaiserslautern, February 2001

Peter Mörters.

Chapter 1

Self-Similarity

In this chapter we introduce the concept of self-similarity. We define self-similar sets and describe two algorithms to put them on a computer screen. Both algorithms are based on fundamental principles of mathematical analysis: the first (deterministic) algorithm exploits the contraction principle, the second (stochastic) algorithm uses the ergodic theorem. A third algorithm, the *escape time algorithm*, will be discussed in the exercises.

1.1 Self-similar sets: Definition and basic properties

1.1.1 The idea of self-similarity.

The basic idea of *self-similarity* is very simple: A set C is called a *self-similar set* if it is the union of small copies of itself.

Example 1.1

(1) *The* Sierpinski gasket *consists of three copies of itself, each contracted by a factor of* $1/2$.

(2) *The* von Koch curve *consists of four copies of itself, each contracted by a factor of* $1/3$.

To formalize the idea of self-similarity we consider compact subsets of a fixed metric space X with metric d. Recall that (X, d) is a *metric space* if X is a set and $d : X \times X \to [0, \infty)$ a mapping such that

(1) $d(x, y) = d(y, x)$ for all $x, y \in X$,

(2) $d(x, y) = 0$ if and only if $x = y$,

(3) $d(x, y) \leq d(x, z) + d(z, y)$ for all $x, y, z \in X$ (*triangle inequality*).

d defines a *metric* or *distance* between points and hence also, for every nonempty set A, the *diameter* by

$$|A| := \sup_{x,y \in A} d(x, y).$$

We have to make precise what we mean by a *small copy* of a set. We offer two possible solutions here: The first definition gives a quite generous meaning to the notion of a small copy and we use it to define self-similar sets in the wider sense, the second is more restrictive and we reserve the word *self-similar set* for this class.

Definition 1.2 *Suppose (X, d) is a metric space. A mapping*

$$\phi : X \longrightarrow X$$

is called contraction, *if there is a number $0 < c < 1$, called the* contraction factor, *such that*

$$d(\phi(x), \phi(y)) < c \cdot d(x, y) \text{ for all } x, y \in X.$$

A family $(\phi_1, \dots, \phi_n)$ of finitely many contractions is called an iterated function system. *A compact set $C \subset X$ is called* attractor of the iterated function system $(\phi_1, \dots, \phi_n)$ *if*

$$C = \bigcup_{i=1}^{n} \phi_i(C), \tag{1.1}$$

where $\phi_i(C) := \{\phi_i(x) : x \in C\}$ is the image of C under ϕ_i. A set is called self-similar in the wider sense *if it is attractor of an iterated function system.*

A particularly interesting case occurs if $\phi_1, \dots, \phi_n$ are similitudes. A mapping $\phi : X \to X$ is a similitude *if there is a number $0 < r < 1$, called the* contraction ratio, *such that*

$$d(\phi(x), \phi(y)) = r \cdot d(x, y) \text{ for all } x, y \in X.$$

Obviously, similitudes are contractions, but not conversely. A set C is a self-similar set *or sometimes* self-similar in the strict sense *if it is the attractor of an iterated function system $(\phi_1, \dots, \phi_n)$ consisting entirely of similitudes.*

Remark 1.3 *Contractions are continuous. If $c_1, \dots, c_n$ are the contraction coefficients of $\phi_1, \dots, \phi_n$, then*

$$c := \max_{i=1}^{n} c_i < 1$$

is called the contraction coefficient of the iterated function system $(\phi_1, \dots, \phi_n)$.
For every contraction ϕ the set $\phi(A)$ is smaller than A in the sense that its diameter is strictly smaller. More precisely, $|\phi(A)| \leq c|A|$. In the case of a similitude ϕ we even have $|\phi(A)| = r|A|$ for $r = \max_{i=1}^{n} r_i$.

Example 1.4 *Every finite set is self-similar in the wider sense, but the only finite sets which are self-similar in the strict sense are the singletons. A line segment $[0, 1]$ is self-similar with similitudes $\phi_1(x) = x/2$ and $\phi_2(x) = (x + 1)/2$.*

These examples are, of course, not the interesting ones. The first nontrivial example will be presented in the next section.

1.1.2 The middle-third Cantor set.

We now give a first nontrivial example of a self-similar set. We define the *ternary* or *middle-third Cantor set* by

$$C := \Big\{ \sum_{i=1}^{\infty} \frac{x_i}{3^i} \, : \, x_i \in \{0,2\} \Big\}. \tag{1.2}$$

Note that

$$0 \le \sum_{i=1}^{\infty} \frac{x_i}{3^i} \le 2 \sum_{i=1}^{\infty} \frac{1}{3^i} = 2\Big(\frac{1}{1-1/3} - 1\Big) = 1\,,$$

hence $C \subset [0,1]$.

Remark 1.5 *To get an intuitive feeling for the set, recall that every real number $x \in [0,1]$ can be expanded in a* 3-adic expansion, *which is a sequence $(x_k \, : \, k \ge 1)$ with* digits $x_k \in \{0,1,2\}$ *such that*

$$x = \sum_{i=1}^{\infty} \frac{x_i}{3^i}\,.$$

Unfortunately, this expansion is not *unique for all x. Using the formula for the sum of a geometric sequence, one can observe, for example, that for every $j \ge 1$ and $x_j \in \{0,1\}$,*

$$\sum_{i=1}^{j} \frac{x_i}{3^i} + \sum_{i=j+1}^{\infty} \frac{2}{3^i} = \sum_{i=1}^{j} \frac{x_i}{3^i} + \frac{2}{3^{j+1}}\frac{1}{1-1/3} = \sum_{i=1}^{j} \frac{x_i}{3^i} + \frac{1}{3^j}\,. \tag{1.3}$$

These are two different expansions of the same number. The Cantor set C is defined as the set of points $x \in [0,1]$ who have a 3-adic expansion, which does not *use the digit 1.*

A popular definition of the Cantor set is based on successively removing open intervals from $[0,1]$. To understand the connection note that every $x \in [0,1]$ with $x_1 = 0$ is in the interval $[0,1/3]$ and every point in this interval has a 3-adic expansion with $x_1 = 0$. Similarly, every $x \in [0,1]$ with $x_1 = 2$ is in the interval $[2/3,1]$ and every point in this interval has a 3-adic expansion with $x_1 = 2$. Hence, the interval $I := (1/3,2/3)$ is disjoint from the Cantor set C, it consists exactly of the points whose 3-adic expansion must use the digit 1 in the first place. This interval is removed from $[0,1]$ in the first stage of the construction of C.

Similarly, we can see that the set $(1/9,2/9) \cup (4/9,5/9) \cup (7/9,8/9)$ consists of those numbers $x \in [0,1]$ which have the digit $x_2 = 1$ in each of their expansions, so they have to be removed. Going on like this, one can get an intuitive construction of the Cantor set. However, this is unpleasant to write down rigorously, so we prefer the more direct definition (1.2).

Lemma 1.6 *C is compact and nonempty.*

Proof: Obviously, C is nonempty. To show that it is closed, pick a point $x \in [0,1] \setminus C$ in the complement. There exists a sequence $(x_k \, : \, k \ge 1)$ of digits in $\{0,1,2\}$ such that

$$x = \sum_{k=1}^{\infty} \frac{x_k}{3^k}\,,$$

a 3-adic expansion of x. As $x \notin C$ there exists a minimal natural number j with $x_j = 1$. Let

$$a = \sum_{i=1}^{j} \frac{x_i}{3^i} \text{ and } \varepsilon := \sum_{i=j+1}^{\infty} \frac{2}{3^i}.$$

We show (i) $x \in (a, a, +\varepsilon)$ and (ii) $(a, a+\varepsilon) \cap C = \emptyset$.

Clearly, $x \in [a, a+\varepsilon]$. If $x = a$ or $x = a + \varepsilon$ (1.3) offers an alternative representation of x avoiding the digit 1, contradicting $x \in C$. Hence (i) is shown.

We have to show that $(a, a+\varepsilon)$ is disjoint from C. Suppose that $\tilde{x} \in (a, a+\varepsilon)$ has a 3-adic expansion $\tilde{x} = \sum_{k=1}^{\infty} \tilde{x}_k / 3^k$. Suppose $\tilde{x}_i \neq x_i$ for some minimal $1 \leq i \leq j$. If $\tilde{x}_i < x_i$ then $\tilde{x} \leq a$, if $\tilde{x}_i > x_i$ then $\tilde{x} \geq a + \varepsilon$. As this contradicts the assumption, we have $\tilde{x}_i = x_i$ for all $1 \leq i \leq j$. In particular, we must have $\tilde{x}_j = x_j = 1$. This implies that $\tilde{x} \notin C$, showing (ii).

We infer that $[0,1] \setminus C$ is open, so C must be closed. It is also bounded and, by the Theorem of Heine-Borel, compact. ∎

Lemma 1.7 *Define $\phi_1 : [0,1] \to [0,1]$ by $\phi_1(x) = x/3$ and $\phi_2 : [0,1] \to [0,1]$ by $\phi_2(x) = x/3 + 2/3$. Then ϕ_1, ϕ_2 are similitudes on the metric space $[0,1]$ with $d(x,y) = |x - y|$, their contraction ratios are $1/3$.*

Proof: Obviously ϕ_1, ϕ_2 map $[0,1]$ into itself and

$$d\big(\phi_1(x), \phi_1(y)\big) = d\big(\phi_2(x), \phi_2(y)\big) = \left|\frac{x}{3} - \frac{y}{3}\right| = \frac{1}{3}\, d(x,y),$$

proving the statement. ∎

Lemma 1.8 *$C = \phi_1(C) \cup \phi_2(C)$.*

Proof: If $x \in [0,1]$, then $x = \sum_{k=1}^{\infty} x_k/3^k$ for a sequence $(x_k : k \geq 1)$ with $x_k \in \{0,1,2\}$. Then

$$\phi_1(x) = \sum_{k=1}^{\infty} \frac{y_k}{3^k} \qquad \text{with } y_1 = 0, y_k = x_{k-1} \text{ for } k \geq 2,$$

$$\phi_2(x) = \sum_{k=1}^{\infty} \frac{y_k}{3^k} \qquad \text{with } y_1 = 2, y_k = x_{k-1} \text{ for } k \geq 2.$$

Hence, $x \in \phi_1(C)$ if and only if x has a 3-adic expansion with $x_1 = 0$ and $x_k \in \{0,2\}$ for all $k \geq 1$. Similarly, $x \in \phi_2(C)$ if and only if x has a 3-adic expansion with $x_1 = 2$ and $x_k \in \{0,2\}$ for all $k \geq 1$. ∎

Theorem 1.9 *The ternary Cantor set is a self-similar set.*

Proof: Lemma 1.6, Lemma 1.7 and Lemma 1.8 imply that the Cantor set C is attractor of the iterated function system (ϕ_1, ϕ_2). ∎

1.1.3 The coding theorem for self-similar sets

In the previous example we have used the fact that there is a mapping

$$\pi : \{0,2\}^{\mathbb{N}} \longrightarrow C\,, \qquad (x_k) \mapsto \sum_{i=1}^{\infty} \frac{x_i}{3^i}\,.$$

A similarly useful construction exists for all sets associated with an iterated function system. It is called the *code map*. In this section we construct the code map.

Definition 1.10 *Let $\Sigma_n := \{1,\dots,n\}^{\mathbb{N}}$ be the set of all sequences with values in $\{1,\dots,n\}$. For any two distinct sequences (x_k) and (y_k) define*

$$d_n\big((x_k),(y_k)\big) = \frac{1}{2^m} \textit{ for } m = \min\{k \,:\, x_k \neq y_k\}\,.$$

d_n defines a metric on Σ_n, which makes Σ_n compact. Σ_n is called the code space in n symbols.

Remark 1.11 *Compactness of Σ_n follows either from the Theorem of Tychonov stating that products of compact sets are compact, or by using the sequential compactness: from every sequence of elements in Σ_n we can extract a convergent subsequence using the famous* diagonal argument.

Theorem 1.12 (Coding Theorem) *Suppose that $(\phi_1,\dots,\phi_n)$ is an iterated function system and on a metric space X with metric d. Suppose K is the attractor of this iterated function system. Then there exists a continuous and surjective mapping*

$$\pi : \Sigma_n \longrightarrow K \textit{ with } \pi((y_k)) = \lim_{k\uparrow\infty} \phi_{y_1}\circ\cdots\circ\phi_{y_k}(x) \textit{ for all } x \in K.$$

such that, for all $(y_1,\dots,y_k) \in \{1,\dots,n\}^k$,

$$\pi\big(\{(x_i) \,:\, x_1 = y_1,\dots,x_k = y_k\}\big) = \phi_{y_1}\circ\cdots\circ\phi_{y_k}(K).$$

π is called the code map. *If additionally all ϕ_k, $1 \le k \le n$, are injective and*

$$\phi_i(K)\cap\phi_j(K) = \emptyset \textit{ for all } i \neq j,$$

then π is a homeomorphism.

Remark 1.13 *The elements of $\{1,\dots,n\}^{\mathbb{N}}$ can be seen as* addresses *or* address sequences. *π maps an address sequence onto a corresponding point, if π fails to be injective, a point may have several addresses.*

Proof: We proceed in five steps.

Step 1: For every $(x_k) \in \Sigma_n$ there exists $x \in K$ such that

$$\bigcap_{k=1}^{\infty} \phi_{x_1}\circ\cdots\circ\phi_{x_k}(K) = \{x\}.$$

To prove Step 1, look at the sets $K_k := \phi_{x_1} \circ \cdots \circ \phi_{x_k}(K)$. Note that

$$\phi_{x_1} \circ \cdots \circ \phi_{x_k}(K) \subset \phi_{x_1} \circ \cdots \circ \phi_{x_{k-1}}(K),$$

so that $K_k \subset K_{k-1}$ for all $k > 1$. Each K_j is compact, because it is the image under a continuous map of the compact set K, and it is nonempty. Recall that the intersection of a decreasing sequence of nonempty, compact sets is nonempty. Hence, there is a nonempty compact set A with

$$\bigcap_{k=1}^{\infty} \phi_{x_1} \circ \cdots \circ \phi_{x_k}(K) =: A.$$

In order to prove that the set A is a singleton, i.e. contains exactly one point, it suffices to show that A has diameter 0. Recall that every compact set has finite diameter. Also recall the definition of the contraction factor c of the iterated function system. We have,

$$\begin{aligned} \big|\phi_{x_1} \circ \cdots \circ \phi_{x_k}(K)\big| &= \sup\Big\{d\big(\phi_{x_1}(x), \phi_{x_1}(y)\big) \,:\, x, y \in \phi_{x_2} \circ \cdots \circ \phi_{x_k}(K)\Big\} \\ &\leq c \sup\Big\{d(x,y) \,:\, x, y \in \phi_{x_2} \circ \cdots \circ \phi_{x_k}(K)\Big\}. \end{aligned}$$

Inductively, we infer

$$|A| \leq \big|\phi_{x_1} \circ \cdots \circ \phi_{x_k}(K)\big| \leq c^k\,|K| \overset{k\to\infty}{\longrightarrow} 0\,.$$

This proves Step 1.

Step 2: Define the mapping $\pi : \Sigma_n \to K$ such that $\pi((x_k))$ is the only element of the set $\bigcap_{k=1}^{\infty} \phi_{x_1} \circ \cdots \circ \phi_{x_k}(K)$. Then π is surjective.

To prove this, let $x \in K$. We construct a sequence (x_k) inductively as follows. As $K = \bigcup_{i=1}^{n} \phi_i(K)$, there exists x_1 with $x \in \phi_{x_1}(K)$. Having constructed $x_1, \ldots, x_k$ with $x \in \phi_{x_1} \circ \cdots \circ \phi_{x_k}(K)$ we find $\tilde{x} \in K$ such that $x = \phi_{x_1} \circ \cdots \circ \phi_{x_k}(\tilde{x})$. Then choose x_{k+1} such that $\tilde{x} \in \phi_{x_{k+1}}(K)$. By definition, $x \in \bigcap_{k=1}^{\infty} \phi_{x_1} \circ \cdots \circ \phi_{x_k}(K)$ and $\pi((x_k))$ is the only point of this set, hence $x = \pi((x_k))$. This proves surjectivity.

Step 3: For all $(y_1, \ldots, y_k) \in \{1, \ldots, n\}^k$,

$$\pi\big(\{(x_i) \,:\, x_1 = y_1, \ldots, x_k = y_k\}\big) = \phi_{y_1} \circ \cdots \circ \phi_{y_k}(K).$$

This implies in particular that

$$\pi((y_k)) = \lim_{k\uparrow\infty} \phi_{y_1} \circ \cdots \circ \phi_{y_k}(x) \text{ for all } x \in K.$$

If $x \in \phi_{y_1} \circ \cdots \circ \phi_{y_k}(K)$, there exists $\tilde{x} \in K$ with $x = \phi_{y_1} \circ \cdots \circ \phi_{y_k}(\tilde{x})$. By Step 2 we can choose a sequence $(\tilde{x}_k)$ with $\tilde{x} = \pi((\tilde{x}_k))$. Then $\pi((x_1, \ldots, x_k, \tilde{x}_1, \tilde{x}_2, \ldots)) = x$, hence $x \in \pi\big(\{(x_i) \,:\, x_1 = y_1, \ldots, x_k = y_k\}\big)$. This shows the inclusion $\supseteq$. Conversely,

$$\begin{aligned} &\pi\big(\{(x_i) \,:\, x_1 = y_1, \ldots, x_k = y_k\}\big) \\ &\subset \bigcup\Big\{\bigcap_{i=1}^{\infty} \phi_{x_1} \circ \cdots \circ \phi_{x_i}(K) \,:\, (x_k) \text{ with } x_1 = y_1, \ldots, x_k = y_k\Big\} \\ &\subset \phi_{y_1} \circ \cdots \circ \phi_{y_k}(K). \end{aligned}$$

This finishes the proof of Step 3.

Step 4: π is uniformly continuous.

Suppose $\varepsilon > 0$ is given. We fix m large enough that $c^m|K| < \varepsilon$. This is possible because $|K| < \infty$ and $c < 1$. If $(x_k), (y_k) \in \Sigma_n$ with $d_n((x_k), (y_k)) \le 1/2^{m+1}$ we infer that $x_1 = y_1, \ldots, x_m = y_m$ and $\pi(x_k)$ and $\pi(y_k)$ are both in the set $\phi_{x_1} \circ \cdots \circ \phi_{x_m}(K)$. The diameter of this set is at most $c^m|K| \le \varepsilon$. Hence, $d(\pi(x_k), \pi(y_k)) \le \varepsilon$, showing uniform continuity.

Step 5: If all ϕ_k, $1 \le k \le n$, are injective and

$$\phi_i(K) \cap \phi_j(K) = \emptyset \text{ for all } i \neq j,$$

then π is a homeomorphism.

To prove the last step, we let (x_i) and (y_i) be different elements from Σ_n. Then, there exists a minimal $j \ge 1$ with $x_j \neq y_j$ and, by our assumptions, the sets

$$\phi_{x_1} \circ \cdots \circ \phi_{x_j}(K) \text{ and } \phi_{y_1} \circ \cdots \circ \phi_{y_j}(K)$$

are disjoint. As $\pi(x_k)$ is in the first and and $\pi(y_k)$ in the second set, they must be different. This proves that π is injective. Hence an inverse mapping π^{-1} of the code map exists. It is automatically continuous. Indeed, if $A \subset \Sigma_n$ is closed, it is compact (because Σ_n is compact) and its inverse image under π^{-1} is $\pi(A)$, which is compact, hence closed. ∎

1.2 Construction of self-similar sets via the contraction principle

The *key questions* about iterated function systems are the following:

- Given an iterated function system $(\phi_1, \ldots, \phi_n)$, is there an attractor?
- Given an iterated function system $(\phi_1, \ldots, \phi_n)$, is the attractor uniquely determined?

It will turn out that on $\mathbb{R}^d$ both questions have an affirmative answer. This allows us later to define a self-similar set in the wider sense by just giving an iterated functions system, or a self-similar set in the strong sense by giving an iterated functions system consisting of similitudes.

1.2.1 The contraction principle

We formulate a condition on a metric space $(M, \mathbf{d})$, which is sufficient for the contraction principle.

Definition 1.14 *A sequence $(X_k : k \ge 1)$ in a metric space $(M, \mathbf{d})$ is called a* Cauchy sequence *if*

$$\lim_{n\to\infty} \sup_{i,j \ge n} d(X_i, X_j) = 0 .$$

A metric space $(M, \mathbf{d})$ is called complete *if every Cauchy sequence in $(M, \mathbf{d})$ is convergent to some $X \in M$.*

Example 1.15 *Closed subset of $\mathbb{R}^d$ are complete metric spaces.*

Recall that $X \in M$ is called a fixed point of $\Phi : M \to M$ if $\Phi(X) = X$. The following theorem, called the *contraction principle* or *Banach fixed point theorem* establishes existence and uniqueness of fixed points for contractions on complete metric spaces.

Theorem 1.16 (Contraction principle) *If $(M, \mathbf{d})$ is a complete metric space and $\Phi : M \to M$ a contraction. Then Φ has exactly one* fixed point *$X \in M$ and for every $X_0 \in M$ the sequence $(X_k : k \geq 1)$ defined by $X_k = \Phi(X_{k-1})$ converges to the fixed point X.*

Proof: For every $X_0 \in M$, we have, using the triangle inequality and geometric series,

$$\begin{aligned}\mathbf{d}\big(X_0, \Phi^k(X_0)\big) &\leq \mathbf{d}\big(X_0, \Phi(X_0)\big) + \mathbf{d}\big(\Phi(X_0), \Phi^2(X_0)\big) \cdots + \mathbf{d}\big(\Phi^{k-1}(X_0), \Phi^k(X_0)\big) \\ &\leq \big(1 + c + c^2 + \cdots + c^{k-1}\big)\, \mathbf{d}\big(X_0, \Phi(X_0)\big) \\ &\leq \frac{\mathbf{d}\big(X_0, \Phi(X_0)\big)}{1-c}.\end{aligned}$$

Hence,

$$\lim_{k\to\infty} \sup_{n,m\geq k} \mathbf{d}\big(\Phi^n(X_0), \Phi^m(X_0)\big) \leq \lim_{k\to\infty} \frac{c^k}{1-c}\mathbf{d}\big(X_0, \Phi(X_0)\big) = 0\,,$$

hence $(\Phi^k(X_0))$ is a Cauchy sequence and we let X be its limit. Letting $k \to \infty$ in the equation

$$\Phi^{k+1}(X_0) = \Phi(\Phi^k(X_0))$$

shows that X is a fixed point. Finally, if X and Y are two fixed points, we have

$$\mathbf{d}\big(X, Y\big) = \mathbf{d}\big(\Phi(X), \Phi(Y)\big) \leq c\mathbf{d}\big(X, Y\big)\,,$$

which implies $\mathbf{d}(X, Y) = 0$, hence $X = Y$. ■

1.2.2 Existence of self-similar sets

We now use the contraction principle to answer the question about existence and uniqueness of attractors of a given iterated function system on metric spaces given as closed subsets of $\mathbb{R}^d$ with the Euclidean distance $d(x, y) = |x - y|$.

Theorem 1.17 *Let $(\phi_1, \ldots, \phi_n)$ be an iterated function system on a nonempty closed subset $X \subset \mathbb{R}^d$. Then there is exactly one nonempty compact set $K \subset X$, which is the attractor of this system, i.e. such that $K = \bigcup_{i=1}^n \phi_i(K)$.*

To prove Theorem 1.17 fix the metric space (X, d) as above and an iterated function system on this space. The **idea of the proof** is to work on the set

$$\mathfrak{K} := \{A \subset X : \text{ nonempty and compact}\}$$

and interpret the attractor K of the iterated function system as solutions of the fixed point equation $\Phi(K) = K$ for the function

$$\Phi(K) := \bigcup_{i=1}^{n} \phi_i(K) \text{ for all } K \in \mathfrak{K}.$$

Existence and uniqueness of the fixed point follows from the contraction principle *if* we can find a metric $\mathfrak{d}$ on $\mathfrak{K}$ such that

(i) $(\mathfrak{K}, \mathfrak{d})$ is a *complete* metric space,

(ii) Φ is a contraction on $(\mathfrak{K}, \mathfrak{d})$.

We now define the *Hausdorff metric* on the set $\mathfrak{K}$ and verify (i) and (ii) above for this metric. Then Theorem 1.17 follows directly from the contraction principle.

Definition 1.18 *If $K \in \mathfrak{K}$ and $\delta > 0$ then*

$$K[\delta] := \big\{x \in X \ : \ \textit{there is } y \in K \textit{ with } d(x,y) \le \delta\big\}$$

*in called the δ-*parallel body *of K. For two sets $K, L \in \mathfrak{K}$ we define the* Hausdorff distance *by*

$$\mathfrak{d}(K,L) := \inf\big\{\delta > 0 \ : \ K \subset L[\delta] \textit{ and } L \subset K[\delta]\big\}.$$

Lemma 1.19 *$\mathfrak{d}$ defines a metric on the set $\mathfrak{K}$ called the Hausdorff metric. The metric space $(\mathfrak{K}, \mathfrak{d})$ is complete.*

Proof: Obviously, $\mathfrak{d}$ is nonnegative and symmetric. Suppose that $\mathfrak{d}(K,L) = 0$, then $L \subset K[\delta]$ for all $\delta > 0$. As K is closed we have $\bigcap_{\delta>0} K[\delta] = K$. Hence $L \subset K$ and, by the symmetric argument, $L \supset K$. We infer that $\mathfrak{d}(K,L) = 0$ implies $K = L$. The triangle inequality follows from the fact that $K \subset L[\delta]$ and $L \subset M[\varepsilon]$ implies $K \subset M[\delta + \varepsilon]$.

The crucial part of the proof is the completeness. Let $(K_n : n \ge 1)$ be a Cauchy sequence in $\mathfrak{K}$. To prove completeness, we may, by passing to a subsequence, if necessary, assume that $\mathfrak{d}(K_n, K_m) < 1/2^n$ for all $m \ge n$. Define

$$K := \bigcap_{j=1}^{\infty} \mathrm{cl}\Big(\bigcup_{i=j}^{\infty} K_i\Big),$$

where cl stands for the closure of a set. Then

$$\emptyset \ne \mathrm{cl}\Big(\bigcup_{i=j}^{\infty} K_i\Big) \subset K_j[1/2^j],$$

hence $\mathrm{cl}\Big(\bigcup_{i=j}^{\infty} K_i\Big)$ is compact, because it is closed and bounded. Recall that the intersection of a decreasing sequence of compacts is nonempty. This implies that

$$\emptyset \ne \bigcap_{j=1}^{\infty} \mathrm{cl}\Big(\bigcup_{i=j}^{\infty} K_i\Big) = K \in \mathfrak{K}.$$

We know already that $K \subset K_j[1/2^j]$ for all j. We intend to show that also $K_j \subset K[1/2^j]$ for all j. Indeed, if $x \in K_j$, there exist, for all $k \geq j$, an $x_k \in K_k$ with $|x - x_k| \leq 1/2^j$. Because (x_k) is bounded, we may pick a subsequence of (x_k) which converges to a limit called y. Note that

$$y \in \mathrm{cl}\Big(\bigcup_{i=k}^{\infty} K_i\Big) \text{ for all } k.$$

Hence, $y \in K$ and $|x - y| \leq 1/2^j$. This implies that $K_j \subset K[1/2^j]$ for all j.

We have shown that $\mathfrak{d}(K, K_j) \leq 1/2^j \longrightarrow 0$, hence K is the limit of our Cauchy sequence and this proves completeness. ■

Lemma 1.20 *The mapping Φ defined by*

$$\Phi(K) = \bigcup_{i=1}^{n} \phi_i(K), \textit{ for all } K \in \mathfrak{K},$$

is a contraction with contraction factor $c = \max_{i=1}^{n} c_i$.

Proof: If $A, B \in \mathfrak{K}$, we have

$$\mathfrak{d}\big(\Phi(A), \Phi(B)\big) = \mathfrak{d}\Big(\bigcup_{i=1}^{n} \phi_i(A), \bigcup_{i=1}^{n} \phi_i(B)\Big) \leq \max_{i=1}^{n} \mathfrak{d}\big(\phi_i(A), \phi_i(B)\big)\,.$$

Note that

$$\mathfrak{d}\big(\phi_i(A), \phi_i(B)\big) \leq c_i \mathfrak{d}\big(A, B\big)\,,$$

and this completes the proof. ■

Theorem 1.17 follows now by applying Theorem 1.16 to the mapping Φ on $(\mathfrak{K}, \mathfrak{d})$ and recalling that the fixed points of Φ are exactly the attractors of the given iterated function system.

The previous proof is not only an abstract existence proof, but also gives a concrete way of constructing the attractor K of a given iterated function system on a computer screen. Recall from Theorem 1.16 that we can get the fixed point $K \in \mathfrak{K}$ of the mapping ϕ as the limit of a sequence obtained by iterated application of Φ with an arbitrary starting set. Hence we have verified the following *deterministic algorithm* for the construction of the attractor of a given iterated function system:

Algorithm 1:

Step 1: Let $k = 0$ and choose an arbitrary nonempty, compact subset X_0 of X.

