Random Walk and the Heat Equation

Gregory F. Lawler

Department of Mathematics, University of Chicago, Chicago, IL 60637

E-mail address: lawler@math.uchicago.edu

Contents

Preface

The basic model for the diffusion of heat is uses the idea that heat spreads randomly in all directions at some rate. The *heat equation* is a deterministic (non-random), partial differential equation derived from this intuition by averaging over the very large number of particles. This equation can and has traditionally been studied as a deterministic equation. While much can be said from this perspective, one also loses much of the intutition that can be obtained by considering the individual random particles.

The idea in these notes is to introduce the heat equation and the closely related notion of harmonic functions from a probabilistic perspective. Our starting point is the random walk which in continuous time and space becomes Brownian motion. We then derive equations to understand the random walk. This follows the modern approach where one tries to use both probabilistic and (deterministic) analytical methods to analyze diffusion.

Besides the random/deterministic dichotomy, another difference in approach comes from choosing between discrete and continuous models. The first chapter of this book starts with discrete random walk and then uses it to define harmonic functions and the heat equations in the discrete set-up. Here one sees that linear functions arise, and the deterministic questions yield problems in linear algebra. In

particular, solutions of the heat equation can be found using diagonalization of symmetric matrices.

The next chapter goes to continuous time and continuous space. We start with the Brownian motion which is the limit of random walk. This is a fascinating object in itself and it takes a little work to show that it exists. We have separated the treatment into Sections 2.1 and 2.6. The idea is that the latter section does not need to be read in order to appreciate the rest of the chapter. The traditional heat equation and Laplace equation are found by considering the Brownian particles. Along the way, it is shown that the matrix diagonalization of the previous chapter turns into a discussion of Fourier series.

The third chapter introduces a fundamental idea in probability, martingales, that is closely related to harmonic functions. The viewpoint here is probabilistic. The final chapter is an introduction to fractal dimension. The goal, which is a bit ambitious, is to determine the fractal dimension of the random Cantor set arising in Chapter 3.

This book is derived from lectures that I gave in the Research Experiences for Undergraduates (REU) program at the University of Chicago. The REU is a summer program in part or in full by about eighty mathematics majors at the university. The students take a number of mini-courses and do a research paper under supervision of graduate students. Many of the students also serve as teaching assistants for one of two other summer programs, one for bright high school students and another designed for elementary and high school teachers. The first two chapters in this book come from my mini-courses in 2007 and 2008, and the last two chapters from my 2009 course.

The intended audience for these lectures was advanced undergraduate mathematics majors who may be considering graduate work in mathematics or a related area. The idea was to present probability and analysis in a more advanced way than found in undergraduate courses. I assume the students have had the equivalent of an advanced calculus (rigorous one variable calculus) course and some exposure to linear algebra. I do not assume that the students have had a course in probability, but I present the basics quickly. I do not assume measure theory, but I introduce many of the important ideas along the

way: Borel-Cantelli lemma, monotone and dominated convergence theorems, Borel measure, conditional expectation. I also try to firm up the students grasp of the advanced calculus along the way.

It is hoped that this book will be interesting to undergraduates, especially those considering graduate studies, as well as to graduate students and faculty whose specialty is not probability or analysis. This book could be used for advanced seminars or for independent reading. There are a number of exercises at the end of each section. They vary in difficult and some of them are at the challenging level that correspond to summer projects for undergraduates at the REU.

I would like to thank Marcelo Alvisio, Laurence Field, and Jacob Perlman for their comments on a draft of this book. The author's research is supported by the National Science Foundation.

Chapter 1

Random Walk and Discrete Heat Equation

1.1. Simple random walk

We consider one of the basic models for random walk, *simple random walk on the integer lattice* $\mathbb{Z}^d$. At each time step, a random walker makes a random move of length one in one of the lattice directions.

1.1.1. One dimension. We start by studying simple random walk on the integers. At each time unit, a walker flips a fair coin and moves one step to the right or one step to the left depending on whether the coin comes up heads or tails. Let S_n denote the position of the walker at time n. If we assume that the walker starts at x, we can write

$$S_n = x + X_1 + \cdots + X_n$$

where X_j equals ± 1 and represents the change in position between time $j-1$ and time j. More precisely, the increments $X_1, X_2, \ldots$ are independent random variables with $\mathbb{P}\{X_j = 1\} = \mathbb{P}\{X_j = -1\} = 1/2$.

Suppose the walker starts at the origin ($x = 0$). Natural questions to ask are:

- On the average, how far is the walker from the starting point?

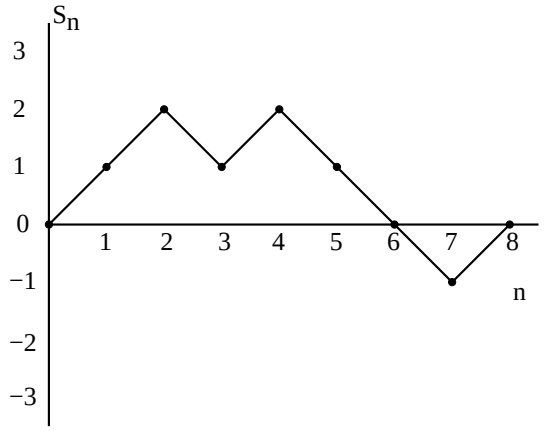

Figure 1. One dimensional random walk with $x = 0$

- What is the probability that at a particular time the walker is at the origin?
- More generally, what is the probability distribution for the position of the walker?
- Does the random walker keep returning to the origin or does the walker eventually leave forever?

Probabilists use the notation $\mathbb{E}$ for *expectation* (also called *expected value, mean, average value*) defined for discrete random variables by

$$\mathbb{E}[X] = \sum_z z\, \mathbb{P}\{X = z\}.$$

The random walk satisfies $\mathbb{E}[S_n] = 0$ since steps of $+1$ and -1 are equally likely. To compute the average *distance*, one might try to

compute $\mathbb{E}\left[|S_n|\right]$. It turns out to be much easier to compute $\mathbb{E}[S_n^2]$,

$$\begin{aligned}\mathbb{E}[S_n^2] &= \mathbb{E}\left[\left(\sum_{j=1}^{n} X_j\right)^2\right] \\ &= \mathbb{E}\left[\sum_{j=1}^{n}\sum_{k=1}^{n} X_j X_k\right] \\ &= \sum_{j=1}^{n}\sum_{k=1}^{n} \mathbb{E}[X_j X_k] = n + \sum_{j\neq k} \mathbb{E}[X_j X_k].\end{aligned}$$

♢ This calculation uses an important property of average values:

$$\mathbb{E}[X+Y] = \mathbb{E}[X] + \mathbb{E}[Y].$$

The fact that the average of the sum is the sum of the averages for random variables even if the random variables are dependent is easy to prove but can be surprising. For example if one looks at the rolls of n regular 6-sided dice, the expected value of the sum is $(7/2)\,n$ whether one takes one die and uses that number n times or rolls n different dice and adds the values. In the first case the sum takes on the six possible values $n, 2n, \ldots, 6n$ with probability $1/6$ each while in the second case the probability distribution for the sum is hard to write down explicitly.

If $j \neq k$, there are four possibilities for the (X_j, X_k); for two of them $X_j X_k = 1$ and for two of them $X_j X_k = -1$. Therefore, $\mathbb{E}[X_j X_k] = 0$ for $j \neq k$ and

$$\mathrm{Var}[S_n] = \mathbb{E}[S_n^2] = n.$$

Here Var denotes the variance of a random variable, defined by

$$\mathrm{Var}[X] = \mathbb{E}\left[(X - \mathbb{E}X)^2\right] = \mathbb{E}[X^2] - (\mathbb{E}X)^2$$

(a simple calculation establishes the second equality). Our calculation illustrates an important fact about variances of sums: if $X_1, \ldots, X_n$ are independent, then

$$\mathrm{Var}[X_1 + \cdots + X_n] = \mathrm{Var}[X_1] + \cdots + \mathrm{Var}[X_n].$$

$\diamondsuit$ The sum rule for expectation and the fact that the cross terms $\mathbb{E}[X_j X_k]$ vanish make it much easier to compute averages of the *square* of a random variable than other powers. In many ways, this is just an analogy of the Pythagorean theorem from geometry: the property $\mathbb{E}[X_j X_k] = 0$, which follows from the fact that the random variables are independent and have mean zero, is the analogue of perpendicularity or orthogonality of vectors.