Step 2: Let $X_{k+1} = \Phi(X_k)$ and increase k by one.

Step 3: If k is below a given threshold, go back to Step 2, otherwise stop.

The *threshold* in the algorithm depends on the contraction rate c and the distance of the starting set X_0 and the fixed point K, the former is known from the iterated function system and the latter can be estimated, for example by the diameter of X. We have

$$\mathfrak{d}(X_k, K) \le c\mathfrak{d}(X_{k-1}, K) \le c^k \mathfrak{d}(X_0, K) \le c^k |X| \le \varepsilon\,,$$

if $k > \log(\varepsilon/|X|)/\log c$.

The main disadvantage of this algorithm is that it is expensive to apply the mapping Φ: in every step of the algorithm each point in the previous picture is mapped onto n new points. We shall discuss a more effective algorithm in Section 1.3.

1.2.3 Existence of self-similar measures

In this section we construct nice probability measures on the attractors of iterated function systems. Having such a measure allows us to pick a point at random from the set, taking the probability measure as the distribution of the random point.

Recall that a vector $(p_1, \dots, p_n)$ is called a *probability vector* if it has nonnegative entries summing up to 1. Such a vector defines a probability distribution P on the set $\{1, \dots, n\}$ by $P\{i\} = p_i$.

Theorem 1.21 (Existence of self-similar measures) *Suppose $\phi = (\phi_1, \dots, \phi_n)$ is an arbitrary iterated function system on a closed subset $X \subset \mathbb{R}^d$ and $K \subset X$ its attractor. Let $p = (p_1, \dots, p_n)$ be a probability vector. Then there exists exactly one probability measure μ on K such that*

$$\mu = \sum_{k=1}^n p_k\, \mu \circ \phi_k^{-1}\,.$$

Such a measure is called the self-similar measure *associated to p and ϕ. We have,*

$$\mu\Big(\phi_{y_1} \circ \dots \circ \phi_{y_k}(K)\Big) \ge \prod_{i=1}^k p_{y_i}\,, \tag{1.4}$$

and if $\phi_i(K) \cap \phi_j(K) = \emptyset$ for all $i \neq j$ and all ϕ_i are injective we have equality. If all $p_i > 0$, then the support of μ is K.

Proof: We use again the contraction principle. Let $\mathfrak{M}$ be the set of all probability measures on K and define

$$\Phi : \mathfrak{M} \to \mathfrak{M}\,, \quad \Phi(\mu) = \sum_{k=1}^n p_k\, \mu \circ \phi_k^{-1}\,.$$

The self-similar measures μ are the fixed points of Φ. It suffices to choose a metric on $\mathfrak{M}$, which makes the space complete and Φ a contraction. Let $\mathcal{L}$ be the class of functions $g : K \to [0, \infty)$ with $|g(x) - g(y)| \le |x - y|$ for all x, y. For $\nu_1, \nu_2 \in \mathfrak{M}$ let

$$\mathbf{d}(\nu_1, \nu_2) = \sup\Big\{\Big|\int g\, d\nu_1 - \int g\, d\nu_2\Big| \,:\, g \in \mathcal{L}\Big\}.$$

Then $\mathbf{d}$ is a metric, which makes $\mathfrak{M}$ complete (see exercise). We show that Φ is a contraction. The first question we have to address is, how the integrals with respect to the measure $\Phi(\nu)$ look like. Indeed,

$$\int g\, d\Phi(\nu) = \sum_{k=1}^{n} p_k \int g\, d\nu \circ \phi_k^{-1} = \sum_{k=1}^{n} p_k \int g \circ \phi_k \, d\nu\,.$$

Now we can check that Φ is a contraction.

$$\begin{aligned}
\mathbf{d}\big(\Phi(\nu_1), \Phi(\nu_2)\big) &= \sup\Big\{\Big| \int g\, d\Phi(\nu_1) - \int g\, d\Phi(\nu_2)\Big| \,:\, g \in \mathcal{L}\Big\} \\
&= \sup\Big\{\Big| \sum_{k=1}^{n} p_k \Big(\int g \circ \phi_k\, d\nu_1 - \int g \circ \phi_k\, d\nu_2\Big)\Big| \,:\, g \in \mathcal{L}\Big\} \\
&\le \sum_{k=1}^{n} p_k \sup\Big\{\Big| \int g \circ \phi_k\, d\nu_1 - \int g \circ \phi_k\, d\nu_2\Big| \,:\, g \in \mathcal{L}\Big\}.
\end{aligned}$$

Now note that

$$|c_k^{-1} g \circ \phi_k(x) - c_k^{-1} g \circ \phi_k(y)| \le c_k^{-1} |\phi_k(x) - \phi_k(y)| \le |x - y|\,,$$

hence $c_k^{-1} g \circ \phi_k \in \mathcal{L}$. Then the previous chain of inequalities can be continued with

$$\begin{aligned}
&\le \sum_{k=1}^{n} p_k c_k \sup\Big\{\Big| \int c_k^{-1} g \circ \phi_k\, d\nu_1 - \int c_k^{-1} g \circ \phi_k\, d\nu_2\Big| \,:\, g \in \mathcal{L}\Big\} \\
&\le \sum_{k=1}^{n} p_k c_k \sup\Big\{\Big| \int g\, d\nu_1 - \int g\, d\nu_2\Big| \,:\, g \in \mathcal{L}\Big\} \\
&\le c\, \mathbf{d}(\nu_1, \nu_2)\,,
\end{aligned}$$

for $c = \max_{k=1}^{n} c_k$. This proves existence and uniqueness of the self-similar measure. To see the additional statement we use the defining property, and argue inductively,

$$\begin{aligned}
\mu\Big(\phi_{y_1} \circ \cdots \circ \phi_{y_k}(K)\Big) &= \sum_{i=1}^{n} p_i \mu \circ \phi_i^{-1} \Big(\phi_{y_1} \circ \cdots \circ \phi_{y_k}(K)\Big) \\
&\ge p_{y_1} \mu\Big(\phi_{y_2} \circ \cdots \circ \phi_{y_k}(K)\Big) \\
&\ge \prod_{i=1}^{k} p_{y_i}\,.
\end{aligned}$$

If $\phi_i(K) \cap \phi_j(K) = \emptyset$ for all $i \ne j$, and all ϕ_i are injective, the parts are disjoint and the measure must add to one. Hence we must have equality. Finally, if $p_i > 0$ for all i, note that every open set $U \subset X$ with $K \cap U \ne \emptyset$ contains a set $\phi_{y_1} \circ \cdots \circ \phi_{y_k}(K)$ for sufficiently large k and hence $\mu(U) > 0$. We have shown that every open set intersecting K has positive measure, which means that K is the support of μ. ■

Remark 1.22 (Random addresses) *Recall that we have the coding map*

$$\pi : \Sigma_n \longrightarrow K\,.$$

Suppose that $X_1, X_2, X_3, \ldots$ is a sequence of independent identically distributed random variables each with law P. A natural way to choose a point from K at random would be to choose an address from the address space Σ_n at random and look at the point with this address. Formally speaking, this means that we take the product probability measure of P on Σ_n and look at its image measure under the mapping π.

It is easy to check that this measure is the self-similar measure of the previous theorem. In the construction we have assumed that, for any given probability measure P on Σ_n, there is a sequence of independent identically distributed random variables with law P. This is usually proved in lectures on probability theory, but is by no means obvious. One way to prove it is analogous to the proof of Theorem 1.21.

1.3 Construction of self-similar sets via the chaos game

1.3.1 The ergodic theorem

The ergodic theorem is one of the most fundamental and useful results in analysis and a link between probability theory and dynamical systems.

The basic heuristic of the ergodic theorem can be stated, very roughly, like this:

> *For a chaotically moving particle at late times the average of a function along the trajectory of the particle equals the overall spatial average.*

Very roughly, our aim will be to construct a randomly moving particle on the metric space X such that, for every set A the time spent by the particle in the set is proportional to $\mu(A)$ for a self-similar measure μ. Then the trajectory of the particle looks like the support of the self-similar measure. From Theorem 1.21 we know that the support of μ is K if all $p_i > 0$.

To turn this loose heuristic idea into a mathematical theorem we fix a suitable framework:

Suppose $(\Omega, \mathfrak{A}, \mathbf{P})$ is a probability space and $f : \Omega \to \Omega$ a *measurable invariant mapping*, i.e.

$$\mathbf{P} \circ f^{-1} = \mathbf{P}.$$

f determines the movement of a particle started in $x_0 \in \Omega$ by

$$x_{k+1} = f(x_k) \text{ for all } k \geq 0.$$

f is called *ergodic* if

$$\mathbf{P}(A) = 0 \text{ or } \mathbf{P}(A) = 1 \text{ for all } A \in \mathfrak{A} \text{ with } A = f^{-1}(A).$$

This condition ensures that the particle does not move on a set significantly smaller than Ω.

Let $\varphi : \Omega \to [0, \infty)$ be a bounded, measurable function. The average of φ along the trajectory $x_1, \ldots, x_n$ of the particle at time n is $(1/n) \sum_{k=0}^{n-1} \varphi(x_k)$, the overall spatial average is $\int \varphi \, d\mathbf{P}$.

The ergodic theorem of G. Birkhoff (1931) states that for ergodic f the averages on the trajectory converge, for almost every starting point, to the space average. We shall give a proof of this fact, which is essentially due to Katznelson and Weiss (1982).

Theorem 1.23 (Ergodic Theorem) *Let* $\mathbf{P}$ *be a probability measure on* $(\Omega, \mathfrak{A})$ *and* $f : \Omega \to \Omega$ *an invariant mapping. Suppose that* $\varphi : \Omega \to [0, \infty)$ *is a bounded, measurable function. Then the limit*

$$\psi(x) := \lim_{k\to\infty} \frac{1}{k} \sum_{j=0}^{k-1} \varphi\big(f^j(x)\big)$$

exists and satisfies $\psi \circ f(x) = \psi(x)$ *for* $\mathbf{P}$*-almost all* x*, and*

$$\int \varphi \, d\mathbf{P} = \int \psi \, d\mathbf{P} \,.$$

Moreover, if f *is ergodic, then for* $\mathbb{P}$*-almost all* x*,* ψ *is constant with* $\psi(x) = \int \varphi \, d\mathbf{P}$*.*

Proof: Let M be the upper bound of φ. Write

$$\alpha_k(x) := \frac{1}{k} \sum_{j=0}^{k-1} \varphi(f^j(x)) \,,$$

the average along the trajectory at time $k-1$. Let

$$\overline{\alpha}(x) := \limsup_{k\to\infty} \alpha_k(x) \,,$$

and recall that this is a measurable function. We also have

$$\overline{\alpha}\big(f(x)\big) = \limsup_{k\to\infty} \frac{1}{k} \sum_{j=1}^{k} \varphi\big(f^j(x)\big) = \overline{\alpha}(x) \,,$$

and, by induction, $\overline{\alpha}(f^k(x)) = \overline{\alpha}(x)$ for all k. We want to show that, for all $\varepsilon > 0$,

$$\int \overline{\alpha}(x) \, d\mathbf{P}(x) \leq \int \varphi(x) \, d\mathbf{P}(x) + \varepsilon. \tag{1.5}$$

To verify (1.5) define

$$\tau(x) = \min\big\{k > 0 \,:\, \alpha_k(x) \geq \overline{\alpha}(x) - \varepsilon\big\} \,,$$

which is finite for all x by definition. Assume first that the measurable function τ is bounded above by some value T. Now the sum α_k can be broken into blocks of length at most T such that the average of the $\varphi(f^j(x))$ over the j in each block is at least $\overline{\alpha}(x) - \varepsilon$. More precisely, for each x we define a sequence $k_1, k_2, \ldots$ inductively, by taking $k_1 = \tau(x)$ and $k_i = \tau(f^{k_1+\cdots+k_{i-1}}(x))$ for $i \geq 2$. Then,

$$\begin{aligned}
\sum_{j=k_1+\cdots+k_{i-1}}^{k_1+\cdots+k_i-1} \varphi(f^j(x)) &= k_i \alpha_{k_i}\left(f^{k_1+\cdots+k_{i-1}}(x)\right) \\
&\geq k_i \left(\overline{\alpha}(f^{k_1+\cdots+k_{i-1}}(x)) - \varepsilon\right) \\
&= k_i \left(\overline{\alpha}(x) - \varepsilon\right) ,
\end{aligned}$$

since $\overline{\alpha}(x) = \overline{\alpha}(f^k(x))$ for all k. Summing over all blocks,

$$\sum_{j=0}^{k-1} \varphi(f^j(x)) \geq k\big(\overline{\alpha}(x) - \varepsilon\big) ,$$

whenever k is of the form $k_1 + \cdots + k_i$. For an arbitrary integer k we let l be the largest integer that $k_1 + \cdots + k_l \leq k$. We get

$$\sum_{j=0}^{k-1} \varphi(f^j(x)) \geq (k - T)\big(\overline{\alpha}(x) - \varepsilon\big) , \tag{1.6}$$

using that $0 \leq k - (k_1 + \cdots + k_l) \leq T$. Because f is invariant, we have

$$\int \varphi \circ f^j \, d\mathbf{P} = \int \varphi \, d\mathbf{P} \circ (f^j)^{-1} = \int \varphi \, d\mathbf{P} .$$

Now we integrate (1.6) and obtain

$$\int \varphi \, d\mathbf{P} = \frac{1}{k} \sum_{j=0}^{k-1} \int \varphi(f^j(x)) \, d\mathbf{P}(x) \geq \frac{k - T}{k} \Big(\int \overline{\alpha} \, d\mathbf{P} - \varepsilon \Big) .$$

Letting $k \uparrow \infty$ we obtain (1.5).

Now we suppose that τ is unbounded. As $\tau(x) < \infty$ for all x, we may choose T so large that $\mathbf{P}(A) < \varepsilon$ for

$$A := \big\{ x \,:\, \tau(x) > T \big\} .$$

The idea is to change the definition of φ on the set A, so that we can apply the previous part to the modified function. At the same time we have to make sure that thanks to the smallness of A we have changed φ so little, that the resulting inequality for the modified function still allows to infer sensible results for the unmodified function.

More precisely, define $\varphi^* : \Omega \to [0, \infty)$ by

$$\varphi^*(x) := \begin{cases} \varphi(x) & \text{if } x \notin A , \\ M & \text{if } x \in A . \end{cases}$$

We define

$$\alpha_k^*(x) := \frac{1}{k} \sum_{j=0}^{k-1} \varphi^*(f^j(x))$$

and

$$\tau^*(x) = \min \big\{ k > 0 \,:\, \alpha_k^*(x) \geq \overline{\alpha}(x) - \varepsilon \big\} .$$

Now we have $\tau^*(x) = 1$ for all $x \in A$ and, as $\alpha_k^* \geq \alpha_k$, for all $x \notin A$ we note that $\tau^*(x) \leq T$. The argument in the bounded case gives the first inequality in the following chain, the others are just definitions

$$\begin{aligned} \int \overline{\alpha}(x) \, d\mathbf{P}(x) &\leq \int \varphi^*(x) \, d\mathbf{P}(x) + \varepsilon \\ &= \int \varphi(x) \, d\mathbf{P}(x) + \int \big(\varphi^*(x) - \varphi(x)\big) \, d\mathbf{P}(x) + \varepsilon \\ &\leq \int \varphi(x) \, d\mathbf{P}(x) + M\varepsilon + \varepsilon . \end{aligned}$$

As $M\varepsilon + \varepsilon$ can be made arbitrarily small, this proves (1.5) in the general case.

As (1.5) holds for all $\varepsilon > 0$, we have

$$\int \overline{\alpha}(x)\, d\mathbf{P}(x) \leq \int \varphi(x)\, d\mathbf{P}(x).$$

Analogously we can define

$$\underline{\alpha}(x) := \liminf_{k\to\infty} \alpha_k(x)\,,$$

and obtain $\int \underline{\alpha}(x)\, d\mathbf{P}(x) \geq \int \varphi(x)\, d\mathbf{P}(x)$. Altogether,

$$\int \big(\underline{\alpha}(x) - \overline{\alpha}(x)\big)\, d\mathbf{P}(x) \geq 0\,,$$

As the integrand is nonpositive, we must have $\underline{\alpha}(x) = \overline{\alpha}(x)$ for $\mathbf{P}$-almost every x. This means, that the limit $\psi(x) = \underline{\alpha}(x) = \overline{\alpha}(x)$ exists and $\psi(f(x)) = \overline{\alpha}(f(x)) = \overline{\alpha}(x) = \psi(x)$, for $\mathbf{P}$-almost every x. By construction, $\int \psi\, d\mathbf{P} = \int \varphi\, d\mathbf{P}$.

Now look at the ergodic case. For every $a \geq 0$ define a set

$$L(a) := \{x\, : \psi(x) \leq a\} \subset \Omega\,.$$

If $f(x) \in L(a)$, then $\psi(x) = \psi(f(x)) \leq a$, hence $x \in L(a)$. In other words, $f^{-1}(L(a)) = L(a)$ and we infer that every set $L(a)$ has measure 0 or 1. Let

$$c := \sup\{a \in [0, M]\, :\, \mathbf{P}(L(a)) = 0\,\}.$$

Then $\psi(x) = c$, $\mathbf{P}$-almost surely. Finally,

$$c = \int c\, d\mathbf{P} = \int \psi\, d\mathbf{P} = \int \varphi\, d\mathbf{P}\,.$$

■

Remark: If you want to know more about ergodic theorems, you are referred to R.M Dudley's book "Real Analysis and Probability", which contains a range of more general ergodic theorems together with applications in probability.

1.3.2 The chaos game algorithm

In this section we shall present a simple *stochastic algorithm* that allows us to draw self-similar sets quickly and efficiently on a computer. The algorithm is based on the ergodic theorem and although verification of the algorithm is nontrivial, it is very easy to run the algorithm.

The algorithm is usually called the *chaos game*. Let $(\phi_1, \ldots, \phi_n)$ be an iterated function system and K the associated set, which we want to draw. To run the algorithm we also choose a vector $(p_1, \ldots, p_n)$ of strictly positive numbers adding to one. Here is how the chaos game is played:

Algorithm 2:

Step 1: Pick a point $y_0 \in X$ and let $k = 1$.

Step 2: Choose randomly an $X_k \in \{1, \ldots, n\}$.
The choice must be made independently from previously chosen random variables such that $\mathbf{P}\{X_k = j\} = p_j$.

Step 3: Let $y_k = \phi_{X_k}(y_{k-1})$ and increase k by one.

Step 4: If k is below a given threshold, go back to Step 2, otherwise stop.

We give some **practical hints for running the algorithm:**

(1) The accuracy of the picture can be regulated by the threshold parameter, which gives the number of points. The algorithm is fast, for good results the number of points should be chosen large.

(2) In practice, it is useful not to display, roughly speaking, the first $\sqrt{k}$ out of k points, which might be too far away from the set K if y_0 is chosen badly.

(3) In many cases, the best results are obtained if p_i is chosen such that the probability p_i of a mapping corresponds to the optimal contraction factor c_i, i.e. such that $\log p_i$ is proportional to $\log c_i$.

The algorithm produces a random set $\{y_1, \ldots, y_k\}$. The following theorem shows in which sense this set is close to the self-similar set K or, more precisely, the self-similar measure μ.

Theorem 1.24 (Chaos game) *Suppose $X \subset \mathbb{R}^d$ is a compact metric space. For every continuous $g : X \to [0, \infty)$ we have, almost surely, that*

$$\lim_{k\to\infty} \frac{1}{k} \sum_{i=0}^{k-1} g(y_i) = \int g \, d\mu .$$

Remark 1.25 *For every set A with $\mu(\partial A) = 0$ the previous result implies that, for large k, the proportion of points from $\{y_1, \ldots, y_k\}$ in A is*

$$\lim_{k\to\infty} \frac{1}{k} \sum_{i=0}^{k-1} 1_A(y_i) = \mu(A) .$$

This equivalence can be shown by approximation of 1_A with continuous functions and is sometimes called the Portmanteau Theorem.

In the **first step** of proof we show that the result of the algorithm does not depend on the starting point. Indeed, if we choose two starting points y_0 and $\tilde{y}_0$ and (y_i) and $(\tilde{y}_i)$ are the resulting sequences with the same random choice of mappings, then

$$d\big(y_k, \tilde{y}_k\big) = d\big(\phi_{X_k} \circ \cdot \circ \phi_{X_1}(y_0), \phi_{X_k} \circ \cdot \circ \phi_{X_1}(\tilde{y}_0)\big) \le c^k d(y_0, \tilde{y}_0) \to 0 .$$

This implies that, for every $\varepsilon > 0$, there is N with $|g(y_k) - g(\tilde{y}_k)| < \varepsilon$ for all $k \geq N$. We infer that

$$\begin{aligned} & \limsup_{k\to\infty} \Big| \frac{1}{k} \sum_{i=0}^{k-1} g(y_i) - \frac{1}{k} \sum_{i=0}^{k-1} g(\tilde{y}_i) \Big| \\ \leq \; & \limsup_{k\to\infty} \frac{1}{k} \sum_{i=0}^{N} |g(y_i) - g(\tilde{y}_i)| + \limsup_{k\to\infty} \frac{1}{k} \sum_{i=N}^{k-1} |g(y_i) - g(\tilde{y}_i)| \\ \leq \; & \varepsilon . \end{aligned}$$

As this holds for all $\varepsilon > 0$, the limit must be zero and the statement of the first step is shown.

In the **second step** we may now restrict our attention to starting points in the set K. It is important to notice that this is just a step in the proof, the algorithm works for all starting points by the argument of Step 1.

Of course, our framework of the ergodic theorem does not seem to involve *random* movements of particles. We have to be very clever in the choice of the space Ω to which we apply the ergodic theorem in order to incorporate this.

The idea is that a point in the space Ω corresponds not only to a point in X, but also to all the list of all sample contractions we shall apply in the future. The mapping f, which determines the movement has to do two things: The first contraction on the list is applied to our point in X *and* the contraction is removed from the list.

A list of contractions from the iterated function system can be represented as an element of the code space Σ_n, more precisely $(x_k) \in \Sigma_n$ represents the list

$$\phi_{x_0}, \phi_{x_1}, \phi_{x_2}, \ldots .$$

By the coding theorem, a point in the attractor K of the iterated function system can also be represented by a sequence in the address space Σ_n. Hence a good choice for the space Ω is the *two-sided code-space* defined by

$$\Sigma_n^* := \{1, \ldots, n\}^{\mathbb{Z}} .$$

Now $(x_k) \in \Sigma_n^*$ represents the point

$$\pi(x_{-1}, x_{-2}, x_{-3}, \ldots)$$

together with the list

$$\phi_{x_0}, \phi_{x_1}, \phi_{x_2}, \ldots$$

of contractions.

The mapping $f : \Sigma_n^* \to \Sigma_n^*$ applies the first contraction on the list, this is ϕ_{x_0}, to the point with address $(x_{-1}, x_{-2}, x_{-3}, \ldots)$. The result, by definition of the code map π, is the point with address $(x_0, x_{-1}, x_{-2}, x_{-3}, \ldots)$. Then ϕ_{x_0} is removed from the list of contractions, in other words we replace $(x_0, x_1, x_2, \ldots)$ by $(x_1, x_2, x_3 \ldots)$. Altogether, applying f to an element $(x_k) \in \Sigma_n$ shifts the sequence (x_k) by one element to the left,

$$f((x_k)) = (y_k) \text{ with } y_k = x_{k+1} .$$

We now observe that for the function

$$\varphi : \Sigma_n^* \to [0,\infty)\,, \varphi\big((x_k)\big) = g\big(\pi(x_{-1}, x_{-2}, x_{-3}, \dots)\big)$$

we have, for $y_0 = \pi(x_{-1}, x_{-2}, \dots)$,

$$\varphi\big(f((x_k))\big) = g\big(\pi(x_0, x_{-1}, x_{-2}, \dots)\big) = g\big(\phi_{x_0}\pi(x_{-1}, x_{-2}, \dots)\big) = g\big(\phi_{x_0}(y_0)\big) = g(y_1)$$

and, by induction,

$$\varphi\big(f^j((x_k))\big) = g\big(\phi_{x_{j-1}} \circ \dots \circ \phi_{x_0}(y_0)\big) = g(y_j)\,.$$

By now we clarified the framework except for the choice of the probability measure P on Σ_n^*. This probability measure P must be defined in such a way that the projections

$$X_j((x_k)) = x_j \text{ for } j = 0, 1, 2, \dots\,,$$

define an independent, identically sequence of P-distributed random variables.

The σ-field $\mathfrak{A}$ on Σ_n^* is generated by the *cylinder sets*

$$\big\{(x_k) \,:\, x_i = y_i \text{ for all } i \in I\big\}, \qquad \text{for } I \subset \mathbb{Z} \text{ finite and } y_i \in \{1, \dots, n\}.$$

We let P be the product measure on Σ_n^*, which satisfies

$$\mathsf{P}\big\{(x_k) \,:\, x_i = y_i \text{ for all } i \in I\big\} = \prod_{i \in I} p_{y_i}\,.$$

Such a measure exists and under this measures the random variables X_i are independent and identically P-distributed.

We have to show that f is invariant and ergodic with respect to P. This requires two useful facts from measure theory, which we formulate as two lemmas.

Lemma A: *Two probability measures on Σ_n^* which agree on all cylinder sets are equal.*

Lemma B: *For every set $A \in \mathfrak{A}$ and $\varepsilon > 0$ there exist pairwise disjoint cylinder sets $C_1, \dots, C_m$ with*

$$\mathsf{P}\Big(A \Delta \bigcup_{i=1}^m C_i\Big) < \varepsilon\,.$$

Lemma A follows from the fact that the cylinder sets are a $\cap$-stable generator of $\mathfrak{A}$ and probability measures, which agree on a $\cap$-stable generator are equal. This result is frequently proved in the lectures on elementary probability theory and a proof can be found in Durrett's book *Probability* or Protter's book *Probability Essentials.*

Lemma B can be proved as follows: Let $\mathfrak{C}$ be the collection of all sets $A \in \mathfrak{A}$, for which the statement is true. This collection is a σ-algebra, this fact can be checked easily. As it obviously contains the cylinder sets, which generate $\mathfrak{A}$, it must coincide with $\mathfrak{A}$.

Step 3. f is invariant and ergodic with respect to P.

To check *invariance* it suffices to show that $\mathsf{P} \circ f^{-1}(E) = \mathsf{P}(E)$ for all cylinder sets E, because cylinder sets are a $\cap$-stable generator of $\mathfrak{A}$. This is easy, because

$$f^{-1}\big\{(x_k) \,:\, x_i = y_i \text{ for } i \in I\big\} = \big\{(x_k) \,:\, x_{i+1} = y_i \text{ for } i \in I\big\}.$$

By definition the set on the right hand side and $\{(x_k) : x_i = y_i \text{ for } i \in I\}$ have the same measure.

To check *ergodicity* let $A \in \mathfrak{A}$ with $A = f^{-1}(A)$ and $\mathbf{P}(A) > 0$. It suffices to show that

$$\mathbf{P}(A \cap E) = \mathbf{P}(A)\mathbf{P}(E) \text{ for all cylinder sets } E. \tag{1.7}$$

Indeed, then the probability measures $\mathbf{P}(A \cap \cdot)/\mathbf{P}(A)$ and $\mathbf{P}$ agree on the cylinder sets and hence they are equal. Then we may plug A itself into the equation and get $\mathbf{P}(A) = \mathbf{P}(A)^2$ implying $\mathbf{P}(A) \in \{0, 1\}$.

To show (1.7) we use Lemma B. We fix an arbitrary cylinder set

$$E = \big\{(x_k) : x_i = y_i \text{ for } i \in I\big\}, \qquad \text{for } I \subset \mathbb{Z} \text{ finite and } y_i \in \{1, \dots, n\}.$$

Suppose $C_1, \dots, C_K$ are pairwise disjoint cylinder sets such that their union approximates A in the sense of Lemma B. Then, for all N,

$$\mathbf{P}\big(A \cap E\big) = \mathbf{P}\big(f^{-N}(A) \cap E\big) = \mathbf{P}\big(A \cap f^{N}(E)\big)$$

and the latter term is

$$\begin{cases} \leq & \mathbf{P}\Big(\bigcup_{l=1}^{K} C_l \cap f^N(E)\Big) + \varepsilon, \\ \geq & \mathbf{P}\Big(\bigcup_{l=1}^{K} C_l \cap f^N(E)\Big) - \varepsilon. \end{cases}$$

Now assume $C = \big\{(x_k) : x_l = \tilde{y}_l \text{ for } l \in L\big\}$ and $L \subset \mathbb{Z}$ finite. Choose N so large that $\sup I - N < \inf L$. Then, using the definition of $\mathbf{P}$,

$$\begin{aligned} \mathbf{P}\big(C \cap f^N(E)\big) &= \mathbf{P}\big\{X_l = \tilde{y}_l \text{ for } l \in L\,,\ X_{i-N} = y_i \text{ for } i \in I\big\} \\ &= \mathbf{P}\big\{X_l = \tilde{y}_l \text{ for } l \in L\big\}\,\mathbf{P}\big\{X_{i+N} = y_i \text{ for } i \in I\big\} \\ &= \mathbf{P}(C)\,\mathbf{P}(E)\,. \end{aligned}$$

Taking the union over all C_l we infer that

$$\mathbf{P}\big(A \cap E\big) \begin{cases} \leq & \big(\mathbf{P}(A) + \varepsilon\big)\mathbf{P}(E) + \varepsilon, \\ \geq & \big(\mathbf{P}(A) - \varepsilon\big)\mathbf{P}(E) - \varepsilon. \end{cases}$$

Letting $\varepsilon \downarrow 0$ gives (1.7) and proves Step 3.

Step 4. We finish the proof by recalling all the steps we have done. We start the algorithm with the randomly chosen point $y_0 = \pi(X_{-1}, X_{-2}, X_{-3}, \dots)$, which has distribution μ by Remark 1.22. However, we recall from Step 1 that the algorithm works independently of this choice. Using the ergodic theorem in the third step, we have, $\mathbf{P}$-almost surely,

$$\begin{aligned} \lim_{k\to\infty} \frac{1}{k} \sum_{i=0}^{k-1} g(y_i) &= \lim_{k\to\infty} \frac{1}{k} \sum_{i=0}^{k-1} g\Big(\phi_{X_{i-1}} \circ \cdots \circ \phi_{X_0}\big(\pi(X_{-1}, X_{-2},, \dots)\big)\Big) \\ &= \lim_{k\to\infty} \frac{1}{k} \sum_{i=0}^{k-1} g\Big(\pi(x_{i-1}, x_{i-2},, \dots)\big)\Big) \\ &= \lim_{k\to\infty} \frac{1}{k} \sum_{i=0}^{k-1} \varphi\big(f^i((x_k))\big) = \int \varphi \, d\mathbf{P} \\ &= \int g \circ \pi(x_{-1}, x_{-2}, \dots)\, d\mathbf{P} = \int g \, d\mu\,. \end{aligned}$$

This finishes the proof of the chaos game theorem.

Chapter 2

Hausdorff dimension

In this chapter we introduce three notions of dimension, the Minkowski dimension, packing dimension and Hausdorff dimension. In particular the latter is crucial in fractal geometry. We show how Hausdorff dimension is used to determine the size of a set and describe techniques to calculate the Hausdorff dimension. In particular, we derive a formula for the Hausdorff dimension of self-similar sets.

2.1 Minkowski, Hausdorff and packing dimension

How can we determine the size of a geometric object? There are two levels on which this question can be answered, the crude level is the level of dimension. It allows, for example, to distinguish a line segment, which is a one-dimensional object, from a planar square, which is a two-dimensional object. But it does not allow to distinguish, for example, line segments of different length.

How can we capture the dimension of a geometric object? One requirement for a useful definition of dimension is that it should be *intrinsic*. This means that it should be independent of an embedding of the object in an ambient space like $\mathbb{R}^d$. Intrinsic notions of dimension can be defined in arbitrary metric spaces.

2.1.1 The Minkowski dimension

Suppose E is a bounded metric space with metric d. Here bounded means that the diameter of E is finite. The example we have in mind is a bounded subset of $\mathbb{R}^d$. Define, for $\varepsilon > 0$,

$$M(E,\varepsilon) = \min\Big\{k \geq 1 \,:\, \text{ there exist } x_1, \dots, x_k \in E \text{ with } E \subset \bigcup_{i=1}^{k} B(x_i,\varepsilon)\Big\},$$

where $B(x,\varepsilon) = \{y \in E \,:\, d(x,y) < \varepsilon\}$ is the open ball around x of radius ε. Intuitively, when E has dimension d the number $M(E,\varepsilon)$ should be approximately C/ε^d. This can be verified in simple cases like line segments, planar squares, etc. This argument motivates the definition of *Minkowski dimension*.

Definition 2.1 *For a bounded metric space E we define the* lower Minkowski dimension *as*

$$\underline{\dim}_M E := \liminf_{\varepsilon \downarrow 0} \frac{\log M(E,\varepsilon)}{\log(1/\varepsilon)},$$

and the upper Minkowski dimension *as*

$$\overline{\dim}_M E := \limsup_{\varepsilon \downarrow 0} \frac{\log M(E,\varepsilon)}{\log(1/\varepsilon)}.$$

We always have $\underline{\dim}_M E \le \overline{\dim}_M E$, but equality need not hold. If it holds we write $\dim_M E = \underline{\dim}_M E = \overline{\dim}_M E$.

We shall see below that these definitions have limitations. Before that, let us calculate the Minkowski dimension of the Cantor set.

Lemma 2.2 *If C is the ternary Cantor set, we have $\dim_M C = \frac{\log 2}{\log 3}$.*

Proof: Let $\alpha = \log 2/\log 3$. For the upper bound it suffices to find an efficient covering of C by balls of radius ε. If $\varepsilon \in (0,1)$ is given, let n be the integer such that $1/3^n < \varepsilon \le 1/3^{n-1}$ and look at the sets

$$\Big[\sum_{i=1}^n \frac{x_i}{3^i}, \sum_{i=1}^n \frac{x_i}{3^i} + \varepsilon\Big] \text{ for } (x_1, \dots, x_n) \in \{0,2\}^n.$$

These sets obviously cover C and each of them is contained in an open ball centred in a point of C of radius ε. Hence

$$M(C,\varepsilon) \le 2^n = 3^{\alpha n} = 3^\alpha \big(3^{n-1}\big)^\alpha \le 3^\alpha (1/\varepsilon)^\alpha.$$

This implies $\overline{\dim}_M C \le \alpha$.

For the lower bound we assume we have a cover by open balls $B(x_k,\varepsilon)$ of radius $\varepsilon \in (0,1)$, and let n be the integer such that $1/3^{n+1} \le \varepsilon < 1/3^n$. Let

$$x_k = \sum_{i=1}^\infty \frac{x_{i,k}}{3^i}.$$

Then

$$B(x_k,\varepsilon) \cap C \subset \Big\{ \sum_{i=1}^\infty \frac{y_i}{3^i} \,:\, y_1 = x_{1,k}, \dots, y_n = x_{n,k} \Big\},$$

and we need at least 2^n sets of the latter type to cover C. Hence,

$$M(C,\varepsilon) \ge 2^n = 3^{\alpha n} = (1/3)^\alpha \big(3^{n+1}\big)^\alpha \ge (1/3)^\alpha (1/\varepsilon)^\alpha.$$

This implies $\underline{\dim}_M C \ge \alpha$. ∎

Our second example indicates an unpleasant limitation of the notion of Minkowski dimension. Observe first that singletons $S = \{x\}$ have Minkowski dimension 0. However, a countable set may have positive dimension, as the following example shows.

Lemma 2.3 *Let* $E := \{1/n \,:\, n \in \mathbb{N}\} \cup \{0\}$. *Then* $\dim_M E = \frac{1}{2}$.

Proof: Given $\varepsilon \in (0,1)$ find the integer n such that $1/(n+1)^2 \le \varepsilon < 1/n^2$. Then the points in $\{1/k \,:\, k > n\} \cup \{0\}$ can be covered by $n+1$ balls of radius ε. n further balls suffice to cover the remaining n points. Hence

$$M(E,\varepsilon) \le 2n+1 \le \tfrac{2n+1}{n}\,(1/\varepsilon)^{1/2}\,,$$

implying $\overline{\dim}_M(E) \le 1/2$. On the other hand, as the distance between two neighbouring points is

$$\frac{1}{k} - \frac{1}{k+1} = \frac{1}{k(k+1)} \ge \frac{1}{(k+1)^2}\,,$$

we always need at least $n-1$ balls of radius ε to cover E, which implies

$$M(E,\varepsilon) \ge n-1 \ge \tfrac{n-1}{n+1}\,(1/\varepsilon)^{1/2}\,,$$

hence $\underline{\dim}_M(E) \ge 1/2$. ■

The previous example shows that the Minkowski dimension does not have the *countable stability property*

$$\dim \bigcup_{k=1}^{\infty} E_k = \sup\big\{\dim E_k \,:\, k \ge 1\big\}.$$

This is one of the properties we expect from a reasonable concept of dimension. There are two ways out of this problem.

(i) One can use a notion of dimension based on covering with balls of varying size. This captures finer details of the set and leads to the notion of *Hausdorff dimension.*

(ii) One can enforce the countable stability property by subdividing every set in countably many bounded pieces and taking the maximal dimension of them. The infimum over the numbers such obtained leads to the notion of *packing dimension.*

2.1.2 The Hausdorff dimension

The Hausdorff dimension and Hausdorff measure were introduced by Felix Hausdorff in 1919. Like the Minkowski dimension, Hausdorff dimension can be based on the notion of a *covering* of the metric space E by balls. A *covering* of E by balls is an at most countable collection of balls

$$B(x_1,r_1), B(x_2,r_2), B(x_3,r_3), \ldots$$

with

$$E \subset \bigcup_{i=1}^{\infty} B(x_i,r_i)\,.$$

For every $s \ge 0$ we say that the *s-value* of the covering is

$$\sum_{i=1}^{\infty} r_i^s\,.$$

In the case of Minkowski dimension we look at coverings of E by finitely many balls $B(x_1,\varepsilon),\dots,B(x_k,\varepsilon)$ with *equal radius* and the s-value is in this case $k\varepsilon^s$. We define the number

$$M^s(E)=\inf\Big\{k\varepsilon^s\,:\,B(x_1,\varepsilon),\dots,B(x_k,\varepsilon)\text{ is a covering of }E\Big\},$$

roughly speaking this corresponds to the value of the most efficient covering by balls of equal radius. We now show that

$$\underline{\dim}_M E=\inf\Big\{s\geq 0\,:\,M^s(E)=0\Big\}=\sup\Big\{s\geq 0\,:\,M^s(E)>0\Big\}.$$

Indeed, if $\underline{\dim}_M E<s$ then, for every $\underline{\dim}_M E<\eta<s$, there exists arbitrarily small $\varepsilon>0$, and a covering by balls of radius ε with $k\leq 1/\varepsilon^\eta$ balls and hence an s-value smaller than

$$\frac{\varepsilon^s}{\varepsilon^\eta}\xrightarrow{\varepsilon\downarrow 0}0\,.$$

We infer that $M^s(E)=0$. If $\underline{\dim}_M E>s$ then, for every $\underline{\dim}_M E>\eta>s$, for sufficiently small $\varepsilon>0$, every covering by balls of radius ε consists of at least $k\geq 1/\varepsilon^\eta$ balls, hence must have an s-value exceeding

$$\frac{\varepsilon^s}{\varepsilon^\eta}\xrightarrow{\varepsilon\downarrow 0}\infty\,.$$

We infer that $M^s(E)>0$, proving the claim.

Looking back at the example of Lemma 2.3 one can see that covering with sets of fixed radius does not capture the fine features of a set if these occur in very different scales at different places. The set $\{1/n\,:\,n\geq 1\}\cup\{0\}$ can be covered much more effectively if we are allowed to decrease the size of the balls as we move from right to left. The terminology of the s-values of a covering allows to formulate a concept of dimension, which is sensitive to this effect.

Definition 2.4 *For every $\alpha\geq 0$ the* (spherical) α-Hausdorff content *of a metric space E is defined as*

$$\mathcal{S}^\alpha(E)=\inf\Big\{\sum_{i=1}^\infty r_i^\alpha\,:\,(B(x_i,r_i))\text{ is a covering of }E\Big\},$$

informally speaking the α-value of the most efficient covering. If $0\leq s\leq t$, and $\mathcal{S}^s(E)=0$, then also $\mathcal{S}^t(E)=0$. Thus we can define

$$\dim E=\inf\Big\{s\geq 0\,:\,\mathcal{S}^s(E)=0\Big\}=\sup\Big\{s\geq 0\,:\,\mathcal{S}^s(E)>0\Big\},$$

the Hausdorff dimension *of the set E.*

Remark 2.5

(1) *The Hausdorff dimension may, of course, be infinite. But from the discussion above it is obvious that $\dim E\leq\underline{\dim}_M E$ and, in particular, subsets of $\mathbb{R}^d$ have Hausdorff dimension no larger than d.*

Lemma 2.6 *The Hausdorff dimension has the countable stability property.*

Proof: Indeed, as $E \subset F$ implies $\dim E \le \dim F$, it is obvious that

$$\dim \bigcup_{k=1}^{\infty} E_k \ge \sup\big\{ \dim E_k \,:\, k \ge 1 \big\}.$$

To see the converse, we use

$$\begin{aligned}
\mathcal{S}^s\Big(\bigcup_{k=1}^{\infty} E_k\Big) &\le \inf\Big\{ \sum_{k=1}^{\infty}\sum_{j=1}^{\infty} r_{j,k}^s \,:\, \big(B(x_{j,k}, r_{j,k}) \,:\, j \ge 1\big) \text{ covers } E_k \Big\} \\
&= \sum_{k=1}^{\infty} \inf\Big\{ \sum_{j=1}^{\infty} r_{j,k}^s \,:\, \big(B(x_{j,k}, r_{j,k}) \,:\, j \ge 1\big) \text{ covers } E_k \Big\} \\
&= \sum_{k=1}^{\infty} \mathcal{S}^s(E_k)\,.
\end{aligned}$$

Hence,

$$\begin{aligned}
\dim \bigcup_{k=1}^{\infty} E_k &\le \sup\Big\{ s \ge 0 \,:\, \mathcal{S}^s\big(\bigcup_{k=1}^{\infty} E_k\big) > 0 \Big\} \\
&\le \sup\Big\{ s \ge 0 \,:\, \sum_{k=1}^{\infty} \mathcal{S}^s\big(E_k\big) > 0 \Big\} \\
&\le \sup_{k=1}^{\infty} \sup\Big\{ s \ge 0 \,:\, \mathcal{S}^s\big(E_k\big) > 0 \Big\}\,.
\end{aligned}$$

This proves the converse inequality. ■

The concept of the α-Hausdorff content plays an important part in the definition of the Hausdorff dimension. However, it does not help distinguishing the size of sets of the same dimension. For example a line segment of unit length and a plus consisting of two orthogonal line segments of unit length have the same 1-Hausdorff content. Therefore one considers a refined concept, the *Hausdorff measure.* Here the idea is to consider only coverings by *small* sets, which need not be balls.

Definition 2.7 *Let X be a metric space and $E \subset X$. For every $\alpha \ge 0$ and $\delta > 0$ define*

$$\mathcal{H}_\delta^\alpha(E) = \inf\Big\{ \sum_{i=1}^{\infty} |E_i|^\alpha \,:\, \bigcup_{i=1}^{\infty} E_i \supset E \text{ with } |E_i| \le \delta \Big\}\,,$$

i.e. we are considering coverings of E by sets of diameter no more than δ. Then

$$\mathcal{H}^\alpha(E) = \sup_{\delta > 0} \mathcal{H}_\delta^\alpha(E) = \lim_{\delta \downarrow 0} \mathcal{H}_\delta^\alpha(E)$$

*is the α-*Hausdorff measure *of the set E.*

Remark 2.8

(1) *It can be shown with some effort that the mapping* $B \mapsto \mathcal{H}^\alpha(B)$ *is a measure on the Borel-σ-field of* X*. This is proved in my script* Maßtheorie.

(2) *If* $X = \mathbb{R}^d$ *and* $\alpha = d$ *the Hausdorff measure* $\mathcal{H}^\alpha$ *is a constant multiple Lebesgue measure. If* α *is an integer and* X *is an embedded α-submanifold, then* $\mathcal{H}^\alpha$ *is the surface measure. This idea can also be used to develop vector analysis on sets with much less smoothness than a differentiable manifold. Some hints can be found in my script* Maßtheorie.

One can express the Hausdorff dimension in terms of the Hausdorff measure.

Lemma 2.9 *For every metric space* E *we have*

$$\begin{aligned} \dim E &= \inf\{s : \mathcal{H}^s(E) = 0\} = \inf\{s : \mathcal{H}^s(E) < \infty\} \\ &= \sup\{s : \mathcal{H}^s(E) > 0\} = \sup\{s : \mathcal{H}^s(E) = \infty\}. \end{aligned}$$

Proof: Let us prove the first equality only, the other arguments are similar. Suppose $\dim E > s$. Then, for all $t \le s$, $c := \mathcal{S}^t(E) > 0$. As an arbitrary set is contained in a ball whose radius is the diameter of the set, we have $\mathcal{H}^t_\delta(E) \ge c > 0$ for all $\delta > 0$. Hence, $\mathcal{H}^t(E) \ge c > 0$ and this implies $\mathcal{H}^t(E) > 0$ for all $t \le s$. We infer that $\inf\{t : \mathcal{H}^t(E) = 0\} \ge s$.

Conversely, if $\dim E < s$, then $\mathcal{S}^s(E) = 0$ and hence, for every $\delta > 0$, there exists a covering by sets $E_1, E_2, \dots$ with $\sum_{k=1}^\infty |E_k|^s < \delta$. These sets have automatically diameter less than $\delta^{1/s}$, hence $\mathcal{H}^s_{\delta^{1/s}}(E) < \delta$ and letting $\delta \downarrow 0$ yields $\mathcal{H}^s(E) = 0$. This proves $\inf\{t : \mathcal{H}^t(E) = 0\} \le s$. ■

2.1.3 The packing dimension

Packing dimension was introduced surprisingly late by Tricot (1982) and should perhaps be called Tricot dimension. Our first approach to packing dimension does not reveal why the name packing dimension is chosen.

The notion of packing dimension can be founded on *regularization* of the upper Minkowski dimension. For every metric space E we define

$$\dim_P E = \inf\Big\{ \sup_{i=1}^{\infty} \overline{\dim}_M E_i \,:\, E = \bigcup_{i=1}^{\infty} E_i \,, E_i \text{ bounded}\Big\}. \tag{2.1}$$

$\dim_P E$ is called the *packing dimension* of E. The reason for this terminology will become clear from Theorem 2.12 below.

Remark 2.10 *We have, for all bounded sets* E*, that* $\dim_P E \le \overline{\dim}_M E$ *and, of course, strict inequality may hold. Obviously, every countable set has packing dimension* 0*, compare with the example in Lemma 2.3. For this definition it is not hard to see that the countable stability property is satisfied.*

Lemma 2.11 *For every metric space* $\dim_P E \ge \dim E$.

Proof: For every decomposition $E = \bigcup_{i=1}^{\infty} E_i$ of E into bounded sets, we have, using countable stability of Hausdorff dimension,

$$\sup_{i=1}^{\infty} \overline{\dim}_M E_i \geq \sup_{i=1}^{\infty} \dim E_i = \dim \bigcup_{i=1}^{\infty} E_i = \dim E\,,$$

and passing to the infimum yields the statement. ■

The importance of the notion of packing dimension comes from a duality relation between Hausdorff and packing dimension: even if one is only interested in Hausdorff dimension, packing dimension is cropping up naturally at many places, see Example 2.14 below.

The natural *dual* operation to covering a set with balls, as we have done in the case of Hausdorff dimension, is the operation of *packing* a set with balls. For every $\delta > 0$, a δ*-packing* of $A \subset E$ is a countable collection of *disjoint* balls

$$B(x_1, r_1), B(x_2, r_2), B(x_3, r_3), \ldots$$

with centres $x_i \in A$ and radii $0 \leq r_i \leq \delta$. For every $s \geq 0$ we introduce the s*-value* of the packing as

$$\sum_{i=1}^{\infty} r_i^s\,.$$

The s*-packing number* of A is defined as

$$P^s(A) = \lim_{\delta \downarrow 0} P_\delta^s \quad \text{for } P_\delta^s = \sup \Big\{ \sum_{i=1}^{\infty} r_i^s \,:\, (B(x_i, r_i)) \text{ a } \delta\text{-packing of } A \Big\}\,.$$

Note that the packing number is defined in the same way as the Hausdorff measure with efficient (small) coverings replaced by efficient (large) packings. A difference is that the packing numbers do *not* define a reasonable measure. However a small modification gives the so-called packing measure,

$$\mathcal{P}^s(A) = \inf \Big\{ \sum_{i=1}^{\infty} P^s(A_i) \,:\, A = \bigcup_{i=1}^{\infty} A_i \Big\}.$$

The packing dimension has an alternative definition analogous to the definition of Hausdorff dimension with Hausdorff measures replaced by packing measures.

Theorem 2.12 *We have*

$$\begin{aligned} \dim_P E &= \inf\{s : \mathcal{P}^s(E) = 0\} = \inf\{s : \mathcal{P}^s(E) < \infty\} \\ &= \sup\{s : \mathcal{P}^s(E) > 0\} = \sup\{s : \mathcal{P}^s(E) = \infty\}\,. \end{aligned}$$

Proof: It is easy to see that the last three equalities hold. Hence, we only need to show the first equality. Define, for every $A \subset E$ and $\varepsilon > 0$,

$$P(A, \varepsilon) = \max \big\{ k : \text{ there are disjoint balls } B(x_1, \varepsilon), \ldots, B(x_k, \varepsilon) \text{ with } x_i \in A \big\}.$$

We first show that

$$P(A,4\varepsilon) \le M(A,2\varepsilon) \le P(A,\varepsilon)\,. \tag{2.2}$$

Indeed, if $k = P(A,\varepsilon)$ let $B(x_1,\varepsilon),\dots,B(x_k,\varepsilon)$ be disjoint balls with $x_i \in A$. Suppose $x \in A \setminus \bigcup_{i=1}^k B(x_i,2\varepsilon)$, then $B(x,\varepsilon)$ is disjoint from all balls $B(x_i,\varepsilon)$ contradicting the choice of k. Hence $B(x_1,2\varepsilon),\dots,B(x_k,2\varepsilon)$ is a cover of A and we have shown $M(A,2\varepsilon) \le P(A,\varepsilon)$.

For the other inequality let $m = M(A,2\varepsilon)$ and $k = P(A,4\varepsilon)$ and choose $x_1,\dots,x_m \in A$ and $y_1,\dots,y_k \in A$ such that

$$A \subset \bigcup_{i=1}^m B(x_i,2\varepsilon) \text{ and } B(y_1,4\varepsilon),\dots,B(y_k,4\varepsilon) \text{ disjoint.}$$

Then each y_j belongs to some $B(x_i,2\varepsilon)$ and no such ball contains more than one such point. Thus $k \le m$, which proves $P(A,4\varepsilon) \le M(A,2\varepsilon)$.

Suppose now that $\inf\{t : \mathcal{P}^t(E) = 0\} < s$. Then there is $t < s$ and $E = \bigcup_{i=1}^\infty A_i$ such that, for every set $A = A_i$, we have $P^t(A) < 1$. Obviously, $P^t_\varepsilon(A) \ge P(A,\varepsilon)\varepsilon^t$. Letting $\varepsilon \downarrow 0$ gives

$$\lim_{\varepsilon\downarrow 0} M(A,\varepsilon)\varepsilon^t \le \lim_{\varepsilon\downarrow 0} P(A,\varepsilon/2)\varepsilon^t \le 2^t P^t(A) < 2^t.$$

Hence $\overline{\dim}_M A \le t$ and by definition $\dim_P E \le t < s$.

To prove the opposite inequality, let

$$0 < t < s < \inf\{r : \mathcal{P}^r(E) = 0\}\,,$$

and $A_i \subset E$ bounded with $E = \bigcup_{i=1}^\infty A_i$. It suffices to show that $\overline{\dim}_M(A_i) \ge t$ for some i. Since $\mathcal{P}^s(E) > 0$ there is i such that $P^s(A_i) > 0$. Let $0 < \alpha < P^s(A_i)$, then for all $\delta \in (0,1)$ we have $P^s_\delta(A_i) > \alpha$ and there exist disjoint balls $B(x_1,r_1), B(x_2,r_2), B(x_3,r_3),\dots$ with centres $x_j \in A_i$ and radii r_j smaller than δ with

$$\sum_{j=1}^\infty r_j^s \ge \alpha\,.$$

For every m let k_m be the number of balls with radius $2^{-m-1} < r_j \le 2^{-m}$. Then,

$$\sum_{m=0}^\infty k_m 2^{-ms} \ge \sum_{j=1}^\infty r_j^s \ge \alpha\,.$$

This yields, for some integer $N \ge 0$,

$$2^{Nt}(1-2^{t-s})\alpha \le k_N\,,$$

since otherwise

$$\sum_{m=0}^\infty k_m 2^{-ms} < \sum_{m=0}^\infty 2^{mt}(1-2^{t-s})2^{-ms}\alpha = \alpha\,.$$

Since $r_j \le \delta$ for all j, we have $2^{-N-1} < \delta$. Moreover,

$$P(A_i, 2^{-N-1}) \ge k_N \ge 2^{Nt}(1-2^{t-s})\alpha\,,$$

which gives

$$\sup_{0\le\varepsilon\le\delta} P(A_i,\varepsilon)\varepsilon^t \ge P(A_i,2^{-N-1})2^{-Nt-t} \ge 2^{-t}(1-2^{t-s})\alpha\,.$$

Letting $\delta\downarrow 0$, and recalling (2.2), we obtain

$$\limsup_{\varepsilon\downarrow 0} M(A_i,\varepsilon)\varepsilon^t \ge \limsup_{\varepsilon\downarrow 0} P(A_i,2\varepsilon)\varepsilon^t > 0\,,$$

and thus $\overline{\dim}_M A_i \ge t$, as required. ■

Exercise 2.13 *Let $\{m_k : k\ge 1\}$ be a rapidly increasing sequence of positive integers such that*

$$\lim_{k\to\infty}\frac{m_k}{m_{k+1}} = 0.$$

Define two subsets of $[0,1]$ by

$$E = \Big\{ \sum_{i=1}^{\infty}\frac{x_i}{2^i} \,:\, x_i\in\{0,1\} \text{ and } x_i = 0 \text{ if } m_k+1\le i\le m_{k+1} \text{ for some even } k\Big\}$$

and

$$F = \Big\{ \sum_{i=1}^{\infty}\frac{x_i}{2^i} \,:\, x_i\in\{0,1\} \text{ and } x_i = 0 \text{ if } m_k+1\le i\le m_{k+1} \text{ for some odd } k\Big\}.$$

Show that

(1) $\dim E = \underline{\dim}_M E = 0$ *and* $\dim F = \underline{\dim}_M F = 0$,

(2) $\dim_P E = \overline{\dim}_M E = 1$ *and* $\dim_P F = \overline{\dim}_M F = 1$,

(3) $\dim(E\times F)\ge 1$.

Example 2.14 *As an example of a result demonstrating the duality between Hausdorff and packing dimension is the* product formula *of Bishop and Peres (1996). In the dimension theory of smooth sets (manifolds, linear spaces) we have the following formula for product sets*

$$\dim(E\times F) = \dim E + \dim F\,.$$

The example discussed in Exercise (2.13) shows that this formula fails for Hausdorff dimension, a reasonable formula for the Hausdorff dimension of product sets necessarily involves information about the packing dimension of one of the factor sets. Bishop and Peres have shown that, for every Borel set $A\subset\mathbb{R}^d$,

$$\dim_P(A) = \sup_B \Big\{ \dim(A\times B) - \dim(B)\Big\}$$

where the supremum is over all compact sets $B\subset\mathbb{R}^d$. One can also show that, if A satisfies $\dim A = \dim_P A$, *then the product formula* $\dim(A\times B) = \dim A + \dim B$ *holds.*

2.2 Techniques for calculating the Hausdorff dimension

From the definition of the Hausdorff dimension it is plausible that in many cases it is relatively easy to give an upper bound on the dimension: just find an efficient covering of the set. However it looks more difficult to give lower bounds, as we must obtain a lower bound on α-values of *all* coverings of the set. In this section we discuss two important techniques to obtain lower bounds for the Hausdorff dimension.

2.2.1 The mass distribution principle

The basic idea of the *mass distribution principle* is the following: if it is possible to distribute a positive amount of mass on a set E in such a manner that its local concentration is bounded above, then the set must be large in a suitable sense.

For the purpose of this method we call a measure μ on the Borel sets of a metric space E a *mass distribution* on E, if

$$0 < \mu(E) < \infty .$$

The intuition here is that a positive and finite mass is spread over the space E.

Theorem 2.15 (Mass distribution principle) *Suppose E is a metric space and $s \geq 0$. If there is a mass distribution μ on E and constants $C > 0$ and $\delta > 0$ such that*

$$\mu(U) \leq C|U|^s$$

for all closed sets U with diameter $|U| \leq \delta$. Then

$$\mathcal{H}^s(E) \geq \frac{\mu(E)}{C} > 0,$$

and hence $\dim E \geq s$.

Proof: Suppose that $\{U_i : i \geq 1\}$ is any covering of E with $|U_i| \leq \delta$. Denote by $V_i =\mathrm{cl}\, U_i$ the closure of the covering sets and observe that $|V_i| = |U_i|$. We have

$$0 < \mu(E) \leq \mu\Big(\bigcup_{i=1}^{\infty} V_i\Big) \leq \sum_{i=1}^{\infty} \mu(V_i) \leq C \sum_{i=1}^{\infty} |V_i|^s = C \sum_{i=1}^{\infty} |U_i|^s .$$

Passing to the infimum over all coverings with diameter $\leq \delta$ yields

$$\mathcal{H}^s_\delta(E) \geq \frac{\mu(E)}{C} ,$$

and letting $\delta \downarrow 0$ gives the statement. ■

Remark 2.16

(1) *The mass distribution principle is the reason why we were so keen on defining probability measures on self-similar sets, see the next section for the application of the mass distribution principle to self-similar sets.*

(2) *The following converse of the mass distribution principle is called* Frostman's Lemma*:*

Suppose that E is a complete separable metric space and $\mathcal{H}^s(E) > 0$. Then there exists a mass distribution μ on E such that $\mu(B) \le |B|^s$ for all bounded Borel sets B.