Finding the probability that the walker is at the origin after n steps is harder than computing $\mathbb{E}[S_n^2]$. However, we can use our computation to give a guess for the size of the probability. Since $\mathbb{E}[S_n^2] = n$, the typical distance away from the origin is of order $\sqrt{n}$. There are about $\sqrt{n}$ integers whose distance is at most $\sqrt{n}$ from the starting point, so one might guess that the probability for being at a particular one should decay like a constant times $n^{-1/2}$. This is indeed the case as we demonstrate by calculating the probability exactly.

It is easy to see that after an odd number of steps the walker is at an odd integer and after an even number of steps the walker is at an even integer. Therefore, $\mathbb{P}\{S_n = x\} = 0$ if $n + x$ is odd. Let us suppose the walker has taken an even number of steps, $2n$. In order for a walker to be back at the origin at time $2n$, the walker must have taken n "+1" steps and n "−1" steps. The number of ways to choose which n steps are $+1$ is $\binom{2n}{n}$ and each particular choice of $2n$ $+1$s and -1s has probability 2^{-2n} of occurring. Therefore,

$$\mathbb{P}\{S_{2n} = 0\} = \binom{2n}{n} 2^{-2n} = \frac{(2n)!}{n!\, n!}\, 2^{-2n}.$$

More generally, if the walker is to be at $2j$, there must be $(n + j)$ steps of $+1$ and $(n - j)$ steps of -1. The probabilities for the number of $+1$ steps are given by the binomial distribution with parameters $2n$ and $1/2$,

$$\mathbb{P}\{S_{2n} = 2j\} = \binom{2n}{n+j} 2^{-2n} = \frac{(2n)!}{(n+j)!\,(n-j)!}\, 2^{-2n}.$$

While these formulas are exact, it is not obvious how to use them because they contain ratios of very large numbers. Trying to understand

the expression on the right hand side leads to studying the behavior of $n!$ as n gets large. This is the goal of the next section.

1.1.2. Stirling's formula. *Stirling's formula* states that as $n \to \infty$,

$$n! \sim \sqrt{2\pi}\, n^{n+\frac{1}{2}}\, e^{-n},$$

where $\sim$ means that the ratio of the two sides tends to 1. We will prove this in the next two subsections. In this subsection we will prove that there is a positive number C_0 such that

$$\lim_{n\to\infty} b_n = C_0, \qquad \text{where} \qquad b_n = \frac{n!}{n^{n+\frac{1}{2}} e^{-n}}, \tag{1.1}$$

and in Section 1.1.3 we show that $C_0 = \sqrt{2\pi}$.

♢ Suppose a_n is a sequence of positive numbers going to infinity and we want to find a positive function $f(n)$ such that $a_n/f(n)$ converges to a positive constant L. Let $b_n = a_n/f(n)$. Then

$$b_n = b_1 \prod_{j=2}^{n} \frac{b_j}{b_{j-1}} = b_1 \prod_{j=2}^{n} [1+\delta_j],$$

where

$$\delta_j = \frac{b_j}{b_{j-1}} - 1,$$

and

$$\lim_{n\to\infty} \log b_n = \log b_1 + \lim_{n\to\infty} \sum_{j=2}^{n} \log[1+\delta_j] = \log b_1 + \sum_{j=2}^{\infty} \log[1+\delta_j],$$

provided that the sum converges. A necessary condition for convergence is that $\delta_n \to 0$. The Taylor's series for the logarithm shows that $|\log[1+\delta_n]| \le c\,|\delta_n|$ for $|\delta_n| \le 1/2$, and hence a sufficient condition for uniform convergence of the sum is that

$$\sum_{n=2}^{\infty} |\delta_n| < \infty.$$

Although this argument proves that the limit exists, it does not determine the value of the limit.