This result was proved by Howroyd (1995) in the present setting, but result for closed subsets of $\mathbb{R}^d$ is older and probably due to Frostman (1935). In the proof the main difficulty arising is that, if $\mathcal{H}^s(E) = \infty$, one has to find a subset $A \subset E$ with $0 < \mathcal{H}^s(A) < \infty$.

2.2.2 The potential theoretic method

In this section we present a lower bound for the Hausdorff dimension, which is particularly interesting in applications to random fractals: The potential theoretic method. It is based on a localization of the mass distribution principle.

Definition 2.17 *Suppose μ is a mass distribution on E and $s \ge 0$. Then, for every $x \in E$, the value*

$$\overline{d}_s(\mu, x) = \limsup_{r \downarrow 0} \frac{\mu(B(x,r))}{r^s}$$

is called the upper s-density *of μ at x.*

Theorem 2.18 (Local mass distribution principle) *If μ is a mass distribution on the metric space E, and $A \subset E$ a Borel set with*

$$\overline{d}_s(\mu, x) < C \text{ for all } x \in A\,,$$

then $\mathcal{H}^s(A) \ge \frac{\mu(A)}{C}$, and, in particular, $\dim A \ge s$.

For the proof we make the following observation:

Lemma 2.19 *The function $x \mapsto \mu\big(B(x,r)\big)$ is upper semicontinuous and hence measurable.*

Proof: We have to show that, for all $a > 0$, the set

$$F = \{x \in E \,:\, \mu(B(x,r)) < a\}$$

is open. Let $x \in F$, then

$$\mu\big(B(x, r+\varepsilon)\big) \downarrow \mu\big(B(x,r)\big) \text{ as } \varepsilon \downarrow 0\,.$$

Hence, there is $\varepsilon > 0$ such that $\mu(B(x, r+\varepsilon)) < a$. We infer that, for all y with $d(y,x) < \varepsilon$,

$$\mu\big(B(y,r)\big) \le \mu\big(B(x, r+\varepsilon)\big) < a\,.$$

Thus $y \in F$ and we have shown the statement. ■

Proof: Now we prove the local mass distribution principle. Fix $\delta > 0$. The set

$$A_\delta = \Big\{x \in A \,:\, \forall 0 < r < \delta \text{ we have } \mu\big(B(x,r)\big) \le Cr^s\Big\} = \bigcap_{\substack{0 \le r \le \delta \\ r \in \mathbb{Q}}} \Big\{x \in A \,:\, \mu\big(B(x,r)\big) \le Cr^s\Big\}$$

is a Borel set. Suppose $\{U_i : i \geq 1\}$ is a covering of A by sets of diameter at most δ. If $x \in U_i \cap A_\delta$, then $U_i \subset B(x, |U_i|)$ and thus $\mu(U_i) \leq \mu\big(B(x,|U_i|)\big) \leq C|U_i|^s$, and hence

$$\mu(A_\delta) \leq \sum_{U_i \cap A_\delta \neq \emptyset} \mu(U_i) \leq C \sum_{i=1}^{\infty} |U_i|^s.$$

Passing to the infimum over all such coverings yield

$$\mu(A_\delta) \leq C\mathcal{H}^s_\delta(A).$$

Now let $\delta \downarrow 0$ and observe that $\mu(A_\delta) \uparrow \mu(A)$ and $\mathcal{H}^s_\delta(A) \uparrow \mathcal{H}^s(A)$. We infer that $\mu(A) \leq C\mathcal{H}^s(A)$, as claimed. ■

We now come to the potential theoretic method and introduce the necessary terminology.

Definition 2.20 *Suppose μ is a mass distribution on a metric space E and $\alpha \geq 0$. The* α-potential *of a point $x \in E$ with respect to μ is defined as*

$$\phi_\alpha(x) = \int \frac{d\mu(y)}{d(x,y)^\alpha}.$$

In the case $E = \mathbb{R}^3$ and $\alpha = 1$, this is the Newton gravitational potential of the mass μ. The α-energy *of μ is*

$$I_\alpha(\mu) = \int \phi_\alpha(x)\, d\mu(x) = \int\int \frac{d\mu(x)\, d\mu(y)}{d(x,y)^\alpha}.$$

The (Riesz) α-capacity *of E is defined as*

$$C_\alpha(E) = \sup\Big\{ I_\alpha(\mu)^{-1} : \mu \textit{ a mass distribution on } E \textit{ with } \mu(E) = 1 \Big\}.$$

The idea of the potential theoretic method is that mass distributions with $I_s(\mu) < \infty$ for large s spread the mass so that at each place the concentration is small. This is only possible on large sets.

Theorem 2.21 (Potential theoretic method) *Let $s \geq 0$ and E a metric space. Suppose there is a mass distribution μ on E with $I_s(\mu) < \infty$. Then $\mathcal{H}^s(E) = \infty$ and, in particular, $\dim E \geq s$.*

Remark 2.22 *In order to get a lower bound on the dimension from this method it suffices to show finiteness of a single integral. This is very convenient for random sets, where it suffices to show that $\mathbb{E}I_s(\mu) < \infty$ for a (random) measure on the random set E in order to get $\dim E \geq s$ almost surely.*

Proof: Let μ be a mass distribution μ on E with $I_s(\mu) < \infty$. Then, clearly, $\mu\{x\} = 0$ for all $x \in E$. Put

$$E_1 = \big\{ x \in E : \overline{d}_s(\mu, x) > 0 \big\},$$

and note, using Lemma 2.19, that this is a Borel set. If $x \in E_1$ there exists $\varepsilon > 0$ and $r_i \downarrow 0$ such that

$$\mu\big(B(x,r_i)\big) \geq \varepsilon r_i^s \,.$$

As $\mu\{x\} = 0$ there exist small $0 < q_i < r_i$ such that $B_i = B(x,r_i) \setminus B(x,q_i)$ satisfies

$$\mu(B_i) \geq \frac{1}{4}\varepsilon r_i^s \,.$$

By passing to a subsequence, if necessary, we may assume $r_{i+1} < q_i$ and hence the sets $B_1, B_2, B_3, \ldots$ are pairwise disjoint. We infer that

$$\phi_s(x) = \int \frac{d\mu(y)}{d(x,y)^s} \geq \sum_{i=1}^{\infty} \int_{B_i} \frac{d\mu(y)}{d(x,y)^s} \geq \frac{1}{4} \sum_{i=1}^{\infty} \varepsilon r_i^s r_i^{-s} = \infty \,.$$

However, because

$$I_s(\mu) = \int \phi_s(x) d\mu(x) < \infty \,,$$

we infer that $\mu(E_1) = 0$. But for all $x \in E \setminus E_1$ we have $\overline{d}_s(\mu, x) = 0$ and thus, for all $C > 0$ by the local mass distribution theorem,

$$\mathcal{H}^s(E) \geq \mathcal{H}^s(E \setminus E_1) \geq \frac{\mu(E \setminus E_1)}{C} = \frac{\mu(E)}{C} \,.$$

Hence, letting $C \downarrow 0$, we get $\mathcal{H}^s(E) = \infty$, as claimed. ■

2.3 The dimension of self-similar sets

We start with an upper bound, which can be given in a quite general situation.

Theorem 2.23 *Let $(\phi_1, \ldots, \phi_n)$ be an iterated function system on a closed subset $X \subset \mathbb{R}^d$ with contraction coefficients $(c_1, \ldots, c_n)$. Then there is exactly one number $s \geq 0$ with*

$$\sum_{i=1}^{n} c_i^s = 1 \,.$$

If K is the attractor of the iterated function system, then $\mathcal{H}^s(K) \leq |K|^s$ and $\dim K \leq s$.

Proof: The function $p(x) = \sum_{i=1}^n c_i^x$ is continuous, strictly decreasing with $p(0) = n \geq 1$ and $\lim_{x\to\infty} p(x) = 0$. Hence, by the intermediate value theorem, there is exactly one $s \geq 0$ with $p(s) = 1$, which is the unique solution of our problem.

By induction from the definition of the attractor, we see that, for every k,

$$K = \bigcup_{\substack{(x_1,\ldots,x_k)\\ \in\{1,\ldots,n\}^k}} \phi_{x_1} \circ \cdots \circ \phi_{x_k}(K) \,.$$

Recall that $c = \max_{i=1}^n c_i < 1$. As

$$\left|\phi_{x_1} \circ \cdots \circ \phi_{x_k}(K)\right| \le c_{x_1} \cdots c_{x_k} \, |K| \le c^k \, |K| \, .$$

Given $\delta > 0$ we can choose k so large that $c^k|K| < \delta$. Then

$$\left\{\phi_{x_1} \circ \cdots \circ \phi_{x_k}(K) \, : \, (x_1, \ldots , x_k) \in \{1, \ldots , n\}^k\right\}$$

is a covering of K with s-value

$$\begin{aligned} \sum_{\substack{(x_1,\ldots,x_k) \\ \in\{1,\ldots,n\}^k}} \left|\phi_{x_1} \circ \cdots \circ \phi_{x_k}(K)\right|^s &\le \sum_{\substack{(x_1,\ldots,x_k) \\ \in\{1,\ldots,n\}^k}} c^s_{x_1} \circ \cdots \circ c^s_{x_k} |K|^s \\ &= \sum_{x_1=1}^n c^s_{x_1} \cdots \sum_{x_k=1}^n c^s_{x_k} |K|^s = |K|^s \, , \end{aligned}$$

using the defining property of s. This implies that $\mathcal{H}^s_\delta(K) \le |K|^s$ and the result follows by letting $\delta \downarrow 0$. ■

When can this upper bound be sharp? First of all the contraction property needs to be sharp. This is the case if the contractions ϕ_i are similitudes. Secondly, the parts $\phi_i(K)$ should not overlap too much for different values of i. This is the case if $\phi_i(K) \cap \phi_j(K) = \emptyset$. This last condition, frequently called the *strong separation condition* is too restrictive, it excludes examples like the Sierpinski gasket and the von Koch curve, where the parts $\phi_i(K)$ touch in single points, but do not overlap. We now formulate a weaker condition, expressing that parts may touch, but do not overlap.

Definition 2.24 *An iterated function system $(\phi_1, \ldots , \phi_n)$ consisting of similitudes $\phi_i : \mathbb{R}^d \to \mathbb{R}^d$ satisfies the* open set condition, *if there is a nonempty, bounded open set $V \subset \mathbb{R}^d$ with $\phi_i(V) \cap \phi_j(V) = \emptyset$ for all $i \ne j$, and $V \supset \bigcup_{i=1}^n \phi_i(V)$.*

Example 2.25

(1) *If $\phi_i(K) \cap \phi_j(K) = \emptyset$ for all $i \ne j$, then the open set condition is satisfied.*

(2) *The von Koch curve and the Sierpinski gasket satisfy the open set condition.*

Proof: In (1) we choose $\varepsilon > 0$ so small that, for the set

$$V = \left\{x \in \mathbb{R}^d \, : \, \text{there is } y \in K \text{ with } d(x, y) < \varepsilon\right\}$$

we have $\phi_i(V) \cap \phi_j(V) = \emptyset$ for all $i \ne j$. Then, recalling the definition of the parallel body,

$$\bigcup_{i=1}^n \phi_i(V) \subset \bigcup_{i=1}^n \phi_i(K[\varepsilon]) \subset \Big(\bigcup_{i=1}^n \phi_i(K)\Big)[c\varepsilon] = K[c\varepsilon] \subset V \, .$$

This shows that the open set condition is fulfilled in the situation of Example (1) and (2) can be demonstrated with a picture. ■

Theorem 2.26 *Let K be a self-similar set, which is the attractor of the iterated function system $(\phi_1, \dots, \phi_n)$ consisting of similitudes $\phi_i : \mathbb{R}^d \to \mathbb{R}^d$ with contraction ratios $0 < r_i < 1$. The unique solution s of*

$$\sum_{i=1}^{n} r_i^s = 1\,,$$

is called the similarity dimension *of the iterated function system. Suppose that the open set condition is fulfilled, then*

$$\dim K = s\,,$$

and, moreover, $0 < \mathcal{H}^s(K) < \infty$.

Example 2.27

(1) *The Hausdorff dimension of the* unit interval, dimension *is the solution s of $\sum_{i=1}^{2}(1/2)^s = 1$, which is 1.*

(2) *The Hausdorff dimension of the* Cantor set *is the solution s of $\sum_{i=1}^{2}(1/3)^s = 1$, which is* $\log 2/\log 3$.

(3) *The Hausdorff dimension of the* Sierpinski gasket *is the solution s of $\sum_{i=1}^{3}(1/2)^s = 1$, which is* $\log 3/\log 2$.

(4) *The Hausdorff dimension of the* von Koch curve *is the solution s of $\sum_{i=1}^{4}(1/3)^s = 1$, which is* $\log 4/\log 3$.

For the **proof of Theorem 2.26** note that, considering Theorem 2.23, it suffices to prove $\mathcal{H}^s(K) > 0$. The first step in the proof is the following lemma.

Lemma 2.28 *Suppose that $\{V_i : i \geq 1\}$ is a family of disjoint open subsets of $\mathbb{R}^d$, such that every V_i contains a ball of radius $a_1 r$ and is contained in a ball of radius $a_2 r$. Then every ball of radius r intersects at most $(1+2a_2)^d a_1^{-d}$ of the sets in $\{cl\,(V_i) : i \geq 1\}$.*

Proof: Suppose $B = B(x,r)$ is a ball of radius r. Suppose that V_i is such that $B \cap \operatorname{cl} V_i \neq \emptyset$. Then

$$V_i \subset B\big(x, (1+2a_2)r\big)\,.$$

Suppose q sets from $\{\operatorname{cl} V_i : i \geq 1\}$ intersect B. By comparison of volume, denoting the Lebesgue measure by ℓ,

$$\begin{aligned}(1+2a_2)^d\, r^d\, \ell(B(0,1)) &= \ell\big(B(x,(1+2a_2)r)\big)\\ &\geq \sum_{\operatorname{cl} V_i \cap B \neq \emptyset} \ell\big(V_i\big)\\ &\geq q(a_1 r)^d \ell\big(B(0,1)\big)\,,\end{aligned}$$

hence $q \leq (1+2a_2)^d a_1^{-d}$. ■

We now **complete the proof of Theorem 2.26.** We use the notation, for $E \subset X$,

$$E[x_1, \ldots, x_k] = \phi_{x_1} \circ \cdots \circ \phi_{x_k}(E).$$

The proof is based on the mass distribution principle. The mass distribution we are using is the self-similar measure μ associated with the iterated function system $(\phi_1, \ldots, \phi_n)$ and the probability vector $(r_1^s, \ldots, r_n^s)$. Here it is vital to note that our choice of s is the only one which makes this vector indeed a probability vector. Recall that

$$\mu\big(K[x_1, \ldots, x_k]\big) \geq \prod_{j=1}^{k} r_{x_j}^s.$$

Now we have to prove that this measure μ satisfies the conditions of the mass distribution principle. Let V be the bounded open set from the open set condition of our iterated function system. Let $a_1 > 0$ be such that V contains a ball of radius a_1, and let a_2 be the radius of a ball containing V.

Let $1 > \rho > 0$. For every $(x_i) \in \Sigma_n$ there is exactly one k with

$$r_{x_1} \cdots r_{x_k} \leq \rho < r_{x_1} \cdots r_{x_{k-1}}.$$

Let

$$\mathcal{S} = \big\{(x_1, \ldots, x_k) \,:\, (x_i) \in \Sigma_n\big\}.$$

By the open set condition, the set system

$$\big\{V[x_1, \ldots, x_k] \,:\, (x_1, \ldots, x_k) \in \mathcal{S}\big\}$$

consists of pairwise disjoint sets. Every set $V[x_1, \ldots, x_k]$ contains a ball of radius

$$a_1 r_{x_1} \cdots r_{x_k} \geq a_1 \min_{j=1}^{n} r_j \cdot \rho$$

and is contained in a ball of radius

$$a_2 r_{x_1} \cdots r_{x_k} \leq a_2 \cdot \rho.$$

Now we use Lemma 2.28 and see that every ball B of radius ρ intersects at most

$$q = (1 + 2a_2)^d a_1^{-d} \big(\min_{j=1}^{n} r_j\big)^{-d}$$

sets from the collection

$$\big\{\mathrm{cl}\, V[x_1, \ldots, x_k] \,:\, (x_1, \ldots, x_k) \in \mathcal{S}\big\}.$$

We now show that this set system covers K. With the notation of the first chapter we have

$$V \supset \bigcup_{i=1}^{n} \phi_i(V) = \Phi(V).$$

Hence, using also continuity of the mappings $\phi_1, \ldots, \phi_n$,

$$\mathrm{cl}\, V \supset \Phi\big(\mathrm{cl}\, V\big) \supset \Phi^2\big(\mathrm{cl}\, V\big) \supset \ldots.$$

From the results of the first chapter we know that

$$\lim_{k\to\infty} \phi^k\big(\mathrm{cl}\, V\big) = K\,,$$

in the Hausdorff metric. Altogether this gives $\mathrm{cl}\, V \supset K$ and

$$\mathrm{cl} V[x_1, \dots, x_k] \supset K[x_1, \dots, x_k]\,.$$

As the collection

$$\big\{K[x_1, \dots, x_k] \,:\, (x_1, \dots, x_k) \in \mathcal{S}\big\}$$

covers the set K, this is also the case for

$$\big\{\mathrm{cl}\, V[x_1, \dots, x_k] \,:\, (x_1, \dots, x_k) \in \mathcal{S}\big\}\,.$$

If now A is a closed set with diameter ρ, then there is a ball B of radius ρ containing A. We infer

$$\mu(A) \le \mu(B \cap K) \le \sum r_{x_1}^s \cdots r_{x_k}^s\,,$$

where the sum extends over all $(x_1, \dots, x_k) \in \mathcal{S}$ with

$$B \cap \mathrm{cl}\, V[x_1, \dots, x_k] \ne \emptyset\,.$$

But, by Lemma 2.28 these are at most q summands and, by definition of k, we infer

$$\mu(A) \le q\rho^s = (1 + 2a_2)^d a_1^{-d} \big(\min_{j=1}^{n} r_j\big)^{-d} \cdot |A|^s\,.$$

As this holds for all closed sets A with diameter less than 1, μ is a mass distribution satisfying the conditions of the mass distribution principle and we obtain $\mathcal{H}^s(K) \ge 1/q$ and, in particular, using also the upper bound of Theorem 2.23, we get $\dim K = s$.

Chapter 3

The geometry of fractal sets

In this chapter we show by example that self-similar sets have a very different geometry compared to the objects of classical geometry. The systematic study of these differences belongs to the mathematical field of *geometric measure theory.* We cannot give a more detailed treatment of this beautiful mathematical field here, but hope to make the reader a little bit curious to see more. An excellent reference for the subject of geometric measure theory is the book "The geometry of sets and measures in Euclidean spaces" by Pertti Mattila.

3.1 Projections of self-similar 1-sets

3.1.1 Projections and irregularity

The principal issue of this chapter is to juxtapose geometric properties of self-similar sets to properties of classical geometric objects. We call a set $C \subset \mathbb{R}^d$ a 1-set if

$$0 < \mathcal{H}^1(C) < \infty\,.$$

Note that this implies that $\dim C = 1$, but is strictly stronger. The most natural case for a comparison is between two 1-sets. We shall treat a striking example in this section.

Suppose that $C \subset \mathbb{R}^2$ is a self-similar set, which is the attractor of n contracting similitudes $\phi_1, \dots, \phi_n$ with contraction ratios $1/n$. We assume that the similitudes ϕ_k do not involve rotations, so that there are vectors $a_k \in \mathbb{R}^2$ with

$$\phi_k(x) = \frac{x}{n} + a_k\,.$$

We also assume that the strong separation condition holds. In this case the Hausdorff dimension of C equals 1, because

$$\sum_{k=1}^{n} \Big(\frac{1}{n}\Big)^{\alpha} = 1 \text{ iff } \alpha = 1.$$

Moreover, C is a 1-set. The easiest examples are *Sierpinski carpets*, based on dividing the unit square into n^2 non overlapping compact subsquares of sidelength $1/n$ and picking a subcollection of n disjoint squares.

It is natural to compare the 1-set C with the 1-sets typical from classical geometry, namely differentiable curves. For clarity, let $\gamma : [0,1] \to \mathbb{R}^2$ be an injective differentiable mapping, and consider its image $\gamma[0,1]$ as the *differentiable curve* . The set $\gamma[0,1]$ is easily seen to be a 1-set.

For a comparison of the 1-sets $\gamma[0,1]$ and C we will first use projections. We call a set $S \subset \mathbb{R}^2$ in the plane *invisible from direction* $\theta \in [0,\pi)$ if the orthogonal projection proj_θ along lines making an angle θ with the x-axis maps S onto a set of Lebesgue measure zero.

Projection of a differentiable curve $\gamma[0,1]$ gives an interval for all directions θ and this interval is degenerate to a single point only in the situation that γ is a straight line and we project parallel to this line. In other words, differentiable curves are visible from all but at most one directions. See how different the situation is for the self-similar set C.

Theorem 3.1 *The self-similar* 1*-sets* C *above are invisible from almost all directions* $\theta \in [0,\pi)$.

3.1.2 Self-similar 1-sets are invisible from almost all directions

Theorem 3.1 follows from general but hard results of geometric measure theory. In this section we give a direct proof, which is new and due to Peres, Simon and Solomyak (2000).

For every projection angle $\theta \in [0,\pi)$ we let L_θ be the line through the origin, which is orthogonal to the projection direction θ. By $C(\theta)$ we denote the *orthogonal projection* of C onto that line, the projection itself we call proj_θ.

Observation: $C(\theta)$ is a self-similar set on the line L_θ, which has *similarity dimension* 1. Indeed, if $a_i(\theta)$ is the projection of the vector a_i onto L_θ, then

$$C(\theta) = \bigcup_{i=1}^{n} \frac{1}{n} C(\theta) + a_i(\theta).$$

Of course, we must expect that the parts of $C(\theta)$ have a certain overlap.

We now formulate and prove some lemmas about self-similar sets K on the line with similarity dimension 1. So suppose that

$$K = \bigcup_{i=1}^{n} \psi_i(K) = \bigcup_{i=1}^{n} \frac{1}{n} K + b_i \subset \mathbb{R}, \tag{3.1}$$

for $\psi_i(x) = x/n + b_i$. Abbreviate $K_i = \psi_i(K)$. Denote by ℓ the Lebesgue measure on the line.

Lemma 3.2 $\ell(K_i \cap K_j) = 0$ *for all* $i \neq j$.

Proof: Fix $i \neq j$. We have

$$\begin{aligned} \ell(K) &= \ell\Big(\bigcup_{l=1}^{n} K_l\Big) \le \sum_{l=1}^{n} \ell(K_l) - \ell\big(K_i \cap K_j\big) = \frac{1}{n}\sum_{l=1}^{n} \ell(K) - \ell\big(K_i \cap K_j\big) \\ &= \ell(K) - \ell\big(K_i \cap K_j\big). \end{aligned}$$

Hence the last term must vanish. ■

Lemma 3.3 *There exists $i \neq j$ such that $K_i \cap K_j \neq \emptyset$.*

Proof: Suppose that all K_i and K_j with $i \neq j$ are disjoint. Then there exists $\varepsilon > 0$ such that the parallel bodies $K_i[\varepsilon]$ and $K_j[\varepsilon]$ are also disjoint. Now

$$\ell\big(K[\varepsilon]\big) = \sum_{i=1}^{n} \ell\big(K_i[\varepsilon]\big) = n\ell\Big(\big(\frac{1}{n}K\big)[\varepsilon]\Big) = n\ell\Big(\frac{1}{n}K[\varepsilon n]\Big) = \ell\big(K[\varepsilon n]\big).$$

This is a contradiction, as K is compact and nonempty. ∎

Lemma 3.4 *If $\ell(K) > 0$ and $\eta < 1$, there exists an interval J such that $\ell\big(K \cap J\big) > \eta|J|$.*

Proof: We can assume without loss of generality that $K \subset (0,1)$. Then $(0,1) \setminus K$ is an open set and its connected components are a collection of open intervals $I_1, I_2, I_3, \ldots$, which we order starting with the biggest one. If this collection is finite, we are done, so assume it is infinite. Clearly,

$$\ell(K) = 1 - \sum_{j=1}^{\infty} |I_j| \,.$$

For a given $\eta < 1$, we choose n so large that

$$\frac{\ell(K)}{\eta} > 1 - \sum_{j=1}^{n} |I_j| \,.$$

The complement $[0,1] \setminus \bigcup_{j=1}^{n} I_j$ is a collection $J_1, \ldots , J_{n+1}$ of disjoint compact intervals whose union contains K. Suppose now that $\ell\big(K \cap J_j\big) \leq \eta|J_j|$ for all j. Then,

$$\eta\Big(1 - \sum_{j=1}^{n} |I_j|\Big) < \ell(K) = \sum_{j=1}^{n+1} \ell\big(K \cap J_j\big) \leq \eta \sum_{j=1}^{n+1} |J_j| = \eta\Big(1 - \sum_{j=1}^{n} |I_j|\Big) \,.$$

As this is a contradiction, we infer that there must exist some j with $\ell\big(K \cap J_j\big) > \eta|J_j|$. ∎

By iterating (3.1) we get

$$K = \bigcup_{u:|u|=m} K_u,$$

where $u = (u_1, \ldots , u_m) \in \{1, \ldots , n\}^m$ are the words of length $|u| = m \geq 1$ and

$$K_u = \psi_{u_1} \circ \cdots \circ \psi_{u_m}(K) = \pi\{(x_k) \,:\, x_1 = u_1, \ldots , x_m = u_m\} \,.$$

The parts may overlap a lot and for any two distinct words u and v of length m, we call K_u and K_v, *ε-relatively close* if

$$K_u = K_v + x \text{ for some } x \text{ with } |x| \leq \varepsilon n^{-m}.$$

The intuitive idea now is that, if too many parts of the set K are ε-relatively close, then K must be small. The following lemma is a consequence of a theorem of Bandt and Graf (1992).

Lemma 3.5 (Bandt-Graf Lemma) *If, for all $\varepsilon > 0$, there exist two distinct words $u \neq v$ of the same length such that K_u and K_v are ε-relatively close, then $\ell(K) = 0$.*

Proof: Suppose that $\ell(K) > 0$. By Lemma 3.4 there exists a nondegenerate interval J such that $\ell(K \cap J) > (0.9)\ell(J)$. Let $\varepsilon = (0.1)\ell(J) > 0$.
By our assumption we find K_u and K_v, which are ε-relatively close. We write

$$K_u = n^{-m} K + b_u \text{ and } K_v = n^{-m} K + b_v \,.$$

Define

$$J_u = \psi_{u_1} \circ \cdots \circ \psi_{u_m}(J) \text{ and } J_v = \psi_{v_1} \circ \cdots \circ \psi_{v_m}(J)\,.$$

Then, simply by scaling,

$$\ell\big(K_u \cap J_u\big) > (0.9)\,\ell\big(J_u\big) \text{ and } \ell\big(K_v \cap J_v\big) > (0.9)\,\ell\big(J_v\big).$$

As K_u and K_v are ε-relatively close, we get that

$$|b_u - b_v| \le (0.1)\ell(J) n^{-m} = (0.1)\ell\big(J_u\big).$$

Hence

$$\ell\big(K_u \cap K_v\big) \ge \ell\big(J_u \cap J_v\big) - (0.1)\Big(\ell(J_u) + \ell(J_v)\Big) \ge (0.6)\ell(J_u) > 0\,,$$

contradicting Lemma 3.2. ■

Having collected the necessary information about self-similar sets on the line we come back to our projection problem. For every $\varepsilon > 0$ define

$$U[\varepsilon] := \big\{\theta \in [0, \pi) \ : \ \text{there exist } u \neq v \text{ such that}$$
$$C(\theta)_u \text{ and } C(\theta)_v \text{ are } \varepsilon\text{-relatively close}\big\}.$$

By the Bandt-Graf Lemma,

$$\text{if } \theta \in \bigcap_{n=1}^{\infty} U[1/n] \text{ then } \ell\big(C(\theta)\big) = 0.$$

It therefore suffices to prove the following lemma.

Lemma 3.6 *For every $\varepsilon > 0$, the set $U[\varepsilon]$ has full Lebesgue measure, i.e. $\ell\big(U[\varepsilon]\big) = \pi$.*

To prove this we need a criterion for a set on the line to have Lebesgue measure zero. The following measure theoretic lemma gives a so-called *porosity condition*.

Lemma 3.7 *Suppose I is an interval and $U \subset I$ a Borel subset such that there exist $a, b > 0$ such that, for every $\theta \in I$, the following condition holds*

$$\text{for all } \delta > 0 \text{ the interval } [\theta - a\delta, \theta + a\delta] \text{ contains an interval}$$
$$\text{of length } \delta b, \text{ which is disjoint from } I \setminus U. \tag{3.2}$$

Then $\ell(I \setminus U) = 0$.

The condition (3.2) says essentially that in every interval the set $I \setminus U$ has a hole of size at least comparable to the length of the interval, i.e. the set $I \setminus U$ is *porous.*

Proof: Divide I into finitely many nonoverlapping compact intervals of length $2a\delta$. From each such interval we can remove an open interval of length δb, which does not contain points from $I \setminus U$. The remainder is a union of compact intervals of total length $|I|(1-b/2a)$. Now apply the same method to each of the remaining intervals, and so forth. After n steps we see that $I \setminus U$ is contained in intervals of total length less than $|I|(1-b/2a)^n$. Letting $n \to \infty$ gives $\ell(I \setminus U) = 0$. ∎

We now come to the **proof of Lemma 3.6**. Without loss of generality we can assume that C is contained in the unit cube $[0,1]^2$, otherwise we can rescale the set and translate it into the unit cube without changing the validity of the statement of the theorem.

We verify the porosity condition (3.2) for the set $U[\varepsilon] \subset [0, \pi)$.

Observe first that it suffices to check (3.2) for δ of the form n^{-m}, m arbitrary. The condition then holds for all δ with a slightly smaller constant b.

Fix $\theta_0 \in (0, \pi)$. By Lemma 3.3 there exist $i \neq j$ such that $C(\theta_0)_i \cap C(\theta_0)_j \neq \emptyset$. In particular, there exist, for every m, words $u = (i, u_2, \ldots, u_m)$ and $v = (j, v_2, \ldots, v_m)$ such that

$$C(\theta_0)_u \cap C(\theta_0)_v \neq \emptyset.$$

Recall that we have denoted $C_v := \phi_{v_1} \circ \cdots \circ \phi_{v_m}(C)$ and that this set can be written as

$$C_v = \left(\frac{1}{n}\right)^m C + a_v \,,$$

for a suitable vector a_v. As $C \subset [0,1]^2$, the set C_v is contained in the cube

$$Q_v := \left(\frac{1}{n}\right)^m [0,1]^2 + a_v \,.$$

Let θ_1 be the angle of the line going through a_u and a_v. Then Q_u and Q_v have the same projection in direction θ_1 and so do C_u and C_v.

As the two cubes Q_u and Q_v have distance ≤ 1 there exists a universal constant $c > 0$ such that, for all $\varepsilon > 0$ and $\theta \in (\theta_1 - cn^{-m}\varepsilon, \theta_1 + cn^{-m}\varepsilon)$, we have $Q_u(\theta) = Q_v(\theta) + x$ for some $|x| \leq \varepsilon n^{-m}$. Hence also $C(\theta)_u = C(\theta)_v + x$ for all θ in this interval and some $|x| \leq \varepsilon n^{-m}$. In other words this interval is contained in $U[\varepsilon]$.

Choose $\delta > 0$ as in the strong separation condition, i.e. such that $\mathrm{dist}(C_k, C_l) > \delta$ for all $1 \leq k \neq l \leq m$. Then, in particular, $\mathrm{dist}(C_u, C_v) > \delta$ and we get, for some constant $d > 0$,

$$|\theta_1 - \theta_0| \leq d \frac{n^{-m}}{d(C_u, C_v)} \leq d \frac{n^{-m}}{d(C_i, C_j)} \leq d \frac{n^{-m}}{\delta} \,.$$

This proves the porosity condition for $a = d/\delta\varepsilon + c\varepsilon$ and $b = 2c\varepsilon$ as above. Hence Lemma 3.6 and Theorem 3.1 are proved.

Example 3.8 Take the unit cube $I_0 = [0,1]^2$ in the plane. Divide I_0 into 16 nonoverlapping compact subcubes and remove all but the four cubes in the corners. Call the resulting set I_1. Proceeding like this we get a sequence $I_1, \ldots, I_n$ of compact sets with I_k consisting of 4^k cubes of sidelength $(1/4)^k$. Theorem 3.1 shows that the intersection $C = \cap I_n$ is invisible. Another way to formulate this is with the help of *Buffon's needle.*

Throw a needle I randomly on $[0,1]^2$. Suppose the length of the needle is at least $\sqrt{8}$, pick the distance r from the origin to the needle uniformly in $[0, \sqrt{2}]$ and locate the centre of the needle at a uniformly chosen point on the circle $\{|z| = r\}$. Formally, let $A(\theta) \subset \mathbb{R}$ be $\mathrm{proj}_\theta\big([0,1]^2\big)$ and define $c = \int_0^\theta \ell(A(\theta))\, d\theta$. Then, for Borel sets $\Theta \subset [0,\pi)$ and $B(\theta) \subset A(\theta)$,

$$\mathbf{P}\Big\{I \text{ has direction } \theta \in \Theta \text{ and } \mathrm{proj}_\theta(I) \in B(\theta)\Big\} = \frac{1}{c} \int_\Theta \ell\big(B(\theta)\big)\, d\theta\,.$$

Such a needle is called *Buffon's needle.* The probability that the random needle hits the set I_n is now

$$F(n) := \frac{1}{c} \int_0^\pi \ell\big(I_n(\theta)\big)\, d\theta\,.$$

The integral on the right is known in geometric measure theory as the *Favard length* of I_n.

The result of Theorem 3.1 shows that the probability that Buffon's needle hits the set C is zero. Equivalently, $\lim_{n\to\infty} F(n) = 0$. It is a difficult unsolved problem to say, even in this example, how fast the convergence is. A recent article of Peres, Simon and Solomyak gives a very large upper bound.

3.1.3 Intersections of self-similar 1-sets and curves

From Theorem 3.1 we can infer an equally striking result about the intersections of the 1-set C with differentiable curves.

Theorem 3.9 *Let C be a self-similar 1-set as above. Then, for every Lipschitz mapping $\gamma : \mathbb{R} \to \mathbb{R}^2$, the intersection of the curve $\gamma(\mathbb{R})$ with C has 1-Hausdorff measure 0.*

Sketch of the argument: If $\mathcal{H}^1(\gamma(\mathbb{R}) \cap C) > 0$ one can select a small interval $I \subset [0,1]$ such that the subsegment $\gamma(I)$ of the curve is almost completely covered by C. Then the projections of $\gamma(I) \cap C$ and $\gamma(I)$ are almost the same and it is clear that $\gamma(I)$ projects on a nondegenerate interval in all but at most one direction. This contradicts the invisibility of C. We omit the further details.

Remarks:

- This property is frequently taken as the definition of a fractal (or irregular set) in geometric measure theory: a 1-set C in $\mathbb{R}^d$ is a *fractal* if and only if the intersection of every Lipschitz curve $\gamma(\mathbb{R})$ with C has vanishing 1-Hausdorff measure. Analogous definitions hold for m-sets, if m is an integer, replacing curves by manifolds.

- A deep result of geometric measure theory, the *Besicovitch projection theorem* states that Theorem 3.9 and 3.1 are equivalent, i.e. a 1-set is invisible from almost all directions if and only if it is a fractal in the above sense.

3.2 Application: The Besicovitch problem

In this section we turn our attention to a question from the beginning of the 20th century, which has a shocking answer. The question is simple:

What is the size of the smallest set in the plane containing a line in every direction?

This problem is frequently referred to as the *Besicovitch problem.* Thinking about planar sets containing a line in every direction, only sets of infinite two-dimensional Lebesgue measure ℓ^2 come to most people's minds. This makes Besicovitch's answer to the above question quite surprising.

Theorem 3.10 (Besicovitch) *There exists a Borel set $B \subset \mathbb{R}^2$ with $\mathcal{L}^2(B) = 0$ containing a line in every direction.*

In this section we will construct such sets. The argument is due to Besicovitch (1964) simplifying his first example from 1928 considerably.

The main ingredient of the proof is the following corollary.

Corollary 3.11 *There is a compact subset $C \subset [0,1]^2$ with $0 < \mathcal{H}^1(C) < \infty$, such that the projection of C onto the x-axis is the whole interval $[0,1]$, but C is invisible from almost all directions.*

Proof: Take the unit cube $I_0 = [0,1]^2$ in the plane. Divide I_0 into 16 nonoverlapping compact subcubes and remove all but the first and third in the top line and the second and fourth in the bottom line. Call the resulting set I_1. Proceeding like this we get a sequence $I_1, \dots, I_n$ of compact sets with I_k consisting of 4^k cubes of sidelength $(1/4)^k$. The intersection $C = \cap I_n$ is a 1-set and Theorem 3.1 shows that it is invisible. It is obvious from the construction that the projection on the x-axis is all of $[0,1]$. ∎

The remainder of the construction is astonishingly easy.

Consider the lines

$$\ell(a,b) = \big\{(x,y) \,:\, y = ax + b\big\}, \text{ for } a, b \in C,$$

and define

$$B' := \bigcup_{(a,b)\in C} \ell(a,b)\,.$$

For $f(a,bx) = (x, ax+b)$ we note that $B' = f(C \times \mathbb{R})$. This implies that B' is the union of countably many compact sets, in particular a Borel set. By assumption, for every $a \in [0,1]$ there is $b \in [0,1]$ with $(a,b) \in C$, hence $\ell(a,b) \subset B'$. Thus B' contains a parallel of every line $y = ax$ with $a \in [0,1]$. Taking B as the union of four suitably rotated copies of B' we obtain a Borel set containing a line in *every* direction. All is left to show is that $\ell^2(B') = 0$.

By Fubini's Theorem it is enough to show that almost every vertical line meets B' in a set of one-dimensional Lebesgue measure zero. Define a modified projection $\pi_t(x,y) = tx + y$. Then

$$\begin{aligned} B' \cap \big\{(t,y) \,:\, y \in \mathbb{R}\big\} &= \bigcup_{(a,b)\in C} \ell(a,b) \cap \big\{(t,y) \,:\, y \in \mathbb{R}\big\} \\ &= \big\{(t, at+b) \,:\, (a,b) \in C\big\} = \{t\} \times \pi_t(C)\,. \end{aligned}$$

We intend to show that

$$\ell\big(\pi_t(C)\big) = 0 \text{ for } \ell\text{-almost all } t \in \mathbb{R}. \tag{3.3}$$

Indeed, from this it follows that $\ell\{y : (t,y) \in B'\} = 0$ and we are done. To show (3.3) we see from elementary trigonometry that, if $\theta = (\theta_1, \theta_2) \in \partial B(0,1)$ is on the sphere and $\theta_2 \neq 0$ and ζ is the angle orthogonal to the line through the origin and θ, then $\mathrm{proj}_\zeta(x,y) = \pi_{\theta_1/\theta_2}(x,y)\theta_2\theta$. Hence (3.3) follows from the fact that $\mathrm{proj}_\zeta(C) = 0$ for almost all $\zeta \in [0,\pi)$. This finishes the proof.

Remark 3.12 *Higher dimensional analogues of Theorem 3.10 are partially open. But there are also some results in the opposite direction, for example Marstrand (1979) has shown that there is no set of Lebesgue measure zero in $\mathbb{R}^3$, which contains a parallel plane for every hyperplane.*

Chapter 4

Random fractals

In many cases of interest the concept of self-similarity arises together with randomness. The idea that a random set can be decomposed into parts, with (up to scaling) the same distribution as the whole set is much more general and useful than the deterministic concept of self-similarity we have discussed so far. However, it is quite difficult to give a satisfactory definition of statistical self-similarity. The best approach in my view is still the contribution of U. Zähle (1988). Here we do not attempt such a general approach, but instead concentrate on a class of random sets with a natural self-similarity, the Galton Watson fractals. These sets are naturally linked to Galton-Watson trees and branching processes and we use the opportunity to explore the relationship and introduce basic ideas about probability on trees. The chapter culminates with a dimension formula for Galton-Watson fractals due to Falconer, Mauldin and Williams. A comprehensive account of the subject of this chapter will appear in the forthcoming book by Lyons and Peres, "Probability on trees and networks".

4.1 Coding fractals by trees

Before entering the area of random fractals we describe a general approach to describe fractals by means of *trees* with weights, so called *capacities* associated to the edges. We investigate how the Hausdorff dimension of the fractal can be derived from the tree and the capacities. In later chapters we make the tree as well as the capacities random, which allows to describe a large class of random fractals.

4.1.1 The b-adic coding

We start with the simplest case, the b-adic coding of subsets of $[0,1]$, which goes back to Furstenberg (1970). Here subsets of $[0,1]$ are coded by subtrees of the b-adic tree.

Definition 4.1 *A tree is a countable set V of* vertices *together with a collection $E \subset V \times V$ of* edges. *For every vertex $v \in V$ the set of* predecessors

$$\{w \in V \,:\, (w,v) \in E\}$$

consists of exactly one element, except for one element $\rho \in V$, which has no predecessor. This element is the root. *For every $v \in V$ there exists a unique sequence*

$$v = v_0, v_1, \ldots, v_{n-1}, v_n = \rho \qquad \text{with } (v_{i+1}, v_i) \in E\,,$$

which after finitely many steps leads to the root. The number n of steps is called the order *of the element $v \in V$. The order of an edge $e = (u, v)$ is the order of its endvertex v. For every vertex $v \in V$ we also assume that the set of* successors *or* children

$$\{w \in V \,:\, (v, w) \in E\}$$

is finite.

To fix some notation, we write $|u|$ for the order of a vertex u and $|e|$ for the order of an edge e. We denote the predecessor of v by $\overline{v}$ and write $u \to v$ or $\overline{v} = u$ for $(u, v) \in E$. Every infinite sequence of edges

$$(\rho, v_0), (v_0, v_1), (v_1, v_2), (v_2, v_3), \ldots \text{ with } (v_i, v_{i+1}) \in E$$

is called a ray. *The set of rays is denoted ∂T.*

Our interest is mainly in infinite trees. We associate infinite trees to nonempty compact subsets of $[0, 1]$.

Definition 4.2 *Suppose $C \subset [0, 1]$ is a nonempty, compact set and $b \geq 2$ an integer. We can consider the system of open b-adic subintervals of $[0, 1]$. More precisely the intervals in*

$$\left(\frac{k}{b^n}, \frac{k+1}{b^n}\right) \text{ for } k \in \{0, \ldots, b^n - 1\}$$

are the b-adic subintervals of order n. We denote by $T_b(C)$ the set of those b-adic intervals whose intersection with C is not empty. This set constitutes the vertices of a tree.

The tree structure is given as follows: Two such intervals or vertices are connected by an edge *if one contains the other one and their orders differ by exactly one. The root of this tree is the only b-adic subintervals of order 0, namely $[0, 1]$. We denote this tree by $T_b(C)$ and call it the b-adic coding of C.*

Example 4.3 *In the example of the ternary Cantor set the 3-adic coding produces a binary tree, which is isomorphic to the code space $\Sigma_2 = \{1, 2\}^{\mathbb{N}}$. Other codings, like 17-adic coding of the Cantor set are more complicated and therefore not particularly suited for the investigation of the Cantor set.*

Our aim is to show that the b-adic coding determines the Hausdorff dimension of the set. Given a tree T we need a notion of size that can be related to the size of the coded set.

Definition 4.4 *Suppose* capacities *are assigned to the edges of a tree T, i.e. there is a mapping $C : E \to [0, \infty)$. A* flow *of strength $c > 0$ through a tree with capacities C is a mapping $\theta : E \to [0, c]$ such that*

- *for the root we have* $\sum_{w\,:\,\overline{w}=\rho}\theta\big(\rho,w\big)=c$, *for every other vertex* $v\neq\rho$ *we have* $\theta\big(\overline{v},v\big)=\sum_{w\,:\,\overline{w}=v}\theta\big(v,w\big)$,
- $\theta(e)\leq C(e)$, *in other words, the flow through the edge* e *is not bigger than the capacity of this edge.*

If T is a tree and $\lambda>0$, we associate capacities of $C(e)=\lambda^{-|e|}$ to the edges of the tree. If λ is too big or the tree is too small it may be that no flow is possible, as the following example shows.

Example 4.5 *Let* T *be the* b*-adic tree, i.e. the tree such that every vertex has exactly* b *successors. If* $\lambda\leq b$*, then there is a flow* θ *through the tree defined by* $\theta(u,v)=b^{-|v|}$*. Indeed, this is a flow of strength* 1*, because*

$$\theta\big(\overline{v},v\big)=b^{-|v|}=bb^{-(|v|+1)}=\sum_{w\,:\,\overline{w}=v}\theta\big(v,w\big).$$

As $\lambda\leq b$ *the flow* b^{-n} *through every edge is smaller than the capacity* λ^{-n} *of the edge.*

If $\lambda>b$*, then there is no such flow. If there is one, let* c *be the strength of the flow. One can see by induction that, for each* n*, one of the* b^n *edges of order* n *must have a flow* $\theta(e)\geq cb^{-n}$*. For sufficiently large* n *this contradicts the condition* $\theta(e)\leq\lambda^{-n}$*.*

In the example of the b-adic tree flow is possible if $\lambda<b$, the number of children of every vertex, but not if $\lambda>b$. This suggests that flows can be used to define a mean number of children per vertex and hence the size of a tree.

Definition 4.6 *If* T *is a tree and* $\lambda>0$*, we associate a capacity of* $C(e)=\lambda^{-|e|}$ *to every edge* $e\in E$*. Define the* branching number *of the tree* T *by*

$$\operatorname{br}T=\sup\Big\{\lambda>0\ :\ \textit{there exists a flow through } T \textit{ with capacities } C(e)=\lambda^{-|e|}\Big\}.$$

The main result of this section relates the branching number of the tree $T_b(C)$ with the Hausdorff dimension of $C\subset[0,1]$.

Theorem 4.7

$$\dim C=\frac{\log\big(\operatorname{br}T_b(C)\big)}{\log b}.$$

The key to this result is a famous result of graph theory, the *max-flow min-cut Theorem* of Ford and Fulkerson (1962). A set Π of edges is called a *cutset* if every ray includes an edge from Π.

Theorem 4.8 (Max-flow min-cut Theorem)

$$\max\Big\{\text{strength}\,(\theta)\ :\ \theta \textit{ a flow with capacities } C\Big\}=\inf\Big\{\sum_{e\in\Pi}C(e)\ :\ \Pi \textit{ a cutset}\Big\}.$$

Proof: The proof is a festival of compactness arguments.

First observe that on the left hand side the infimum is indeed a maximum, because if $\{\theta_n\}$ is a sequence of flows with capacities C, then at every edge we have a bounded sequence $\{\theta_n(e)\}$ and by the diagonal argument we may pass to a subsequence such that $\lim \theta_n(e)$ exists simultaneously for all $e \in E$. This limit is obviously again a flow with capacities C.

Secondly observe that every cutset Π contains a finite subset $\Pi' \subset \Pi$, which is still a cutset. Indeed, if this was not the case, we had for every positive integer j a ray $e_1^j, e_2^j, e_3^j, \ldots$ with $e_i^j \notin \Pi$ for all $i \leq j$. By the diagonal argument we find a sequence j_k and edges e_l of order l such that $e_l^{j_k} = e_l$ for all $k \geq l$. Then $e_1, e_2, \ldots$ is a ray and $e_l \notin \Pi$ for all l, which is a contradiction.

Now let θ be a flow with capacities C and Π an arbitrary cutset. We let A be the set of vertices v such that there is a sequence of edges $e_1, \ldots, e_n \notin \Pi$ with $e_1 = (\rho, v_1)$, $e_n = (v_{n-1}, v)$ and $e_j = (v_{j-1}, v_j)$. By our previous observation this set is finite. Let

$$\phi(v,e) := \begin{cases} 1 & \text{if } e = (v,w) \text{ for some } w \in V\,, \\ -1 & \text{if } e = (w,v) \text{ for some } w \in V\,. \end{cases}$$

Then, using the definition of a flow and finiteness of all sums,

$$\begin{aligned} \text{strength}\,(\theta) &= \sum_{e \in E} \phi(\rho, e)\theta(e) = \sum_{v \in A} \sum_{e \in E} \phi(v,e)\theta(e) \\ &= \sum_{e \in E} \theta(e) \sum_{v \in A} \phi(v,e) \leq \sum_{e \in \Pi} \theta(e) \\ &\leq \sum_{e \in \Pi} C(e)\,. \end{aligned}$$

This proves the first inequality.

For the reverse inequality we restrict attention to finite trees. Let T_n be the tree consisting of all vertices V_n and edges E_n of order $\leq n$ and look at cutsets Π consisting of vertices in E_n. A flow of strength $c > 0$ through the tree T_n with capacities C is a mapping $\theta : E_n \to [0, c]$ such that

- for the root we have $\sum_{w\,:\,\overline{w}=\rho} \theta\big(\rho, w\big) = c$,
- for every vertex $v \neq \rho$ with $|v| < n$ we have $\theta\big(\overline{v}, v\big) = \sum_{w\,:\,\overline{w}=v} \theta\big(v, w\big)$,
- $\theta(e) \leq C(e)$.

We shall show that

$$\begin{aligned} &\max\Big\{\text{strength}\,(\theta) \,:\, \theta \text{ a flow in } T_n \text{ with capacities } C\Big\} \\ &\quad\geq \min\Big\{\sum_{e \in \Pi} C(e) \,:\, \Pi \text{ a cutset in } T_n\Big\}\,. \end{aligned} \tag{4.1}$$

Once we have this, we get a sequence (θ_n) of flows in T_n with capacities C and strength at least $c = \min\Big\{\sum_{e \in \Pi} C(e) \,:\, \Pi \text{ a cutset in } T\Big\}$. By using the diagonal argument once more we can get a subsequence such that the limits of $\theta_n(e)$ exist for every edge, and the result is a flow θ with capacities C and strength at least c, as required.

To prove (4.1) let θ be a flow of maximal strength c with capacities C in T_n and call a sequence $\rho = v_0, v_1, \ldots, v_n$ with $(v_i, v_{i+1}) \in E_n$ an *augmenting sequence* if $\theta(v_i, v_{i+1}) < C(v_i, v_{i+1})$. If there are augmenting sequences, we can construct a flow $\tilde{\theta}$ of strength $> c$ by just increasing the flow through every edge of the augmenting sequence by a sufficiently small $\varepsilon > 0$. As θ was maximal this is a contradiction. Hence there is a minimal cutset Π consisting entirely of edges in E_n with $\theta(e) \geq C(e)$. Let A, as above, be the collection of all edges which are connected to the root by edges not in Π. As before, we have

$$\text{strength}\,(\theta) = \sum_{e \in E} \theta(e) \sum_{v \in A} \phi(v,e) = \sum_{e \in \Pi} \theta(e) \geq \sum_{e \in \Pi} C(e)\,,$$

where in the penultimate step we use minimality. This proves (4.1) and finishes the proof. ■

Proof of Theorem 4.7: The crucial principle behind all this is

a cover of C by b-adic intervals is essentially the same as a cutset of $T_b(C)$.

To see this, recall that the vertices $e \in E$ correspond (in fact, are) exactly those open b-adic intervals, which hit C. The length of the interval e is exactly $b^{-|e|}$. Let D be the set of all b-adic points k/b^n for some integer k, n. Note that D is countable, so that $C \setminus D$ and C have the same Hausdorff dimension. If $x \in C \setminus D$, there exists a decreasing sequence of open, dyadic intervals $I_1, I_2, I_3, \ldots$ such that the order of I_n is n with $\bigcap_n I_n = \{x\}$. This is exactly a ray in $T_b(C)$. If Π is a cutset there must be some $e \in \Pi$ with $e = I_n$, by definition of a cutset. In other words, the intervals in the cutset are a cover of $C \setminus D$ and, for $s = \log\lambda / \log b$ the s-value of this cover is

$$\sum_{e \in \Pi} b^{-s|e|} = \sum_{e \in \Pi} \lambda^{-|e|}\,. \tag{4.2}$$

First suppose that $\lambda < \operatorname{br} T_b(C)$. Then there is a flow through $T_b(C)$ with $C(e) = \lambda^{-|e|}$ with strength$\,(\theta) = c > 0$. By the Ford-Fulkerson Theorem for every cutset Π we have

$$\sum_{e \in \Pi} \lambda^{-|e|} \geq c > 0\,,$$

Note that an arbitrary set $E_i \subset [0,1]$ with $b^{-n-1} \leq |E_i| < b^{-n}$ is contained in at most two open b-adic intervals of order n. Hence, if $E_1, E_2, E_3, \ldots$ is an arbitrary cover of C, we can get a cover of C by open b-adic intervals $I_1, I_2, \ldots$ with s-value

$$\sum_{i=1}^{\infty} |I_i|^s \leq 2b^s \sum_{i=1}^{\infty} |E_i|^s\,.$$

We note that, using (4.2),

$$\sum_{i=1}^{\infty} |E_i|^s \geq \frac{1}{2b^s} \sum_{i=1}^{\infty} |I_i|^s = \frac{1}{2b^s} \sum_{e \in \Pi} \lambda^{-|e|} \geq \frac{c}{2b^s} > 0\,,$$

hence (here the original definition of Hausdorff dimension, based on Hausdorff content is used!)

$$\dim C \geq s = \frac{\log \lambda}{\log b},$$

and letting λ tend to $\operatorname{br} T_b(C)$ gives

$$\dim C \geq \frac{\log \operatorname{br} T_b(C)}{\log b}.$$

For the converse inequality we suppose that $\dim C > s$. Then there exists $c > 0$ such that every cover of C (or, in fact $C \setminus D$) has s-value at least c. Suppose moreover that Π is an arbitrary cutset of $T_b(C)$. Then $\{e \,:\, e \in \Pi\}$ is a cover of $C \setminus D$ and hence $\sum_{e \in \Pi} \lambda^{-|e|} \geq c$ with $s = \log \lambda / \log b$. By the max-flow min-cut Theorem we have

$$\max \left\{ \text{strength}\,(\theta) \,:\, \theta \text{ a flow with capacities } C(e) = \lambda^{-|e|} \right\} \geq c > 0\,,$$

hence a flow through $T_b(C)$ with capacities $C(e) = \lambda^{-|e|}$ exists and we infer that $\operatorname{br} T_b(C) \geq \lambda = b^s$. Hence

$$s \leq \frac{\log \operatorname{br} T_b(C)}{\log b}$$

and the result follows for s tending to $\dim C$. ■

4.1.2 More general coding by trees

There is a more general way than b-adic coding to relate trees and compact sets, which is also very useful in higher dimensions. Denote the interior of a set I by int I. For any infinite tree T we identify a ray $e_1, e_2, e_3, \ldots \in \partial T$ and the corresponding sequence of vertices $v_0, v_1, v_2, \ldots$ with $e_i = (v_{i-1}, v_i)$.

Definition 4.9 *Let T be an infinite tree. To each vertex $v \in T$ we associate a compact set $\emptyset \neq I_v \subset \mathbb{R}^d$ as follows*

$$I_v = \operatorname{cl}\big(\operatorname{int} I_v\big), \tag{4.3}$$

$$(u, v) \in E \Longrightarrow I_v \subset I_u, \tag{4.4}$$

$$\overline{u} = \overline{v} \text{ and } u \neq v \Longrightarrow \operatorname{int} I_u \cap \operatorname{int} I_v = \emptyset, \tag{4.5}$$

$$\text{for all } \xi = (v_0, v_1, v_2, \ldots) \in \partial T \text{ we have } \lim_{n \to \infty} |I_{v_n}| = 0, \tag{4.6}$$

$$c_1 := \inf_{v \neq \rho} \frac{|I_v|}{|I_{\overline{v}}|} > 0, \tag{4.7}$$

$$c_2 := \inf_{v} \frac{\ell(\operatorname{int} I_v)}{|I_v|^d} > 0. \tag{4.8}$$

Then we associate a set $I(T)$ to the tree T with marks $\{I_v \,:\, v \in V\}$, by

$$I(T) := \bigcup_{\xi \in \partial T} \bigcap_{v \in \xi} I_v\,.$$

We call $I(T)$ a tree fractal.

Before giving examples of tree fractals we show the following simple facts about sets associated to trees.

Lemma 4.10

(1) $\lim_{n\to\infty} \max_{|v|=n} |I_v| = 0,$

(2) $I(T) = \bigcap_{n\geq 1} \bigcup_{|v|=n} I_v\,.$

Proof: Suppose the contrary of (1). Then there exists $\varepsilon > 0$ and a sequence of vertices v^k with $|v^k| = j(k)$ and $\lim_{k\to\infty} j(k) = \infty$ such that $|I_{v^k}| \geq \varepsilon$ for all k. Let

$$\rho = v_0^k, \dots, v_{j(k)}^k = v^k \text{ with } (v_{i-1}^k, v_i^k) \in E.$$

By the usual diagonal argument we can find $v_i \in V$ and a sequence n_k with $v_i^{n_k} = v_i$ for all $i \leq k$. Then $\xi = v_0, v_1, \dots$ is a ray with $|I_{v_k}| \geq \varepsilon$ for all k, contradicting (4.6).