To start, it is easy to check that $b_1 = e$ and if $n \ge 2$,

$$\frac{b_n}{b_{n-1}} = e \left(\frac{n-1}{n}\right)^{n-\frac{1}{2}} = e \left(1-\frac{1}{n}\right)^{n} \left(1-\frac{1}{n}\right)^{-1/2}. \tag{1.2}$$

Let $\delta_n = (b_n/b_{n-1}) - 1$. We will show that $\sum |\delta_n| < \infty$.

♢ One of the most important tools for determining limits is Taylor's theorem with remainder, a version of which we now recall. Suppose f is a C^{k+1} function, i.e., a function with $k+1$ derivatives all of which are continuous functions. Let $P_k(x)$ denote the kth order Taylor series polynomial for f about the origin. Then, for $x > 0$

$$|f(x) - P_k(x)| \le a_k\, x^{k+1},$$

where

$$a_k = \frac{1}{(k+1)!} \max_{0\le t\le x} |f^{(k+1)}(t)|.$$

A similar estimate is derived for negative x by considering $\tilde{f}(x) = f(-x)$. The Taylor series for the logarithm gives

$$\log(1+u) = u - \frac{u^2}{2} + \frac{u^3}{3} - \cdots,$$

which is valid for $|u| < 1$. In fact, the Taylor series with remainder tells us that for every positive integer k

$$\log(1+u) = P_k(u) + O(|u|^{k+1}), \tag{1.3}$$

where $P_k(u) = u - (u^2/2) + \cdots + (-1)^{k+1}(u^k/k)$. The $O(|u|^{k+1})$ denotes a term that is bounded by a constant times $|u|^{k+1}$ for small u. For example, there is a constant c_k such that for all $|u| \le 1/2$,

$$|\log(1+u) - P_k(u)| \le c_k\, |u|^{k+1}. \tag{1.4}$$

We will use the $O(\cdot)$ notation as in (1.3) when doing asymptotics — in all cases this will be shorthand for a more precise statement as in (1.4).

We will show that $\delta_n = O(n^{-2})$, i.e., there is a c such that

$$|\delta_n| \le \frac{c}{n^2}.$$

To see this consider $(1-\frac{1}{n})^n$ which we know approaches e^{-1} as n gets large. We use the Taylor series to estimate how fast it converges. We write

$$\begin{aligned}
\log\left(1-\frac{1}{n}\right)^n &= n\log\left(1-\frac{1}{n}\right) \\
&= n\left[-\frac{1}{n} - \frac{1}{2n^2} + O(n^{-3})\right] \\
&= -1 - \frac{1}{2n} + O(n^{-2}),
\end{aligned}$$

and

$$\log\left(1-\frac{1}{n}\right)^{-1/2} = \frac{1}{2n} + O(n^{-2}).$$

By taking logarithms in (1.2) and adding the terms we finish the proof of (1.1). In fact (see Exercise 1.19) we can show that

$$n! = C_0\, n^{n+\frac{1}{2}}\, e^{-n}\left[1+O(n^{-1})\right]. \tag{1.5}$$

1.1.3. Central limit theorem. We now use Stirling's formula to estimate the probability that the random walker is at a certain position. Let S_n be the position of a simple random walker on the integers assuming $S_0 = 0$. For every integer j, we have already seen that the binomial distribution gives

$$\mathbb{P}\{S_{2n} = 2j\} = \binom{2n}{n+j} 2^{-2n} = \frac{2n!}{(n+j)!(n-j)!}\, 2^{-2n}.$$

Let us assume that $|j| \le n/2$. Then plugging into Stirling's formula and simplifying gives

$$\mathbb{P}\{S_{2n} = 2j\} \sim \tag{1.6}$$

$$\frac{\sqrt{2}}{C_0}\left(1-\frac{j^2}{n^2}\right)^{-n}\left(1+\frac{j}{n}\right)^{-j}\left(1-\frac{j}{n}\right)^{j}\left(\frac{n}{n^2-j^2}\right)^{1/2}.$$

In fact (if one uses (1.5)), there is a c such that the ratio of the two sides is within distance c/n of 1 (we are assuming $|j| \le n/2$).