If $x \in I(T)$, there exists a ray $\xi = v_0, v_1, v_2, \dots \in \partial T$ with $x \in I_{v_n}$ for every n, hence

$$x \in \bigcap_{n\geq 1} \bigcup_{|v|=n} I_v\,.$$

Conversely, if $x \in \bigcap_{n\geq 1} \bigcap_{|v|=n} I_v$ and $x \in I_v$ for $v \in V$ with $|v| = n$, then we find

$$\rho = v_0^n, v_1^n, \dots, v_n^n = v \text{ with } (v_{i-1}^n, v_i^n) \in E.$$

By the usual diagonal argument we can find $v_i \in V$ and a sequence n_k with $v_i^{n_k} = v_i$ for all $i \leq k$. Then $\xi = v_0, v_1, \dots$ is a ray with $x \in \bigcap_{v\in\xi} I_v$. This proves the second part. ∎

Example 4.11 Suppose that K is the attractor of an iterated function system $(\phi_1, \dots, \phi_n)$ consisting of similitudes with contraction ratios $(r_1, \dots, r_n)$. Suppose the open set condition is satisfied with some bounded open set V. Then K is a tree fractal. Indeed, let T be the n-adic tree, which can naturally be identified with $\bigcup_{m=0}^{\infty}\{1, \dots, n\}^m$. To the vertex $(x_1, \dots, x_m)$ we associate the compact set

$$I[x_1, \dots, x_m] = \mathrm{cl}\big(\phi_{x_1} \circ \dots \circ \phi_{x_m}(V)\big).$$

The conditions (4.3) to (4.8) are easy to check using the relations cl int cl = cl and int cl int = int. To see that $I(T)$ is the attractor of the iterated function system we note that all sets $I[x_1, \dots, x_m]$ are contained in $I[x_1, \dots, x_{m-1}]$ and $I[x_1] \subset \mathrm{cl} V$. Hence $\Phi^n(\mathrm{cl}V)$ is decreasing and, by the contraction principle, in the Hausdorff metric,

$$K = \lim_{n\to\infty} \Phi^n(\mathrm{cl}V) = \bigcap_{n=1}^{\infty} \Phi^n(\mathrm{cl}V).$$

Hence, our statement follows from the second part of the previous lemma.

Example 4.12 If C is an arbitrary compact set $C \subset [0,1]$, then we use the b-adic coding tree $T_b(C)$ and attach the closed b-adic interval corresponding to each vertex. Then it easy to check conditions (4.3) to (4.8) and that C itself is the associated set. In a similar vein, using b-adic cubes, one can see that every compact subset of $[0,1]^d$ can be coded as a tree fractal.

The most interesting examples of tree fractals arise in a random context, we defer this to one of the following sections. We now give a formula for the Hausdorff dimension of tree fractals in terms of the tree and associated capacities.

Theorem 4.13 *If $I(T)$ is the tree fractal associated to the tree $T = (V, E)$ and sets $\{I_v : v \in V\}$. For every $s \geq 0$ define capacities $C_s(e) = |I_v|^s$ if $e = (\overline{v}, v)$. Then*

$$\dim I(T) = \inf \Big\{ s : \inf_{\Pi} \sum_{e \in \Pi} C_s(e) = 0 \Big\} = \sup \Big\{ s : \text{ there is a flow with capacities } C_s \Big\}.$$

Note that the second equality follows from the max-flow min-cut theorem and that the statement includes Theorem 4.7 as a special case.

Proof: The inequality $\leq$ is easy, because every cutset Π creates a covering

$$\Big\{ I_v : (\overline{v}, v) \in \Pi \Big\}$$

with s-value $\sum_{e \in \Pi} C_s(e)$. To argue conversely, we consider an arbitrary cover

$$I(T) \subset \bigcup_{i=1}^{\infty} E_i$$

by sets of diameter $|E_i| \leq |I_\rho|$. Define, for each i,

$$\Pi_i := \big\{ e : |I_v| \leq |E_i| < |I_{\overline{v}}| \text{ for } e = (\overline{v}, v) \big\}.$$

Clearly, Π_i is a cutset and

$$I(T) \cap E_i \subset I(T) \subset \bigcup_{(\overline{v}, v) \in \Pi_i} I_v.$$

For $(\overline{v}, v) \in \Pi_i$ we have

$$\ell\big(\mathrm{int} I_v\big) \geq c_2 |I_v|^d \geq c_2 c_1^d |I_{\overline{v}}|^d > c_2 c_1^d |E_i|^d.$$

Look at the sets

$$V_i = \big\{ e \in \Pi_i : I_v \cap E_i \neq \emptyset, \, e = (\overline{v}, v) \big\}.$$

Each such set has at most $q \leq \ell(B(0,1)) 2^d / (c_2 c_1^d)$ elements. Indeed, pick $x \in E_i$ and let $B = \mathrm{cl}\, B(x, 2|E_i|)$. Then $I_v \subset B$ if $I_v \cap E_i \neq \emptyset$ and

$$\ell(B(0,1))(2|E_i|)^d \geq \ell(B) \geq \sum_{(\overline{v}, v) \in V_i} \ell(\mathrm{int} I_v) \geq q c_2 c_1^d |E_i|^d,$$

from which our claim readily follows.

Now

$$I(T) \subset \bigcup_{i=1}^{\infty} \Big(E_i \cap \bigcup_{(\overline{v},v)\in V_i} I_v \Big) \subset \bigcup_{i=1}^{\infty} \bigcup_{(\overline{v},v)\in V_i} I_v.$$

Then the cutset $\Pi = \bigcup V_i$ satisfies

$$\sum_{(\overline{v},v)\in \Pi} |I_v|^s \le \sum_{i=1}^{\infty} \sum_{(\overline{v},v)\in V_i} |I_v|^s \le q \sum_{i=1}^{\infty} |E_i|^s.$$

Hence, if there is a covering with arbitrarily small s-value, there exists a cutset with arbitrarily small $\sum_{e\in\Pi} C_s(e)$. This implies the statement. ■

4.2 Galton-Watson trees

4.2.1 The definition of Galton-Watson trees

In the previous section we have seen that the Hausdorff dimension of tree fractals depends only on the diameters $|I_v|$ of the sets attached to the vertices. Our aim in the following is to study *random tree fractals* arising by

- taking a random tree as a basis of the tree fractal,
- assigning random sets to each vertex of the tree.

If we do this in a proper way (i.e. such that conditions (4.3) to (4.8) are satisfied) we get a random fractal and a dimension formula, which relates the Hausdorff dimension to flows on the random tree — see Section 4.4. We thus start by studying random trees and flows on random trees.

We consider a natural class of random trees, the *Galton-Watson trees*. Basically they are family trees starting with a single particle, the root. The distribution of the offspring number of the root is given by a sequence $(p_k : k \ge 1)$ with $p_k \ge 0$ and $\sum_{k=0}^{\infty} p_k = 1$. In other words, the root has k children with probability p_k. Each of these children (if there are any) has children with the same offspring distribution and the offspring numbers of the children are independent of each other and of their parent. This goes on forever or until there are no more children. We now give a formal definition.

Definition 4.14 *Suppose a sequence $(p_k : k \ge 1)$ with $p_k \ge 0$ and $\sum_{k=0}^{\infty} p_k = 1$ is given. Let $\big(N(i_1, \dots, i_n) : n \ge 0, (i_1, \dots, i_n) \in \mathbb{N}^n\big)$ be a countable family of iid random variables with distribution given by $\mathbf{P}\{N = k\} = p_k$. The set of vertices of the associated* Galton-Watson tree *is formally given by*

$$V := \Big\{ (i_1, \dots, i_n) \in \mathbb{N}^n \,:\, n \ge 0 \text{ and } 1 \le i_k \le N(i_1, \dots, i_{k-1}) \text{ for all } 1 \le k \le n \Big\}.$$

V includes the empty tuple ρ. For all vertices $v = (i_1, \dots, i_n) \in V$ we call n the generation *of v, we write $|v| = n$ and let $V_n := \{v \in V : |v| = n\}$. The set of edges is formally defined as*

$$E := \Big\{ \big((i_1, \dots, i_n), (i_1, \dots, i_{n+1})\big) \,:\, n \ge 0 \text{ and } (i_1, \dots, i_{n+1}) \in V \Big\}.$$

It is not hard to see that this defines a tree with root ρ. The generation sizes of the Galton-Watson tree are defined by $Z_n := \#V_n$ and satisfy $Z_0 = 1$ and

$$Z_{n+1} := \sum_{v \in V_n} N(v) \text{ for } n \geq 0.$$

The stochastic process $\{Z_n : n \geq 0\}$ is the Galton-Watson process *associated to $(p_k : k \geq 1)$.*

A vital tool in dealing with Galton-Watson trees is the *probability generating function* of the *offspring variable* $N := N(\rho)$. It is defined by

$$f(s) := \mathbb{E}\{s^N\} = \sum_{k=0}^{\infty} p_k s^k, \text{ for every } s \in [0,1],$$

where we agree on putting $0^0 := 1$. The first simple properties of f are collected in the following lemma.

Lemma 4.15 *$f : [0,1] \to [0,1]$ is convex and monotonically increasing. We have $f(0) = \mathsf{P}\{N = 0\}$, $f(1) = 1$ and $f'(1) = \mathbb{E}N$.*

Proof: It is easy to see that f maps $[0,1]$ into $[0,1]$ and is monotonically increasing. Clearly, $f(0) = p_0$ and $f(1) = 1$. To see the convexity we use convexity of the functions $g_k(x) = x^k$,

$$f\big(\lambda x + (1-\lambda)y\big) = \sum_{k=0}^{\infty} g_k\big(\lambda x + (1-\lambda)y\big)\, \mathsf{P}\{N = k\} \tag{4.9}$$

$$\leq \sum_{k=0}^{\infty} \big(\lambda g_k(x) + (1-\lambda) g_k(y)\big)\, \mathsf{P}\{N = k\} = \lambda f(x) + (1-\lambda) f(y) \tag{4.10}$$

for all $x, y \in [0,1]$ and $0 \leq \lambda \leq 1$. f is given as a power series which we differentiate term by term,

$$f'(s) = \sum_{k=1}^{\infty} p_k k s^{k-1}, \text{ for every } s \in (0,1).$$

The left derivative of f always exists (∞ is a possible value) for convex functions because the differential quotient is increasing. It is also continuous from the left and thus $f'(1) = \sum_{k=1}^{\infty} p_k k = \mathbb{E}N$. ■

We are interested in the event $\{\exists n \quad Z_n = 0\}$, which is called *extinction*. The complementary event is called *nonextinction* or *survival*. Nonextinction means that the Galton-Watson tree is infinite.

4.2.2 The survival theorem

Our representations of fractals are using infinite trees, therefore we are interested in offspring distributions where this event has positive probability. These offspring distributions are characterized by the *survival theorem* for Galton-Watson trees. Note that the excluded case $p_1 = 1$ is trivial.

Theorem 4.16 (Survival Theorem) *Suppose the offspring distribution satisfies* $p_1 \neq 1$. *Then*

$$q := \mathsf{P}\big\{ \text{ extinction } \big\} \text{ is the smallest fixed point of } f : [0,1] \to [0,1]\,.$$

We also have

$$\mathsf{P}\big\{ \text{ extinction } \big\} = 1 \text{ if and only if } \mathbb{E}N \leq 1.$$

The proof uses the probability generating function and we first calculate the probability generating function of Z_n.

Lemma 4.17 *For all* $s \in [0,1]$ *we have* $\mathbb{E}\big\{ s^{Z_n} \big\} = f^n(s)$ *where the right hand side is the nth iterate of* f. *In particular,* $\mathsf{P}\{Z_n = 0\} = f^n(0)$.

Proof: Let $\mathfrak{V}_n$ be the collection of all finite subsets of $\mathbb{N}^n$, which is a countable set. We have

$$\begin{aligned}
\mathbb{E}\big\{ s^{Z_n} \big\} &= \sum_{V \in \mathfrak{V}_{n-1}} \mathbb{E}\Big\{ s^{\sum_{v \in V} N(v)} 1_{\{V_{n-1}=V\}} \Big\} \\
&= \sum_{V \in \mathfrak{V}_{n-1}} \mathbb{E}\Big\{ \prod_{v \in V} s^{N(v)} 1_{\{V_{n-1}=V\}} \Big\}.
\end{aligned}$$

Now $\{V_{n-1} = V\}$ depends only on $N(v)$ with $|v| \leq n-2$ and is hence independent of the (again independent) random variables $N(v)$ for $|v| = n-1$. Hence we can continue

$$\begin{aligned}
&= \sum_{V \in \mathfrak{V}_{n-1}} \mathbb{E}\Big\{ \prod_{v \in V} s^{N(v)} 1_{\{V_{n-1}=V\}} \Big\} \\
&= \sum_{V \in \mathfrak{V}_{n-1}} \prod_{v \in V} \mathbb{E}\Big\{ s^{N} \Big\} \mathsf{P}\{V_{n-1} = V\} \\
&= \sum_{V \in \mathfrak{V}_{n-1}} f(s)^{\# V} \mathsf{P}\{V_{n-1} = V\} \\
&= \mathbb{E}\Big\{ f(s)^{Z_{n-1}} \Big\},
\end{aligned}$$

and the result follows by induction starting from $\mathbb{E}\{s^{Z_0}\} = \mathbb{E}\{s^N\} = f(s)$. ■

Lemma 4.18 *The extinction probability is* $q = \lim_{n\to\infty} f^n(0)$.

Proof: Since { extinction } is the increasing union of the events $\{Z_n = 0\}$ we have

$$q = \lim_{n\to\infty} \mathsf{P}\{Z_n = 0\} = \lim_{n\to\infty} f^n(0).$$

■

Lemma 4.19 *Assume $p_1 \neq 1$. Then*

(1) *$f(x) \leq x$ for all $x \in [q, 1]$.*

(2) *$q = 1$ if and only if $f'(1) \leq 1$.*

(3) *Either $q = 1$ is the only fixed point of $f : [0,1] \to [0,1]$ or there are exactly two fixed points $0 \leq q < 1$ and 1.*

Proof: If $f'(1) \leq 1$ the graph of f is above the diagonal and as $p_1 \neq 1$ it intersects the diagonal only in $x = 1$. By convexity, f' is monotonically increasing and hence $f'(s) < 1$ for all $s \in [0, 1)$. This means that f is a contraction and, by the contraction principle and the previous lemma, $q = \lim_{n\to\infty} f^n(0)$ exists and is the only fixed point of f. 1 is a fixed point, hence $q = 1$.

If $f'(1) > 1$ the graph of f is below the diagonal in some interval $(\varepsilon, 1]$ for some $0 \leq \varepsilon < 1$. There must be one (and by convexity only one) point $\overline{q} \in [0, 1)$ where f intersects the diagonal. Hence, we have exactly two fixed points, $\overline{q}$ and 1. $f : [0, \overline{q}] \to [0, \overline{q}]$ is a contraction and the contraction principle again states that $q = \lim_{n\to\infty} f^n(0) = \overline{q}$. ■

Remark: The previous proof is better understood by drawing a picture of the possible cases of f.

Proof of the survival theorem: Assume $p_1 \neq 1$. Then the last lemma says that q is in any case the smallest fixed point of $f : [0,1] \to [0,1]$. Moreover, $q = 1$ if and only if $f'(1) \leq 1$ and $f'(1) = \mathbb{E}N$ by Lemma 4.15. ■

4.2.3 The Galton-Watson zero-one law

We also need a 01-law for Galton-Watson trees. Let A be a set of trees or, equivalently, a property of trees. We say that A is *inherited* if

- every finite tree is in A, and
- if the tree $T \in A$ and v is a vertex of the tree, then the subtree $T(v)$ whose root is v is also in A.

Recall the definition of elementary conditional probabilities. For an event A with $\mathsf{P}(A) > 0$ we have

$$\mathsf{P}(B \,|\, A) = \frac{\mathsf{P}(A \cap B)}{\mathsf{P}(A)} \quad \text{and} \quad \mathbb{E}\{X \mid A\} = \frac{\mathbb{E}\{X 1_A\}}{\mathsf{P}(A)}.$$

Theorem 4.20 (Galton-Watson 01-Law) *For a Galton-Watson tree every inherited property A has either probability q or probability 1. In particular, $\mathsf{P}\big\{T \in A \,\big|\, \text{nonextinction}\,\big\} \in \{0, 1\}$.*

Proof: For a tree T let N be the number of children of the root. If $N = k$ let $T_1, \ldots, T_k$ be the subtrees whose roots are the children of the root of T. Formally, T_i is the subgraph with vertices

$$\Big\{(i, i_2, \ldots, i_n) \in \mathbb{N}^n \,:\, n \geq 0 \text{ and } 1 \leq i_k \leq N(i, 1_2, \ldots, i_{k-1}) \text{ for all } 1 < k \leq n\Big\}.$$

Then

$$\mathbb{P}\{T \in A\} = \sum_{k=0}^{\infty} \mathbf{P}\big\{T \in A \text{ and } N = k\big\} \le \sum_{k=0}^{\infty} \mathbf{P}\big\{T_1 \in A, \dots, T_k \in A \text{ and } N = k\big\},$$

because A is inherited. Because $\{T_1 \in A\}, \dots, \{T_k \in A\}$ and $\{N = k\}$ are independent we can continue with

$$\begin{aligned}\sum_{k=0}^{\infty} \mathbf{P}\big\{T_1 \in A, \dots, T_k \in A \text{ and } N = k\big\} &= \sum_{k=0}^{\infty} \prod_{i=1}^{k} \mathbf{P}\big\{T_i \in A\,\big\}\,\mathbf{P}\big\{N = k\big\} \\ &= \mathbb{E}\Big\{\mathbf{P}\{T \in A\}^N\Big\} = f\big(\mathbf{P}\{T \in A\}\big),\end{aligned}$$

using also that $\{T_i \in A\}$ and $\{T \in A\}$ have the same probability. Hence, altogether $\mathbf{P}\{T \in A\} \le f\big(\mathbf{P}\{T \in A\}\big)$. On the other hand $\mathbf{P}\{T \in A\} \ge q$ since every *finite* tree is in A. By Lemma 4.19 (1) we infer that $f(\mathbf{P}\{T \in A\}) \le \mathbf{P}\{T \in A\}$. Therefore $\mathbf{P}\{T \in A\}$ must be a fixed point and hence is in $\{q, 1\}$. Finally, if $\{T \in A\}$ has probability 1, then

$$\mathbf{P}\big(\{T \in A\} \cap \{\text{ nonextinction }\}\big) = \mathbf{P}\{\text{ nonextinction }\}\,.$$

If $\{T \in A\}$ has probability q,

$$\mathbb{P}\big(\{T \in A\} \cap \{\text{ nonextinction }\}\big) = \mathbf{P}\{T \in A\} - q = 0.$$

This finishes the proof. ■

4.3 Galton-Watson networks

4.3.1 The definition of Galton-Watson networks

We now associate random capacities to the edges of the tree, leading to *Galton-Watson networks*. We ask whether there exists a flow in such a network — after Theorem 4.13 we expect that this question is closely linked to the Hausdorff dimension of random tree fractals.

Definition 4.21 *Let*

$$\Omega := \Big\{(n, a_1, \dots, a_n) \,:\, n \in \mathbb{N}, a_i \in (0,1]\Big\} \subset \mathbb{N} \otimes \bigcup_{n=0}^{\infty} (0,1]^n$$

and P be an arbitrary probability distribution on Ω. A random element $L = (N, A_1, \dots, A_N)$ of Ω is interpreted as a random offspring number N and a weight *A_i associated to the edge e_i leading to the ith child. As in the definition of Galton-Watson trees we can formally construct a tree with weighted edges from an iid collection $\{L(i_1, \dots, i_n) \,:\, n \ge 0, (i_1, \dots, i_n) \in \mathbb{N}^n\}$ of P-distributed random variables. If $e = (v_{n-1}, v_n)$ is an edge in the tree we let $A(e)$ be the associated weight.*

If $v_0, \dots, v_n$ is the unique path from the root $v_0 = \rho$ to v_n with $e_k = (v_{k-1}, v_k) \in E$, then we define the capacity *of the edge $e = e_n$ by*

$$C(e) = \prod_{i=1}^{n} A(e_i).$$

The Galton-Watson tree with these associated capacities is called the Galton-Watson network *generated by P.*

4.3.2 Falconer's theorem on flows in Galton-Watson networks

The most interesting question about Galton-Watson networks is the following:

> *Find a criterion —in terms of P alone— for the existence of a flow on the Galton-Watson network generated by P.*

The problem of survival of a Galton-Watson tree, which we solved in the previous section, is a special case of this: let all $A_i \equiv 1$, then flow is possible if and only if the tree is infinite.

The best possible answer to our question is given by the following theorem of Falconer (1986). Note that the excluded case is trivial.

Theorem 4.22 (Falconer's Theorem) *Suppose P is the distribution of $(N, A_1, \dots, A_N)$ and assume that $\sum_{i=1}^{N} A_i \neq 1$ with positive probability. Let*

$$\gamma := E\Big\{ \sum_{i=1}^{N} A_i \Big\}.$$

(1) *If $\gamma \leq 1$ then almost surely* no *flow is possible.*

(2) *If $\gamma > 1$ then flow is possible almost surely on nonextinction of the tree.*

We give a proof, which is due to Falconer (part (1)) and Lyons and Peres (part (2)). The second part of the proof uses the idea of *percolation*.

Proof of (1): If v is a vertex in the tree let $A(v)$ be the weight of the unique edge ending in v. We equip the subtree with root v with capacities $e \mapsto C_v(e) := C(e)/A(v)$ and let $\theta(v)$ be the maximal strength of a flow in this subtree. Abbreviating $\theta := \theta(\rho)$ we have

$$\theta = \sum_{v : \overline{v} = \rho} \big(A(v) \wedge (A(v)\theta(v)) \big) = \sum_{v : \overline{v} = \rho} A(v) \big(1 \wedge \theta(v)\big). \tag{4.11}$$

Now suppose that $\gamma \leq 1$ and the random variable $L = (N, A_1, \dots, A_N)$ determines the children of the root and weights attached to the edges connecting the root to its children. Using independence, and the fact that θ and $\theta(v)$ have the same distribution for every edge v,

$$\begin{aligned}
\mathbb{E}\{\theta\} &= \sum_{n=0}^{\infty} \mathbb{E}\Big\{\theta 1_{\{N=n\}}\Big\} = \sum_{n=0}^{\infty}\sum_{v=1}^{n} \mathbb{E}\Big\{A(v)(1\wedge\theta(v))1_{\{N=n\}}\Big\} \\
&= \sum_{n=0}^{\infty}\sum_{v=1}^{n} \mathbb{E}\Big\{A(v)1_{\{N=n\}}\Big\}\mathbb{E}\Big\{1\wedge\theta(v)\Big\} = \sum_{n=0}^{\infty}\sum_{v=1}^{n} \mathbb{E}\Big\{A(v)1_{\{N=n\}}\Big\}\mathbb{E}\Big\{1\wedge\theta\Big\} \\
&= \sum_{n=0}^{\infty} \mathbb{E}\Big\{\sum_{v=1}^{N} A(v)1_{\{N=n\}}\Big\}\mathbb{E}\Big\{1\wedge\theta\Big\} = \gamma\mathbb{E}\{1\wedge\theta\} \le \mathbb{E}\{1\wedge\theta\}.
\end{aligned}$$

Hence $\theta \le 1$ almost surely and $\mathbb{P}\{\theta > 0\} > 0$ only if $\gamma = 1$. This already shows that no flow is possible if $\gamma < 1$. In the case $\gamma = 1$ we write, for every nonnegative random variable X,

$$\|X\| := \inf\Big\{a \,:\, \mathbb{P}\{X > a\} = 0\Big\}.$$

By independence, we get from (4.11) that

$$\|\theta\| = \Big\|\sum_{v=1}^{N} A(v)\Big\|\;\|\theta\|.$$

hence, if $\|\theta\| > 0$ we have $\|\sum_{v=1}^{N} A(v)\| = 1$. As $\mathbb{E}\{\sum_{v=1}^{N} A(v)\} = \gamma = 1$ we must have $\sum_{v=1}^{N} A(v) = 1$, which is the excluded case. Hence $\theta = 0$ almost surely, which means that no flow is possible. This finishes the proof of the first part. ■

Proof of (2): We first look at a fixed deterministic tree T with deterministic weights $A(e)$ attached to the edges. We introduce a family of random variables on this tree T as follows. Independently for every edge $e \in E$ we let

$$X(e) = \begin{cases} 1 & \text{with probability } A(e) \\ 0 & \text{with probability } 1 - A(e). \end{cases}$$

The intuition is that an edge e is *open* if $X(e) = 1$ and otherwise closed. We consider the subtree $T^* \subset T$ consisting of all edges which are connected to the root by a path of open edges. Let

$$Q(T) := P\big\{T^* \text{ is infinite }\big\}.$$

For any cutset Π note that $\sum_{e\in\Pi} C(e)$ is the expected number of edges in Π, which are also in T^*. Note that

$$\sum_{e\in\Pi} C(e) = E\Big\{\sum_{e\in\Pi} 1_{\{e\in T^*\}}\Big\} \ge P\Big\{e \in T^* \text{ for some } e \in \Pi\Big\} \ge P\Big\{T^* \text{ is infinite }\Big\}.$$

If $\theta(T)$ is the maximal strength of a flow in T, then the last inequality together with the max-flow min-cut theorem shows that

$$Q(T) > 0 \implies \theta(T) > 0. \tag{4.12}$$

Now we use this result again for a Galton-Watson network T, in this way performing a two-step experiment: first sampling the network T and the reducing it to T^*. As a result of the experiment, T^* is another Galton-Watson tree whose offspring number has the same law as

$$\sum_{\overline{v}=\rho} X(v).$$

Denoting by $e_1, \ldots, e_N$ the edges starting from the root we get for the mean offspring number of T^*,

$$\begin{aligned}
\mathbb{E}\Big\{ \sum_{i=1}^{N} X(e_i) \Big\} &= \mathbb{E}\Big\{ \sum_{n=1}^{\infty} \sum_{i=1}^{n} X(e_i) 1_{\{N=n\}} \Big\} = \sum_{n=1}^{\infty} \sum_{i=1}^{n} \mathbb{E}\Big\{ X(e_i) 1_{\{N=n\}} \Big\} \\
&= \sum_{n=1}^{\infty} \sum_{i=1}^{n} \mathbb{E}\Big\{ A(e_i) 1_{\{N=n\}} \Big\} = \sum_{n=1}^{\infty} \mathbb{E}\Big\{ \sum_{i=1}^{N} A(e_i) 1_{\{N=n\}} \Big\} \\
&= \mathbb{E}\Big\{ \sum_{i=1}^{N} A(e_i) \Big\} = \gamma.
\end{aligned}$$

If $\gamma > 1$, by the survival criterion for Galton-Watson trees, we have

$$0 < q = \mathbf{P}\big\{ T^* \text{ is infinite } \big\} = \mathbb{E}\big\{ Q(T) \big\}.$$

Hence $Q(T) > 0$ with positive probability, and by (4.12) we infer that $\theta(T) > 0$ with positive probability. In other words, $\mathbf{P}\{\theta(T) = 0\} < 1$. As the event $\{\theta(T) = 0\}$ is inherited, we infer from the Galton-Watson 01-law that

$$\mathbf{P}\big\{ \theta(T) = 0 \,|\, \text{ nonextinction } \big\} = 1.$$

This had to be shown. ■

4.4 Galton-Watson fractals

4.4.1 Examples of Galton-Watson fractals

In this section we combine Falconer's Theorem on flows in Galton-Watson networks and the Hausdorff dimension formula of Theorem 4.13 to obtain a formula for a large class of random fractals, the *Galton-Watson fractals.* We discuss several examples.

Definition 4.23 *Let $L = (N, A_1, \ldots, A_N)$ be a random variable with $0 < A_i < 1$ almost surely, which is generating a Galton-Watson network T. Suppose compact sets I_v are randomly assigned to the vertices of the Galton-Watson tree in such a way that the following conditions are satisfied almost surely.*

$$I_v = \mathrm{cl}\big(\mathrm{int} I_v\big), \tag{4.13}$$

$$(u, v) \in E \Longrightarrow I_v \subset I_u, \tag{4.14}$$

$$\overline{u} = \overline{v} \text{ and } u \neq v \Longrightarrow \mathrm{int}\, I_u \cap \mathrm{int}\, I_v = \emptyset, \tag{4.15}$$

$$\inf_v \frac{\ell(\mathrm{int} I_v)}{|I_v|^d} \geq c_2 > 0. \tag{4.16}$$

Moreover, assume that the normalized diameters

$$C\big((\overline{v},v)\big) = \frac{|I_v|}{|I_\rho|}$$

are the capacities of the Galton-Watson network. The the random tree fractal $I(T)$ is called a Galton-Watson fractal *associated to the network T.*

The main result of this chapter is the following dimension formula for Galton-Watson fractals found independently by Falconer (1986) and Mauldin and Williams (1986).

Theorem 4.24 (Dimension formula for Galton-Watson fractals) *Almost surely on non-extinction of the Galton-Watson fractal,*

$$\dim I(T) = \min\Big\{\alpha \,:\, \mathbb{E}\Big\{\sum_{i=1}^{N} A_i^\alpha\Big\} \le 1\Big\}.$$

Before giving the (by now well-prepared) proof, we give a couple of examples.