What does this look like as n tends to infinity? Let us first consider the case $j = 0$. Then we get that

$$\mathbb{P}\{S_{2n} = 0\} \sim \frac{\sqrt{2}}{C_0\, n^{1/2}}.$$

We now consider j of order $\sqrt{n}$. Note that this confirms our previous heuristic argument that the probability should be like a constant times $n^{-1/2}$, since the typical distance is of order $\sqrt{n}$.

Since we expect S_{2n} to be of order $\sqrt{n}$, let us write an integer j as $j = r\sqrt{n}$. Then the right hand side of (1.6) becomes

$$\frac{\sqrt{2}}{C_0\sqrt{n}}\left(1-\frac{r^2}{n}\right)^{-n}\left[\left(1+\frac{r}{\sqrt{n}}\right)^{-\sqrt{n}}\right]^{r}$$

$$\times \left[\left(1-\frac{r}{\sqrt{n}}\right)^{-\sqrt{n}}\right]^{-r} \left(\frac{1}{1-(r^2/n)}\right)^{1/2}.$$

♢ We are about to use the well known limit

$$\left(1+\frac{a}{n}\right)^n \longrightarrow e^a \quad n\to\infty.$$

In fact, using the Taylor's series for the logarithm, we get for $n \geq 2a^2$,

$$\log\left(1+\frac{a}{n}\right)^n = a + O\left(\frac{a^2}{n}\right),$$

which can also be written as

$$\left(1+\frac{a}{n}\right)^n = e^a \left[1+O(a^2/n)\right].$$

As $n\to\infty$, the right hand side of (1.6) is asymptotic to

$$\frac{\sqrt{2}}{C_0\sqrt{n}}\, e^{r^2}\, e^{-r^2}\, e^{-r^2} = \frac{\sqrt{2}}{C_0\sqrt{n}}\, e^{-j^2/n}.$$

For every $a<b$,

$$\lim_{n\to\infty} \mathbb{P}\{a\sqrt{2n} \leq S_{2n} \leq b\sqrt{2n}\} = \lim_{n\to\infty} \sum \frac{\sqrt{2}}{C_0\sqrt{n}}\, e^{-j^2/n}, \tag{1.7}$$

where the sum is over all j with $a\sqrt{2n} \leq 2j \leq b\sqrt{2n}$. The right hand side is the Riemann sum approximation of an integral where the intervals in the sum have length $\sqrt{2/n}$. Hence the limit is

$$\int_a^b \frac{1}{C_0}\, e^{-x^2/2}\, dx.$$

This limiting distribution must be a probability distribution, so we can see that

$$\int_{-\infty}^{\infty} \frac{1}{C_0}\, e^{-x^2/2}\, dx = 1.$$

This gives the value $C_0 = \sqrt{2\pi}$ (see Exercise 1.21), and hence Stirling's formula can be written as

$$n! = \sqrt{2\pi}\, n^{n+\frac{1}{2}}\, e^{-n} \left[1+O(n^{-1})\right].$$

The limit in (1.7) is a statement of the *central limit theorem (CLT)* for the random walk,

$$\lim_{n\to\infty} \mathbb{P}\{a\sqrt{2n} \le S_{2n} \le b\sqrt{2n}\} = \int_a^b \frac{1}{\sqrt{2\pi}}\, e^{-x^2/2}\, dx.$$

1.1.4. Returns to the origin.

♢ Recall that the sum

$$\sum_{n=1}^{\infty} n^{-a}$$

converges if $a > 1$ and diverges otherwise.