Example 4.25 Divide $[0,1]$ into three equal parts and keep each independently with probability $p \in (0,1)$. Repeat this with the remaining intervals. Now the Galton-Watson network has a generating random variable $(N, A_1, \dots, A_N)$ with N binomial with parameters $n = 3$ and p and A_i deterministic with $A_i = 1/3$. By the survival theorem, the probability that the resulting random set is nonempty is positive if and only if $p > 1/3$, also

$$\mathbb{E}\Big\{\sum_{i=1}^{N} A_i^\alpha\Big\} = \Big(\frac{1}{3}\Big)^\alpha \mathbb{E}N = \frac{3p}{3^\alpha}.$$

This is ≤ 1 if and only if $\alpha \ge 1 + \frac{\log p}{\log 3}$. Hence almost surely on nonextinction,

$$\dim I(T) = 1 + \frac{\log p}{\log 3}.$$

Picking $p = 2/3$ produces the same dimension as the classical Cantor set.

Example 4.26 We generalize the previous example and deal with *percolation fractals*. Fix $p \in (0,1)$ and a positive integer n. Divide $[0,1]^d$ into n^d equal subcubes. Keep each independently with probability p. Apply the same procedure to the remaining cubes until infinity. Now the Galton-Watson network has a generating random variable $(N, A_1, \dots, A_N)$ with N binomial with parameters n^d and p and A_i deterministic with $A_i = 1/n$. The probability that the resulting random set is nonempty is positive if and only if $p > 1/n^d$. Moreover,

$$\mathbb{E}\Big\{\sum_{i=1}^{N} A_i^\alpha\Big\} = \Big(\frac{1}{n}\Big)^\alpha \mathbb{E}N = \frac{n^d p}{n^\alpha}.$$

This is ≤ 1 if and only if $\alpha \ge d + \frac{\log p}{\log n}$. Hence almost surely on nonextinction,

$$\dim I(T) = d + \frac{\log p}{\log n}.$$

Percolation fractals are sometimes used as models for porous media, like soil.

Example 4.27 Remove from $[0, 1]$ a central portion leaving two intervals of random length $A_1, A_2 \in (0, 1/2)$. We repeat this, independently, for each remaining interval. Now the Galton-Watson network has a generating random variable $(N, A_1, \dots, A_N)$ with $N = 2$ deterministic and we have

$$\dim I(T) = \alpha \text{ for } 1 = \mathbb{E}\{A_1^\alpha + A_2^\alpha\} = 2\mathbb{E}\{A_1^\alpha\}.$$

If A_1 is uniform on $[0, 1/2]$ we have

$$\mathbb{E}\{A_1^\alpha\} = 2\int_0^{1/2} t^\alpha\, dt = \frac{2^{-\alpha}}{\alpha+1},$$

and we get the equation $2^{-\alpha+1} = \alpha + 1$, which can be solved numerically (according to Lyons-Peres with a solution between 0.45 and 0.46).

Example 4.28 Suppose M and N are random integers with $M \geq 2$ and $0 \leq N \leq M^2$. Divide the unit square of $\mathbb{R}^2$ into M^2 equal squares and keep N of them in some random manner. Repeat this procedure for the remaining squares. The generating random variable is $(N, 1/M, \dots, 1/M)$ and the probability q of extinction of the associated network is the smallest solution of $\mathbb{E}\{q^N\} = q$. Almost surely on nonextinction,

$$\dim I(T) = \min\Big\{\alpha : \mathbb{E}\{N/M^\alpha\} \leq 1\Big\}.$$

4.4.2 The dimension formula for Galton-Watson fractals

We study a further example in the next chapter, and give the proof of the dimension formula for Galton-Watson fractals, Theorem 4.24, now.

Proof of Theorem 4.24. Since $I_\rho \supset \bigcup_{|v|=1} I_v$ and the interiors of theses sets are disjoint, we have $\ell(I_\rho) \geq \sum_{|v|=1} \ell(\text{int} I_v)$ and hence

$$\ell(B(0,1))\, |I_\rho|^d \geq \ell(I_\rho) \geq c_2 \sum_{|v|=1} |I_v|^d = c_2 |I_\rho|^d \sum_{|v|=1} A_v^d,$$

which implies

$$\mathbb{E}\Big\{\sum_{|v|=1} A_v^d\Big\} \leq \frac{\ell(B(0,1))}{c_2}.$$

By the Lebesgue dominated convergence theorem, $\alpha \mapsto \mathbb{E}\{\sum_{|v|=1} A_v^\alpha\}$ converges, for $\alpha \to \infty$, to zero. We infer that the set

$$\Big\{\alpha : \mathbb{E}\Big\{\sum_{i=1}^N A_i^\alpha\Big\} \leq 1\Big\}$$

is nonempty. Moreover, by the monotone convergence theorem, the function is continuous from the right on $[0, \infty)$, hence the minimum exists.

Assume first that there is $\varepsilon > 0$ with $\varepsilon \leq A_i \leq 1 - \varepsilon$ almost surely. Then the last two conditions on tree fractals (4.6) and (4.7) are satisfied, indeed we have for every ray $(v_0, v_1, v_2, \dots)$ that

$$\lim_{n\to\infty} |I_{v_n}| \leq |I_\rho| \lim_{n\to\infty} (1-\varepsilon)^n = 0,$$

and also

$$\inf_{v\neq\rho} \frac{|I_v|}{|I_{\overline{v}}|} = \inf_{v\neq\rho} A_v \geq \varepsilon.$$

Hence the dimension formula for tree fractals, Theorem 4.13, states that

$$\dim I(T) = \inf\Big\{\alpha\,:\,\inf_{\Pi}\sum_{e\in\Pi}\big|I_{(\overline{v},v)}\big|^{\alpha}=0\Big\} = \inf\Big\{\alpha\,:\,\inf_{\Pi}\sum_{e\in\Pi}C(e)^{\alpha}=0\Big\}.$$

Now the capacities $C(e)^\alpha$ come from a Galton-Watson network with generating random variable $(N, A_1^\alpha, \dots, A_N^\alpha)$. Falconer's Theorem 4.22 says that

$$\inf_{\Pi}\sum_{e\in\Pi}C(e)^{\alpha}=0 \text{ almost surely, if } \mathbb{E}\Big\{\sum_{i=1}^{N}A_i^\alpha\Big\}<1,$$

but

$$\inf_{\Pi}\sum_{e\in\Pi}C(e)^{\alpha}>0 \text{ almost surely on nonextinction, if } \mathbb{E}\Big\{\sum_{i=1}^{N}A_i^\alpha\Big\}>1,$$

which gives the result.

Now consider the general case, i.e. the random variables A_i may take on values arbitrarily close to 0 or 1 with positive probability. Falconer's Theorem gives directly that

$$\mathbb{E}\Big\{\sum_{i=1}^{N}A_i^\alpha\Big\}<1 \text{ implies } \inf_{\Pi}\sum_{(\overline{v},v)\in\Pi}|I_v|^\alpha=0, \text{ almost surely,}$$

which means there exists a covering of $I(T)$ with arbitrarily small α-value. Hence $\mathcal{H}^\alpha(I(T)) = 0$ almost surely. Note that this argument also implies that, if $\mathbb{E}\{\sum_{i=1}^N A_i^\alpha\} = 1$ and $\sum_{i=1}^N A_i^\alpha \neq 1$ with positive probability, then $\mathcal{H}^\alpha(I(T)) = 0$ almost surely (see the corollary below).

For the other direction consider the Galton-Watson subnetwork T_ε consisting of those vertices connected to the root by edges $e = (\overline{v}, v)$ with $\varepsilon \leq A(e) \leq 1-\varepsilon$. Then $I(T_\varepsilon) \subset I(T)$. From the previous argument we get that

$$\mathbb{E}\Big\{\sum_{n=1}^{N}A_i^\alpha 1_{\{\varepsilon\leq A_i\leq 1-\varepsilon\}}\Big\}>1 \text{ implies } \dim I(T_\varepsilon)\geq\alpha \text{ almost surely on nonextinction of } T_\varepsilon.$$

This in turn implies, of course,

$$\dim I(T)\geq\alpha \text{ almost surely on nonextinction of } T_\varepsilon.$$

Note that $\mathsf{P}\{T_\varepsilon \text{ survives}\} \uparrow \mathsf{P}\{T \text{ survives}\}$ as $\varepsilon \downarrow 0$. This can be proved by a simple, purely analytical argument based on the survival theorem. Also

$$\lim_{\varepsilon\downarrow 0}\mathbb{E}\Big\{\sum_{i=1}^{N}A_i^\alpha 1_{\{\varepsilon\leq A_i\leq 1-\varepsilon\}}\Big\} = \mathbb{E}\Big\{\sum_{i=1}^{N}A_i^\alpha\Big\},$$

by monotone convergence. Hence

$$\mathbb{E}\Big\{\sum_{i=1}^{N}A_i^\alpha\Big\}>1 \text{ implies } \dim I(T)\geq\alpha \text{ almost surely on nonextinction of } T.$$

This implies the full statement of the theorem. ■

The proof of the upper bounds above gives a little more than we need.

Corollary 4.29 *If* $\dim I(T) = \alpha$ *and* $\sum_{i=1}^{N} A_i^\alpha \neq 1$ *with positive probability, then* $\mathcal{H}^\alpha(I(T)) = 0$ *almost surely.*

Remark: Compare the following two random fractals: On the one hand Example 4.25 with $p = 2/3$, on the other hand the random fractal obtained by dividing $[0,1]$ into three nonoverlapping intervals of length $1/3$ and choosing two of the three intervals at random, proceeding like this until infinity.

In both cases we obtain fractals of Hausdorff dimension $\alpha = \log 2/\log 3$. To see this in the second case just observe that the 3-adic coding tree of the fractal is the dyadic tree, exactly as in the case of the ordinary ternary Cantor set. The previous corollary indicates a significant difference between the two examples. Whereas for the first case, by the corollary, the α-Hausdorff measure is zero, one can show that in the second case the α-Hausdorff measure is strictly positive. This can be seen from the fact that there exists a flow on the coding tree with capacities $C_\alpha(\overline{v}, v) = |I_v|^\alpha$ in the second example, whilst there is none in the first.

Chapter 5

Fractal properties of Brownian motion

Brownian motion, the most important of all stochastic processes, gives rise to several random sets, which can be studied as fractals: its graph, range or level sets all have an intricate geometric nature and their deeper fractal properties give important insight in the process itself.

5.1 Brownian motion: basic properties

5.1.1 The definition of Brownian motion

The motivation for studying Brownian motion is that it is the (up to affine transformations) only continuous random function $X : [0, \infty) \to \mathbb{R}^d$ with the following properties,

- for every $h > 0$ the displacements $X(t + h) - X(t)$ have the same distribution at every t,
- the displacements $X(t + h) - X(t)$ are independent of the past $\{X(s) : 0 \leq s \leq t\}$,
- the mean displacement is zero.

We now give a definition of one-dimensional Brownian motion in terms of its explicit distribution and then work out properties of its paths from this definition. The fact that Brownian motion exists (this is not obvious!) and that it is already uniquely determined by the features above (this is a variant of the central limit theorem) is usually shown in lectures about probability theory, see for example my script on the subject.

Definition 5.1 *A family* $\{X(t) : t \geq 0\}$ *of random variables on a probability space* $(\Omega, \mathcal{A}, \mathbf{P})$ *is called a (one-dimensional)* Brownian motion *if*

(1) $X : [0, \infty) \to \mathbb{R}, t \mapsto X(t)$ *is a continuous function,*

(2) *for all* $0 \leq t_1 \leq t_2 \leq \cdots \leq t_m$ *the increments* $X(t_2) - X(t_1), \ldots, X(t_m) - X(t_{m-1})$ *are independent,*

(3) *for every $t \geq 0$ and $h > 0$ the increment $X(t+h) - X(t)$ is normally distributed with mean value 0 and variance h, i.e.*

$$\mathsf{P}\big\{X(t+h) - X(t) \leq x\big\} = \frac{1}{\sqrt{2\pi h}} \int_{-\infty}^{x} \exp\big(\frac{-u^2}{2h}\big)\, du \text{ for all } x \in \mathbb{R}.$$

Brownian motion is usually called a *self-similar process.* This refers to the following property, which indicates a *statistical* self-similarity.

Lemma 5.2 (Scaling invariance) *If $\{X(t) : t \geq 0\}$ is a Brownian motion and $s > 0$, then $\{Y(t) : t \geq 0\}$ with $Y(t) := s^{-1/2} X(ts)$ is also a Brownian motion.*

Proof: The first two properties are clearly fulfilled. The increments $Y(t+h) - Y(t) = h^{-1/2}X(ts+hs) - X(ts)$ are normally distributed with mean 0 and variance $(h^{-1/2})^2 hs = s$, as required. ■

Brownian motion has a second very useful invariance property, the invariance under *time-inversion.* Again there is a transformation on the space of functions, which changes the individual Brownian random functions but leaves their distribution unchanged. The following statement is not hard to prove, but we omit the proof here (see e.g. my script on probability theory).

Lemma 5.3 (Time inversion property) *Suppose $\{X(t) : t \geq 0\}$ is a Brownian motion. Then the process $\{Y(t) : t \geq 0\}$ defined by*

$$Y(t) = \begin{cases} 0, & \text{if } t = 0; \\ tX(1/t), & \text{if } t > 0; \end{cases}$$

is also a Brownian motion.

5.1.2 Lévy's modulus of continuity

Brownian motion is continuous by definition, but we now prove a theorem that makes the degree of continuity of the paths of Brownian motion more precise. Pay attention to the order of the *almost surely* and the *for each* in the following statement and note that a change of this order would give a (correct, but) much weaker statement. The result is the most important ingredient in the calculation of the upper bound for the dimension of the graph and range of Brownian motion.

Theorem 5.4 (Lévy's modulus of continuity) *There exists and absolute constant c and a random variable $\varepsilon > 0$ such that, almost surely, for each $0 < h \leq \varepsilon$ and $0 \leq t \leq t + h \leq 1$,*

$$\Big|X(t+h) - X(t)\Big| \leq \sqrt{ch \log(1/h)}\,.$$

Remark: Lévy has shown that the constant can be chosen as anything bigger than 2, but not 2 itself. Our proof produces a bigger constant, but reading between the lines you can see how it can be improved to give the best value of the constant.

Proof: Fix $c > 2$. We first keep t and $h \le 1/e$ fixed and calculate,

$$\begin{aligned}
\mathsf{P}\Big\{|X(t+h) - X(t)| > \sqrt{ch\log(1/h)}\Big\} &= \frac{2}{\sqrt{2h\pi}} \int_{\sqrt{ch\log(1/h)}}^{\infty} \exp\big(-u^2/2h\big)\,du \\
&= \frac{2}{\sqrt{2\pi}} \int_{\sqrt{c\log(1/h)}}^{\infty} \exp\big(-u^2/2\big)\,du \\
&\le \frac{2}{\sqrt{2c\pi\log(1/h)}} \int_{\sqrt{c\log(1/h)}}^{\infty} u e^{-u^2/2}\,du \\
&= \frac{2}{\sqrt{2c\pi\log(1/h)}}\, e^{-u^2/2}\Big|_{\sqrt{c\log(1/h)}}^{\infty} \\
&= \frac{2h^{c/2}}{\sqrt{2c\pi\log(1/h)}} \le h^{c/2}.
\end{aligned}$$

We now focus for a moment on the special case that $[t, t+h]$ is a dyadic interval $[k/2^j, (k+1)/2^j]$. Then, for all $N \in \mathbb{N}$,

$$\begin{aligned}
&\mathsf{P}\Big\{\big|X\big(\tfrac{k}{2^j}\big) - X\big(\tfrac{k}{2^j} + h\big)\big| > \sqrt{ch\log(1/h)} \text{ for some } j \ge N, 0 \le k < 2^j \text{ and } h = 2^{-j}\Big\} \\
&\qquad \le \sum_{j=N}^{\infty} 2^j 2^{-jc/2} = \frac{1}{1 - 2^{1-(c/2)}} 2^{N(1-c/2)} \longrightarrow 0 \text{ as } N \to \infty.
\end{aligned}$$

Hence, almost surely, there exists a random N with

$$\big|X\big(\tfrac{k}{2^j}\big) - X\big(\tfrac{k}{2^j} + h\big)\big| \le \sqrt{ch\log(1/h)} \text{ for all } j \ge N, 0 \le k < 2^j \text{ and } h = 2^{-j}.$$

Choose $\varepsilon = 2^{-N}$. If $0 < h \le \varepsilon$ and $0 \le t < t+h \le 1$, we can write $[t, t+h]$ up to the endpoints as the union of intervals

$$\Big[\frac{k}{2^j}, \frac{k+1}{2^j}\Big] \text{ with } j > N, 0 \le k \le 2^j - 1,$$

such that for every j at most two intervals in the collection have length 2^{-j}. Picking n with $2^{-n} \le h < 2^{-n+1}$ we have

$$\begin{aligned}
\big|X(t+h) - X(t)\big| &\le \sum \big|X\big(k/2^j\big) - X\big((k+1)/2^j\big)\big| \\
&\le 2\sum_{j=n}^{\infty} \sqrt{c2^{-j}\log(2^j)} \\
&\le \sqrt{2^{-n}\log(2^n)}\Big(2\sqrt{c}\sum_{j=0}^{\infty} 2^{-j/2}\sqrt{\frac{j+n}{n}}\Big).
\end{aligned}$$

The last bracket is bounded by the absolute constant $2\sqrt{c}\sum_{j=0}^{\infty} 2^{-j/2}\sqrt{j+1}$. ■

This theorem is sharp in the sense that the function $\sqrt{h\log(1/h)}$ cannot be replaced by any function decreasing faster as $h \downarrow 0$, see my script on probability. For Hölder-continuity of Brownian motion we have the following consequence.

Corollary 5.5 *For every $\alpha < 1/2$ Brownian motion $X : [0,1] \to \mathbb{R}$ is almost surely α-Hölder continuous.*

Proof: Observe that, for every $\alpha < 1/2$, there is $\varepsilon_1 > 0$ such that $\sqrt{ch\log(1/h)} < h^\alpha$ for all $0 < h < \varepsilon_1$. Let $\varepsilon > 0$ as in the theorem (i.e. random) and $\delta = \varepsilon \wedge \varepsilon_1$. If $0 \le s < t \le 1$, then there exist $s = t_0 < \ldots < t = t_n$ with $n \le 1/\delta$ and $t_k - t_{k-1} < \varepsilon$. Hence by the triangle inequality,

$$|X(t) - X(s)| \le \sum_{k=1}^{n} |X(t_k) - X(t_{k-1})| \le \sum_{k=1}^{n} \sqrt{c(t_k - t_{k-1})\log(1/(t_k - t_{k-1}))} < n(t-s)^\alpha.$$

■

5.2 The dimension of graph and range of Brownian motion

5.2.1 The dimension of the graph of one-dimensional Brownian motion

If $f : [0,1] \to \mathbb{R}$ is a function, then the *graph* of the function is the set

$$G_f := \Big\{ \big(t, f(t)\big) \,:\, t \in [0,1] \Big\} \subset [0,1] \times \mathbb{R}.$$

The first theorem of this section determines the Hausdorff dimension of the graph of a Brownian motion.

Theorem 5.6 *If $\{X(t) : t \ge 0\}$ is a Brownian motion and $X : [0,1] \to \mathbb{R}$ the function given by $t \mapsto X(t)$, then almost surely*

$$\dim G_X = \frac{3}{2}.$$

The proof has two parts:

- The *upper bound* for the dimension is based on the Hölder continuity established in the last section. Using the full statement of Theorem 5.4 instead of the corollary, one could even show that $\mathcal{H}^{1/2}(G_f) = 0$.
- The *lower bound* for the dimension is based on the potential theoretic method.

The upper bound follows immediately from Corollary 5.5 and the following easy lemma.

Lemma 5.7 *If $f : [0,1] \to \mathbb{R}$ is α-Hölder continuous, then* $\dim G_f \le 2 - \alpha$.

Remark: If $f : [0,1] \to \mathbb{R}$ is continuously differentiable, then its derivatives has an upper bound $C > 0$ and by the mean value theorem, we have $|f(x) - f(y)| \le C|x-y|$ and the function is 1-Hölder continuous (i.e. Lipschitz continuous). By our theorem, $\dim G_f \le 1$. On the other hand, there is a surjective Lipschitz mapping $\pi : G_f \to [0,1]$ with $\pi(x,y) = x$, which means that $\dim G_f \ge \dim[0,1] = 1$. We infer that the graph of every continuously differentiable function has Hausdorff dimension 1. Hence our theorem implies that almost surely Brownian motion is not continuously differentiable.

Proof of Lemma 5.7: Let $\beta > 2 - \alpha$. Divide $[0,1]$ into n nonoverlapping subintervals $[(k-1)/n, k/n)$ of length $1/n$. Let m_k be the minimum of f on $[(k-1)/n, k/n)$ and M_k the maximum. By the Hölder continuity we have $M_k - m_k \le C/n^\alpha$, hence $Cn^{1-\alpha} + 1$ vertical squares of sidelength $1/n$ cover the graph over the subinterval and the β-value of the total covering is

$$n\Big(Cn^{1-\alpha} + 1\Big)\sqrt{2}(1/n)^\beta \le \text{const } n^{2-\alpha-\beta},$$

which converges to zero as $n \to \infty$. Hence $\mathcal{H}^\beta(G_f) = 0$. ■

Proof of Theorem 5.6: Applying Lemma 5.7 in Corollary 5.5 we get, almost surely,

$$\dim G_X \le \inf_{0<\alpha<1/2} (2-\alpha) = \frac{3}{2}.$$

For the converse inequality we use the potential theoretic method, and therefore we need a *mass distribution* on the graph G_X. We define such a mass distribution as the image under $t \mapsto (t, X(t))$ of a a uniformly distributed point on $[0,1]$. More formally, let

$$\mu_X(A) = \ell\Big\{t \in [0,1] \,:\, (t, X(t)) \in A\Big\} \text{ for } A \subset \mathbb{R}^2 \text{ Borel.}$$

μ_X is obviously a mass distribution on G_X and we have to show that, for all $1 < \alpha < 3/2$, the α-energy of μ_X is finite. Recall that this implies the desired lower bound.

The α-energy of μ_X is, by definition,

$$\iint |x-y|^{-\alpha}\, d\mu_X(x)\, d\mu_X(y) = \int_0^1 \int_0^1 \big[(X(s) - X(t))^2 + (s-t)^2\big]^{-\alpha/2}\, ds\, dt\,,$$

which is a random variable. To show that it is almost surely finite it is sufficient (but would not be necessary) to show that

$$\mathbb{E}\bigg\{ \int_0^1 \int_0^1 \big[(X(s) - X(t))^2 + (s-t)^2\big]^{-\alpha/2}\, ds\, dt \bigg\} < \infty. \tag{5.1}$$

So it suffices to show that a *single integral* is finite in order to establish a lower bound. This, however, requires some preparation, so fix $t \in [0,1]$ and $h > 0$ and calculate, for $\alpha > 1$,

$$\begin{aligned}
\mathbb{E}\Big\{ \big((X(t+h)-X(t))^2 + h^2\big)^{-\alpha/2} \Big\} &= \frac{1}{\sqrt{2\pi h}} \int_{-\infty}^{\infty} (u^2+h^2)^{-\alpha/2} \exp\big(\frac{-u^2}{2h}\big)\, du \\
&= \frac{1}{\sqrt{2\pi}} \int_0^{\infty} (rh+h^2)^{-\alpha/2} r^{-1/2} \exp\big(\frac{-r}{2}\big)\, dr \\
&\le \frac{1}{\sqrt{2\pi}} \Big[\int_0^h (h^2)^{-\alpha/2} r^{-1/2}\, dr + \int_h^{\infty} (rh)^{-\alpha/2} r^{-1/2}\, dr \Big] \\
&\le \frac{1}{\sqrt{2\pi}} \Big[h^{-\alpha} 2\sqrt{h} + h^{-\alpha/2} \frac{2}{1-\alpha} \big(-h^{(1-\alpha)/2} \big) \Big] \\
&\le C \cdot h^{\frac{1}{2}-\alpha},
\end{aligned}$$

for $C = (2\pi)^{-1/2}\big(2 + \frac{2}{\alpha-1}\big) > 0$. Using this and Fubini's Theorem, we get

$$\begin{aligned}
&\mathbb{E}\Big\{ \int_0^1 \int_0^1 \big[(X(s)-X(t))^2 + (s-t)^2\big]^{-\alpha/2}\, ds\, dt \Big\} \\
&= \int_0^1 \int_0^1 \mathbb{E}\Big\{ \big(|X(s)-X(t)|^2 + (s-t)^2\big)^{-\alpha/2} \Big\}\, ds\, dt \\
&\le C \int_0^1 \int_0^1 |s-t|^{1/2-\alpha}\, ds\, dt < \infty,
\end{aligned}$$

for $1 < \alpha < 3/2$. Hence, almost surely, there exists a mass distribution of finite α-energy on G_X and we infer that $\dim G_X \ge 3/2$. ■

5.2.2 The dimension of the range of higher dimensional Brownian motion

Next we want to study the *image* or *range* of a Brownian motion. This is not particularly interesting for a Brownian motion in dimension one, as the image is just the interval

$$\big[\min_{t\in[0,1]} X(t), \max_{t\in[0,1]} X(t) \big],$$

which has dimension one and is of little geometrical interest. Things start to look different in higher dimensions, though.

Definition 5.8 *If $X_1, \dots, X_d$ are independent one dimensional Brownian motions, then the process*

$$X = \big\{ \big(X_1(t), \dots, X_d(t)\big) \,:\, t \ge 0 \big\}$$

with values in $\mathbb{R}^d$ is called a d-dimensional Brownian motion.

We discuss some properties of the distribution of the increments of d-dimensional Brownian motion. First note that the displacements $X(t+h) - X(t)$ are independent from the past $\{X(s) \,:\, 0 \le s \le t\}$ and have the same distribution for every t. From

$$\mathbf{P}\big\{ X_i(t+h) - X_i(t) \in (a_i, b_i] \big\} = \frac{1}{\sqrt{2\pi h}} \int_{a_i}^{b_i} \exp\Big(\frac{-x_i^2}{2h}\Big)\, dx_i \text{ for all } a_i < b_i,$$

we get for any halfopen rectangle

$$Q = (a_i, b_i] \times \cdots \times (a_d, b_d]$$

that

$$\begin{aligned} \mathbb{P}\{X(t+h) - X(t) \in Q\} &= \prod_{i=1}^{d} \frac{1}{\sqrt{2\pi h}} \int_{a_i}^{b_i} \exp\Big(\frac{-x_i^2}{2h}\Big)\, dx_i \\ &= \Big(\frac{1}{2\pi h}\Big)^{d/2} \int_Q \exp\Big(\frac{-\|x\|^2}{2h}\Big)\, dx\,, \end{aligned}$$

using independence and Fubini's Theorem. As every open set is the union of countably many disjoint halfopen rectangles Q, we have

$$\begin{aligned} \mathbb{P}\{\|X(t+h) - X(t)\| < \rho\} &= \Big(\frac{1}{2\pi h}\Big)^{d/2} \int_{B(0,\rho)} \exp\Big(\frac{-\|x\|^2}{2h}\Big)\, dx \\ &= \gamma(d)\Big(\frac{1}{h}\Big)^{d/2} \int_0^{\rho} \exp\Big(\frac{-r^2}{2h}\Big) r^{d-1}\, dr\,, \end{aligned}$$

for a constant $\gamma(d)$ depending only on d, using spherical coordinates.

Lemma 5.9 *Let $0 < \alpha < 1/2$. There exists a random variable C such that, for all $0 \le s, t \le 1$,*

$$\|X(t) - X(s)\| \le C|t-s|^{\alpha}.$$

Proof: This was proved in the one-dimensional case in Corollary 5.5. If c_i is the (random) α-Hölder constant for X_i, we get

$$\|X(t) - X(s)\|^2 = \sum_{i=1}^{d} |X_i(t) - X_i(s)|^2 \le \Big(\sum_{i=1}^{d} c_i^2\Big)|t-s|^{2\alpha},$$

and the result follows with $C = \sqrt{\sum_{i=1}^d c_i^2}$. ■

Theorem 5.10 *Let $d \ge 2$. For a d-dimensional Brownian motion X we have, almost surely,*

$$\dim\Big\{X(t) \,:\, t \in [0,1]\Big\} = 2\,.$$

Proof: We have seen before that α-Hölder mappings can increase the Hausdorff dimension of a set by no more than a factor of $1/\alpha$. Hence,

$$\dim\Big\{X(t) \,:\, t \in [0,1]\Big\} \le \inf_{0<\alpha<1/2} (1/\alpha) \dim[0,1] = 2\,.$$

For the upper bound we use the potential theoretic method again. We define a mass distribution μ_X on the compact set $I_X = \{X(t) \,:\, t \in [0,1]\}$ as the image measure under $t \mapsto X(t)$ of the uniform measure on $[0,1]$. More precisely, let

$$\mu_X(A) = \ell\Big\{t \in [0,1] \,:\, X(t) \in A\Big\} \text{ for } A \subset \mathbb{R}^d \text{ Borel.}$$

μ_X is obviously a mass distribution on I_X and we have to show that, for all $1 < \alpha < 2$, the α-energy of μ_X is finite. Recall that this implies the desired lower bound.

Again we need a calculation to prepare things.

$$\begin{aligned}
\mathbb{E}\Big\{ \|X(t+h) - X(t)\|^{-\alpha} \Big\} &= \gamma(d)\Big(\frac{1}{h}\Big)^{d/2} \int_0^\infty r^{-\alpha} \exp\Big(\frac{-r^2}{2h}\Big) r^{d-1}\, dr\,, \\
&= \gamma(d) h^{-d/2} \int_0^\infty (sh)^{-\alpha/2+d/2-1/2} \exp\Big(\frac{-s}{2}\Big) \frac{h}{2\sqrt{sh}}\, ds \\
&= \gamma(d) h^{-\alpha/2} \frac{1}{2} \int_0^\infty s^{-\alpha/2+d/2-1} \exp\Big(\frac{-s}{2}\Big)\, ds \\
&= C \cdot h^{-\alpha/2},
\end{aligned}$$

for some constant C independent of h and t. Now the α-energy of the mass distribution μ_X is

$$\iint |x-y|^{-\alpha}\, d\mu_X(x)\, d\mu_X(y) = \int_0^1 \int_0^1 \|X(s) - X(t)\|^{-\alpha}\, ds\, dt\,,$$

which is a random variable. To show that it is almost surely finite it is sufficient to show that

$$\mathbb{E}\Big\{ \int_0^1 \int_0^1 \|X(s) - X(t)\|^{-\alpha}\, ds\, dt \Big\} < \infty. \tag{5.2}$$

This integral can be evaluated using our preparation and Fubini's Theorem.

$$\begin{aligned}
\mathbb{E}\Big\{ \int_0^1 \int_0^1 \|X(s) - X(t)\|^{-\alpha}\, ds\, dt \Big\} &= \int_0^1 \int_0^1 \mathbb{E}\Big\{ \|X(s) - X(t)\|^{-\alpha} \Big\}\, ds\, dt \\
&= \int_0^1 \int_0^1 c|s-t|^{-\alpha/2}\, ds\, dt < \infty\,,
\end{aligned}$$

which proves that $\dim I_X \geq \alpha$ for all $\alpha < 2$. ■

Remark: Again our finer continuity result is the key to the (easy) proof that the range of Brownian motion has 2-Hausdorff measure zero.

5.3 The dimension of the zeroset of Brownian motion

The problem of this section is to calculate the dimension of the *zeroset* of a one-dimensional Brownian motion defined as

$$Z = \big\{ t \in [0,1] \,:\, X(t) = 0 \big\}.$$

This is much more difficult than the problems we were facing in the previous section, mainly because we do not have a natural measure readily defined on Z, as this was the case in the previous two examples. So one way to deal with this problem would be to define a good measure on Z. Such a measure exists (the *local time*) but is quite hard to construct.

Alternative ways are based on fine knowledge about distributional properties of Brownian motion and we are going to sketch such a way here, unfortunately I cannot give full details. The idea of the proof is based on Galton-Watson fractals and is due to Graf, Mauldin and Williams (1988).

5.3.1 Brownian motion and Brownian bridge

A first step is to make the problem more symmetric: as Brownian motion satisfies $X(0) = 0$ we would like also to have $X(1) = 0$. We therefore look at a process called *Brownian bridge*, which is more symmetric than Brownian motion.

Definition 5.11 *Let X be a Brownian motion, then define the* Brownian bridge *as the process*

$$\{B(t) \,:\, t \geq 0\} \text{ with } B(t) := X(t) - tX(1).$$

Some properties of the Brownian bridge are easy to check from this definition. The finite dimensional marginals are given by the following density: For $0 = t_0 \leq t_1 \leq \ldots \leq t_n \leq t_{n+1} = 1$,

$$\mathbf{P}\{B(t_1) = dx_1, \ldots, B(t_n) = dx_n\} = \frac{1}{p_1(0,0)} \prod_{i=1}^{n+1} p_{t_i - t_{i-1}}(x_{i-1}, x_i)\,, \tag{5.3}$$

where $x_0 = x_{n+1} = 0$ and $p_t(x,y) = (2\pi t)^{-1/2} \exp(-(y-x)^2/2t)$. From this density we can see that a Brownian bridge is symmetric in the following sense.

Lemma 5.12 *If $\{B(t) \,:\, 0 \leq t \leq 1\}$ is a Brownian bridge, then so is $\{\tilde{B}(t) \,:\, 0 \leq t \leq 1\}$ with $\tilde{B}(t) = B(1-t)$.*

An intuitive interpretation of (5.3) is the following:

The Brownian bridge is Brownian motion conditioned on $X(1) = 1$.

This becomes at least plausible when we write (5.3) in the physicist's way

$$\mathbf{P}\{B(t_1) = dx_1, \ldots, B(t_n) = dx_n\} = \frac{\mathbf{P}\{X(t_1) = dx_1, \ldots, X(t_{n+1}) = dx_{n+1}\}}{\mathbf{P}\{X(t_{n+1}) = dx_{n+1}\}}\Big|_{x_{n+1}=0},$$

and recall the definition of conditional probabilities. With elementary probability alone we cannot make this statement rigorous, because the event $\{X(1) = 1\}$ has probability zero. However, we do not use this, instead we build on another relationship of Brownian bridge and Brownian motion.

Lemma 5.13 *Let $T = \sup\{t < 1 \,:\, X(t) = 0\}$. Then the process*

$$\{\sqrt{1/T}\, X(tT) \,:\, 0 \leq t \leq 1\}$$

is a Brownian bridge.

Proof: We use a bit more of probability than some people in the audience might know. But if you have an intuitive feeling for the notion of a *stopping time* and are willing to use the *strong Markov property* in the sense that "at stopping times Brownian motion essentially starts from the beginning" you can manage.

We start with a Brownian motion X and we have successfully finished the proof if we find another Brownian motion $\tilde{Z}$ with

$$\sqrt{1/T}\, X(tT) = \tilde{Z}(t) - t\tilde{Z}(1) \text{ for } 0 \le t \le 1. \tag{5.4}$$

To do this first take the time-inverted Brownian motion $\tilde{X}$ defined by

$$\tilde{X}(t) = \begin{cases} 0, & \text{if } t = 0; \\ tX(1/t), & \text{if } t > 0. \end{cases}$$

$\tilde{X}$ is again a Brownian motion and

$$T = \sup\{t < 1 \,:\, X(t) = 0\} = \frac{1}{S} \text{ for } S = \sup\{s > 1 \,:\, \tilde{X}(s) = 0\}.$$

Now S is a stopping time for $\tilde{X}$ (whereas T is *not* a stopping time for X!) and hence the process Y defined by

$$Y(t) = \tilde{X}(S+t) - \tilde{X}(S) = \tilde{X}(S+t)$$

is another Brownian motion. We use the substitution $1/t = 1 + s$ or $s = 1/t - 1$ and calculate

$$\begin{aligned} \sqrt{1/T}\, X(tT) &= \sqrt{1/T}\,(tT)\tilde{X}\big(1/(tT)\big) = \sqrt{1/S}\, t\tilde{X}(S/t) \\ &= \sqrt{1/S}\, t\tilde{X}\big(S + sS\big) = t\sqrt{1/S}\, Y(sS). \end{aligned}$$

Now, because S is independent of Y and by Lemma 5.2 the process R defined by $R(s) = \sqrt{1/S}\, Y(sS)$ is a Brownian motion, and it is obvious that there is yet another Brownian motion Z with $R(s-1) = Z(s) - Z(1)$ for all $s \ge 1$. We call its time-inversion $\tilde{Z}$, and this is, finally, another Brownian motion. We have,

$$\begin{aligned} t\sqrt{1/S}\, Y(sS) &= tR(s) = tR\big(1/t - 1\big) \\ &= t\big(Z(1/t) - Z(1)\big) = \tilde{Z}(t) - t\tilde{Z}(1), \end{aligned}$$

which is (5.4) and we are done. ■

Note that we now know two different ways of constructing a Brownian bridge from a Brownian motion and for a given sample path of Brownian motion the two bridges have different sample paths! We use the lemma to prove the following.

Corollary 5.14 *If B is a Brownian bridge and X a Brownian motion, then*

$$\dim\{t \in [0,1] \,:\, B(t) = 0\} \overset{d}{=} \dim\{t \in [0,1] \,:\, X(t) = 0\}.$$

Proof: By the previous lemma there is a random variable T and a Brownian bridge B on the same probability space as X such that

$$\{t \in [0,1] \,:\, B(t) = 0\} = \{t \in [0,1] \,:\, \sqrt{1/T}\, X(tT) = 0\} = \{s/T \,:\, X(s) = 0\,,\, s \in [0,1]\}.$$

Because $x \mapsto x/T$ and its inverse $x \mapsto xT$ are both Lipschitz, the set on the right has the same Hausdorff dimension as the zero set of Brownian motion. ■

5.3.2 The zeroset of a Brownian bridge as a Galton-Watson fractal

We now study the dimension of the zero set of a Brownian bridge. Note that $\mathbf{P}\{B(1/2) = 0\} = 0$, this is evident from the fact that $B(1/2)$ has a density. We define two random variables

$$T_1 = \sup\{t < 1/2 \,:\, B(t) = 0\} \text{ and } T_2 = \inf\{t > 1/2 \,:\, B(t) = 0\}.$$

By Lemma 5.12 the random variables T_1 and $1 - T_2$ have the same distribution (but they are not independent!) The interval (T_1, T_2) does not contain any zeros, and we remove it from $[0, 1]$, which leaves us with two random intervals $[0, T_1]$ at the left and $[T_2, 1]$ on the right. We quote the following fact, similar to Lemma 5.13.

Theorem 5.15 *Let* $T_1 = \sup\{t < 1/2 \,:\, B(t) = 0\}$. *Then the process*

$$\big\{\sqrt{1/T_1}\, B(tT_1) \,:\, 0 \le t \le 1\big\}$$

is a Brownian bridge, which is independent of $\{B(t) \,:\, t \ge T_1\}$.

Proof: Similar to Lemma 5.13 but perhaps a bit more awkward, hence omitted. ■

Now we can represent the zero set of the Brownian bridge as a Galton-Watson fractal: we start with the interval $[0, 1]$ and remove a centred random interval $[T_1, 1 - T_2]$. To the left of the removed interval, we have an independent Brownian bridge

$$\big\{\sqrt{1/T_1} B(tT_1) \,:\, 0 \le t \le 1\big\}.$$

By the symmetry Lemma 5.12, we also have an independent Brownian bridge

$$\big\{\sqrt{1/(1 - T_2)} B(1 - tT_2) \,:\, 0 \le t \le 1\big\}.$$

to the right of the removed interval. If we apply the same procedure on each of the remaining bridges, we iteratively construct the zero set of the Brownian bridge by removing all gaps. The essence of all this is the following:

Theorem 5.16 *The zero set of a Brownian bridge B is a Galton Watson fractal with generating random variable* $L = (2, T_1, 1 - T_2)$. *Hence* $\dim\{t \in [0, 1] \,:\, B(t) = 0\} = \alpha$, *where α is the unique solution of*

$$\mathbb{E}\{T_1^\alpha + (1 - T_2)^\alpha\} = 1.$$

We can now calculate the dimension by evaluating this expectation for the right α.

Theorem 5.17

$$\mathbb{E}\{\sqrt{T_1} + \sqrt{1 - T_2}\} = 1.$$

Proof: By symmetry of the Brownian bridge, T_1 and $1-T_2$ have the same distribution, hence it suffices, to show that $\mathbb{E}\{\sqrt{1-T_2}\} = 1/2$. We have, using the definition of the Brownian bridge and the time inversion property of Brownian motion,

$$\begin{aligned} T_2 &= \inf\big\{1/2 \le t \le 1 \,:\, B(t) = 0\big\} \\ &= \inf\big\{1/2 \le t \le 1 \,:\, X(t) - tX(1) = 0\big\} \\ &\stackrel{d}{=} \inf\big\{1/2 \le t \le 1 \,:\, tX(1/t) - tX(1) = 0\big\} \\ &= \inf\big\{1/2 \le t \le 1 \,:\, X(1/t) - X(1) = 0\big\} \\ &= 1/\sup\big\{1 \le s \le 2 \,:\, X(s) - X(1) = 0\big\}\,. \end{aligned}$$

Because $\{X(s) - X(1) \,:\, s \ge 1\}$ has the same distribution as $\{X(s-1) \,:\, s \ge 1\}$, we have

$$T_2 \stackrel{d}{=} \frac{1}{1 + \sup\big\{0 \le t \le 1 \,:\, X(t) = 0\big\}}$$

and, in particular,

$$\mathbb{E}\sqrt{1-T_2} = \int_0^1 \sqrt{\frac{x}{1+x}}\, f(x)\, dx\,,$$

where f is the density of the random variable L giving the last visit to zero by a Brownian motion before 1,

$$L = \sup\big\{0 \le t \le 1 \,:\, X(t) = 0\big\}.$$

This density can be calculated using the *reflection principle* of Brownian motion.

Claim: $f(x) = \dfrac{1}{\pi}\dfrac{1}{\sqrt{x(1-x)}}$.

Proof of the claim: Define $T_a := \inf\{t > 0 \,:\, X(t) = a\}$, which is a stopping time. This means, loosely speaking, that at time T_a Brownian motion starts anew and behaves independently of what happened in the past. We get the distribution of T_a from the reflection principle, let $a \ge 0$,

$$\mathbf{P}\{T_a \le t\} = \mathbf{P}\{\sup_{0 \le s \le t} X(s) \ge a\} = 2\mathbf{P}\{X(t) \ge a\} = 2\int_a^\infty \frac{1}{\sqrt{2\pi t}} \exp(-x^2/2t)\, dx\,,$$

change variables $x = (\sqrt{t/s})a$, $dx/ds = -a\sqrt{t}s^{-3/2}/2$,

$$= 2\int_t^0 \frac{1}{\sqrt{2\pi t}} \exp(-a^2/2s) \frac{-\sqrt{t}a}{2s^{3/2}}\, ds = \int_0^t \frac{1}{\sqrt{2\pi s^3}} a \exp(-a^2/2s)\, ds.$$

The latter integrand is hence the density of T_a. We write $\mathbf{P}_x$ for probabilities referring to Brownian motion with start in x. We use the Markov property,

$$\begin{aligned}
\mathbb{P}\{L \le s\} &= \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi s}} \exp(-x^2/2s)\mathbf{P}_x\{T_0 > 1-s\}\, dx \\
&= 2\int_0^{\infty} \frac{1}{\sqrt{2\pi s}} \exp(-x^2/2s) \int_{1-s}^{\infty} \frac{1}{\sqrt{2\pi r^3}} x \exp(-x^2/2r)\, dr\, dx \\
&= \frac{1}{\pi} \int_{1-s}^{\infty} \frac{1}{\sqrt{sr^3}} \int_0^{\infty} x \exp(-x^2(r+s)/2rs)\, dx\, dr \\
&= \frac{1}{\pi} \int_{1-s}^{\infty} \frac{1}{\sqrt{sr^3}} \frac{rs}{r+s}\, dr \\
&= \frac{1}{\pi} \int_{1-s}^{\infty} \Big(\frac{(r+s)^2}{rs}\Big)^{1/2} \frac{s}{(r+s)^2}\, dr, \\
&= \frac{1}{\pi} \int_0^{s} \frac{1}{\sqrt{x(1-x)}}\, dx\,,
\end{aligned}$$

substituting $x = s/(r+s)$ with $dx = -s/(r+s)^2 dr$. This proves the claim.

It remains to calculate one more integral.

$$\mathbb{E}\sqrt{1-T_2} = \frac{1}{\pi} \int_0^1 \sqrt{\frac{x}{1+x}}\, \frac{1}{\sqrt{x(1-x)}}\, dx = \frac{1}{\pi} \int_0^1 \frac{1}{\sqrt{1-x^2}}\, dx = \frac{1}{\pi} \arcsin(1) = 1/2.$$

■

We have now shown the the dimension of the zero set of a Brownian bridge is $1/2$ and can infer the desired result from this.

Theorem 5.18 *If X is a Brownian motion and $t > 0$ we have, almost surely,*

$$\dim\big\{s \in [0,t] \,:\, X(s) = 0\big\} = \frac{1}{2}.$$

Proof: The arguments we have produced so far show this for $t = 1$. We can use the scaling invariance of Lemma 5.2 to get the result for every $t > 0$. Indeed, for every $a > 0$ and $Y(s) := a^{-1/2}X(sa)$,

$$Z[t] = \big\{s \in [0,t] \,:\, X(s) = 0\big\} \overset{d}{=} \big\{s \in [0,t] \,:\, Y(s) = 0\big\} = \big\{s \in [0,t] \,:\, X(sa) = 0\big\} = Z[at]/a.$$

Because $x \mapsto x/a$ and its inverse are both Lipschitz, we have $\dim Z[at]/a = \dim Z[at]$. We infer that

$$\dim Z[t] \overset{d}{=} \dim Z[at],$$

which implies the statement by choosing $a = 1/t$. ■

5.4 Intersections of independent Brownian motions

The range of Brownian motion is *not* a Galton-Watson fractal in any obvious way. However there is a fascinating link between the range of a Brownian motion in $\mathbb{R}^3$ and *percolation fractals*, which was first discovered by Yuval Peres in 1996. Pictures show that percolation fractals and the range of a Brownian motion *look* quite different. The connection is given by the notion of *intersection equivalence*.

Definition 5.19 *Two random sets A and B in $\mathbb{R}^d$ are* intersection equivalent *in the open set $U \subset \mathbb{R}^d$, if there are constants $0 < c < C < \infty$, such that for any Borel set $\Lambda \subset U$,*

$$c\mathbf{P}\{A \cap \Lambda \neq \emptyset\} \leq \mathbf{P}\{B \cap \Lambda \neq \emptyset\} \leq C\mathbf{P}\{A \cap \Lambda \neq \emptyset\}. \tag{5.5}$$

Peres showed that the range of a Brownian motion and certain percolation fractals are actually intersection equivalent. We focus our attention on the case of dimension $d = 3$. For dimension $d \geq 4$ and $d = 2$ similar results are available.

We now recall the definition of a percolation fractal. Suppose a parameter $p \in (0,1)$ is given. We consider the natural tiling of $[0,1]^3$ by 2^3 nonoverlapping closed cubes of sidelength $1/2$. Let $\mathcal{S}_1$ be a random subcollection of these cubes, where each cube has probability p of belonging to $\mathcal{S}_1$ and these events are mutually independent. Note that the cardinality of $\mathcal{S}_1$ is binomial with parameters $n = 2^3$ and p.

In general, if $\mathcal{S}_k$ is a (random) collection of cubes of sidelength 2^{-k} we tile each of them by 2^3 nonoverlapping closed cubes of sidelength $(1/2)^{k+1}$ and include each of them independently with probability p into $\mathcal{S}_{k+1}$. Finally, define

$$Q(p) = \bigcap_{k=1}^{\infty} \bigcup_{Q \in \mathcal{S}_k} Q,$$

the *percolation fractal* with retention parameter p. Note the following:

Lemma 5.20 *If $Q(p)$ and $Q(q)$ are independent percolation fractals, then $Q(p) \cap Q(q)$ is a percolation fractal with retention parameter pq.*

To find the right parameter p for a percolation fractal equivalent to the range of a Brownian motion we compare the number of dyadic subcubes of sidelength 2^{-n} in the unit cube which intersect $Q(p)$ and $X[0,1]$.

Let $N(n)$ be the number of dyadic subcubes of $[0,1]^3$ of sidelength 2^{-n}, which intersect $X[0,1]$. Recall the definition of Minkowski dimension,

$$\dim_M X[0,1] = \lim_{n\to\infty} \frac{\log N(n)}{n \log 2}.$$

Because $X[0,1]$ has Minkowski dimension two, if we start the Brownian motion at a point in $[0,1]^3$, we have

$$\lim_{n\to\infty} \frac{\log N(n)}{n \log 2} = 2\,.$$

If $M(n)$ be the number of dyadic subcubes of sidelength 2^{-n}, which intersect $Q(p)$, then

$$\lim_{n\to\infty} \frac{\log N(n)}{n \log 2} = \dim_M Q(p) = 3 + \frac{\log p}{\log 2}.$$

by Example 4.26. Hence

$$\frac{\log p}{\log 2} = -1 \text{ hence } p = \frac{1}{2}.$$

This means that for large n the sets $Q(1/2)$ and $X[0,1]$ hit a similar number of subcubes. Of course, such a calculation is still pretty far from intersection equivalence. Still, Peres proved the following result.

Theorem 5.21 *Let X be a Brownian motion and $\varepsilon > 0$. Then the range $X[\varepsilon, \infty)$ and $Q(1/2)$ are intersection-equivalent in the unit cube $(0,1)^3$.*

Remarks:

- In order to avoid problems related to the deterministic starting point of Brownian motion, we take the range only after a small waiting time $\varepsilon > 0$.
- The proof, which is too complicated to be included here, does not establish a direct link between percolation fractal and Brownian motion, but uses estimates of the hitting probabilities

$$\mathsf{P}\{Q(p) \cap \Lambda \neq \emptyset\} \text{ and } \mathsf{P}\{X[\varepsilon, \infty) \cap \Lambda \neq \emptyset\}$$

  obtained from *potential theory.*

This link between Brownian motion and percolation fractals can be used to answer the following question:

Do the ranges of n independent Brownian motions have a common point? In other words, do n independent Brownian motions intersect each other?

This problem was originally solved by Dvoretzky, Erdös, Kakutani and Taylor in the 1950s with a completely different approach, in the 1980s Kahane gave a fractal geometric proof. The present proof of Peres uses intersection equivalence reduces the problem to the problem of intersection of independent percolation fractals, which is much easier.

Theorem 5.22 *Suppose that X_1 and X_2 are two independent Brownian motions. Then the sets $X_1(0,\infty) \cap X_2(0,\infty) \neq \emptyset$ with positive probability. If X_3 is a third independent Brownian motion, then $X_1(0,\infty) \cap X_2(0,\infty) \cap X_3(0,\infty) = \emptyset$ almost surely.*

Remarks:

- Using a 01-law one can show that the event $X_1(0,\infty) \cap X_2(0,\infty) \neq \emptyset$ holds even almost surely.

- This result is special to dimension three. In dimension four and bigger we have $X_1(0,\infty)\cap X_2(0,\infty)=\emptyset$ almost surely, whereas in dimension two,

$$X_1(0,\infty)\cap X_2(0,\infty)\cap\ldots\cap X_k(0,\infty)\neq\emptyset$$

almost surely, for all $k\in\mathbb{N}$. Proofs of these statements, which are similar to our proof in the case of dimension three, can be given.

Proof of Theorem 5.22: Fix $\varepsilon>0$. Then $X_1[\varepsilon,\infty)$ is intersection equivalent to $Q(1/2)$ and $X_2[\varepsilon,\infty)$ is independent of $X_1[\varepsilon,\infty)$ and intersection equivalent to $Q(1/2)$. It is not hard to show that this implies that $X_1[\varepsilon,\infty)\cap X_2[\varepsilon,\infty)$ is intersection equivalent to $Q(1/2)\cap Q'(1/2)$, where $Q(1/2)$ and $Q'(1/2)$ are independent percolation fractals with retention parameter $1/2$. Hence

$$\mathbf{P}\big\{X_1[\varepsilon,\infty)\cap X_2[\varepsilon,\infty)\neq\emptyset\big\}\geq c\mathbf{P}\big\{Q(1/2)\cap Q'(1/2)\neq\emptyset\big\}.$$

Now note that $Q(1/2)\cap Q'(1/2)$ is again a percolation fractal, but this time with retention parameter $1/4$. Hence the probability on the right is the survival probability of a Galton-Watson tree with binomial offspring distribution with parameters $n=2^3$ and $p=1/4$. The expectation of such a random variable is $np=2^3/4=2>1$ and hence, the survival probability is positive by the survival theorem. Hence

$$\mathbf{P}\big\{X_1(0,\infty)\cap X_2(0,\infty)\neq\emptyset\big\}\geq\mathbf{P}\big\{X_1[\varepsilon,\infty)\cap X_2[\varepsilon,\infty)\neq\emptyset\big\}>0.$$

This proves the first part of the statement. To see that second part we first note that

$$\mathbf{P}\big\{X_1(0,\infty)\cap X_2(0,\infty)\cap X_3(0,\infty)\neq\emptyset\big\}=\lim_{\varepsilon\downarrow 0}\mathbf{P}\big\{X_1[\varepsilon,\infty)\cap X_2[\varepsilon,\infty)\cap X_3[\varepsilon,\infty)\neq\emptyset\big\}.$$

It thus suffices to show that the latter probability is zero. As above $X_1[\varepsilon,\infty)\cap X_2[\varepsilon,\infty)\cap X_3[\varepsilon,\infty)$ is intersection equivalent to $Q(1/2)\cap Q'(1/2)\cap Q''(1/2)$, where $Q(1/2), Q'(1/2)$ and $Q''(1/2)$ are all independent percolation fractals with retention parameter $1/2$. Their intersection is a percolation fractal with retention parameter $(1/2)^3$ and

$$\mathbf{P}\big\{Q(1/2)\cap Q'(1/2)\cap Q''(1/2)\neq\emptyset\big\}$$

is the survival probability of a Galton Watson tree with binomial offspring distribution for $n=2^3$ and $p=(1/2)^3$. The expected offspring number is then $np=1$ and this implies, by the survival theorem, that the survival probability is 0. ∎

The issue of intersections of Brownian motions and also of random walks is one of the most fascinating areas of probability theory, which continues to pose challenging problems. A good introductory text by a protagonist of the field is Lawler's book *Intersections of random walks.*

Chapter 6

References

Textbooks on fractal geometry and related fields:

Falconer, Kenneth
Fractal geometry: mathematical foundations and applications.
Chichester: John Wiley & Sons. xxii, 288 p. (1997).

Falconer, Kenneth
Techniques in fractal geometry.
Chichester: John Wiley & Sons. xvii, 256 p. (1997).

Lyons, Russell & Peres, Yuval
Probability on trees and networks.
To appear. Preliminary version currently available at
http://php.indiana.edu/ rdlyons/prbtree/prbtree.html

Mattila, Pertti
Geometry of sets and measures in Euclidean spaces. Fractals and rectifiability.
1st paperback ed. Cambridge Studies in Advanced Mathematics. 44.
Cambridge: Cambridge University Press. xii, 343 p. (1999).

Peres, Yuval
Probability on trees: An introductory climb.
In: Bertoin, J. (ed.) et al., Lectures on probability theory and statistics.
Ecole d'été de Probabilités de Saint-Flour XXVII-1997, Saint-Flour, France, July 7-23, 1997.
Berlin: Springer. Lect. Notes Math. 1717, 193-280 (1999).

Rogers, C.A.
Hausdorff measures. With a foreword by K. Falconer.
Paperback ed. Cambridge Mathematical Library.
Cambridge: Cambridge University Press (1999).

Background in probability and ergodic theory:

DUDLEY, RICHARD M.
Real analysis and probability. Wadsworth & Brooks/Cole Mathematics Series. Pacific Grove, CA, 436 p. (1989).

DURRETT, RICHARD
Probability. Theory and examples.
Wadsworth & Brooks/Cole Statistics/Probability Series. Pacific Grove, CA, 453 p. (1991).

LAWLER, GREGORY F.
Intersections of random walks.
Probability and Its Applications. Boston: Birkhaeuser. 230 p. (1996).

MÖRTERS, PETER
Lecture notes on probability.
Universität Kaiserslautern, MI Lecture Notes Volume 6. (2000)

Selected research papers:

FALCONER, KENNETH J.
Random fractals.
Math. Proc. Camb. Philos. Soc. 100, 559-582 (1986).

FALCONER, KENNETH J.
Cut-set sums and tree processes.
Proc. Am. Math. Soc. 101, 337-346 (1987).

GRAF, SIEGFRIED; MAULDIN, R.DANIEL & WILLIAMS, S.C.
The exact Hausdorff dimension in random recursive constructions.
Mem. Am. Math. Soc. 71, No.381, 121 p. (1988).

HOWROYD, JOHN D.
On dimension and on the existence of sets of finite positive Hausdorff measure.
Proc. Lond. Math. Soc., III. Ser. 70, No.3, 581-604 (1995).

LYONS, RUSSELL
Random walks and percolation on trees.
Ann. Probab. 18, No.3, 931-958 (1990).

Mauldin, R.Daniel & Williams, S.C.
Random recursive constructions: asymptotic geometric and topological properties.
Trans. Am. Math. Soc. 295, 325-346 (1986).

Peres, Yuval
Intersection-equivalence of Brownian paths and certain branching processes.
Commun. Math. Phys. 177, No.2, 417-434 (1996).

Peres, Yuval; Simon, Karoly & Solomyak, Boris
Self-similar sets of zero Hausdorff measure and positive packing measure.
Isr. J. Math. 117, 353-379 (2000).

Taylor, S.James; Tricot, Claude
Packing measure, and its evaluation for a Brownian path.
Trans. Am. Math. Soc. 288, 679-699 (1985).

Zähle, Ulrich
Self-similar random measures I: Notion, carrying Hausdorff dimension, and hyperbolic distribution. Probab. Theory Relat. Fields 80, No.1, 79-100 (1988).

Index