We now consider the number of times that the random walker returns to the origin. Let $J_n = 1\{S_n = 0\}$. Here we use the *indicator function* notation: if E is an event, then 1_E or $1(E)$ is the random variable that takes the value 1 if the event occurs and 0 if it does not occur. The total number of visits to the origin by the random walker is

$$V = \sum_{n=0}^{\infty} J_{2n}.$$

Note that

$$\mathbb{E}[V] = \sum_{n=0}^{\infty} \mathbb{E}[J_{2n}] = \sum_{n=0}^{\infty} \mathbb{P}\{S_{2n} = 0\}.$$

We know that $\mathbb{P}\{S_{2n} = 0\} \sim c/\sqrt{n}$ as $n \to \infty$. Therefore,

$$\mathbb{E}[V] = \infty.$$

It is possible, however, for a random variable to be finite yet have an infinite expectation, so we need to do more work to prove that V is actually infinite.

♢ A well known random variable with infinite expectation is that obtained from the St. Petersburg's Paradox. Suppose you play a game where you flip a coin until you get a tails. If you get k heads before flipping the tails, then your payoff is 2^k. The probability that you get exactly k heads is the probability of getting k consecutive heads followed by a tails which is 2^{k+1}. Therefore, the expected payoff in this game is

$$2^0 \cdot \frac{1}{2} + 2^1 \cdot \frac{1}{2^2} + 2^2 \cdot \frac{1}{2^3} + \cdots = \frac{1}{2} + \frac{1}{2} + \frac{1}{2} + \cdots = \infty.$$

Since the expectation is infinite, one should be willing to spend any amount of money in order to play this game once. However, this is clearly not true and here lies the paradox.

Let q be the probability that the random walker ever returns to the origin after time 0. We will show that $q = 1$ by first assuming $q < 1$ and deriving a contradiction. Suppose that $q < 1$. Then we can give the distribution for V. For example, $\mathbb{P}\{V = 1\} = (1 - q)$ since $V = 1$ if and only if the walker never returns after time zero. More generally,

$$\mathbb{P}\{V = k\} = q^{k-1}\,(1 - q), \quad k = 1, 2, \ldots$$

This tells us that

$$\mathbb{E}[V] = \sum_{k=1}^{\infty} k\,\mathbb{P}\{V = k\} = \sum_{k=1}^{\infty} k\,q^{k-1}\,(1 - q) = \frac{1}{1 - q} < \infty.$$

But we know that $\mathbb{E}[V] = \infty$. Hence it must be the case that $q = 1$. We have established the following.

Theorem 1.1. *The probability that a (one-dimensional) simple random walker returns to the origin infinitely often is one.*

Note that this also implies that if the random walker starts at $x \neq 0$, then the probability that it will get to the origin is one.

♢ Another way to compute $\mathbb{E}[V]$ in terms of q is to argue that

$$\mathbb{E}[V] = 1 + q\,\mathbb{E}[V].$$

The 1 represents the first visit; q is the probability of returning to the origin; and the key observation is that the expected number of visits after the first visit given that there is a second visit is exactly the expected number of visits starting at the origin. Solving this simple equation gives $\mathbb{E}[V] = (1 - q)^{-1}$.

1.1.5. Several dimensions. We now consider a random walker on the d-dimensional integer grid

$$\mathbb{Z}^d = \{(x_1, \ldots, x_d) : x_j \text{ integers}\}.$$

At each time step, the random walker chooses one of its $2d$ nearest neighbors, each with probability $1/2d$, and moves to that site. We again let

$$S_n = x + X_1 + \cdots + X_n$$

denote the position of the particle. Here $x, X_1, \ldots, X_n, S_n$ represent points in $\mathbb{Z}^d$, i.e., they are d-dimensional vectors with integer components. The increments $X_1, X_2, \ldots$ are unit vectors with one component of absolute value 1. Note that $X_j \cdot X_j = 1$ and if $j \neq k$, then $X_j \cdot X_k$ equals 1 with probability $1/(2d)$; equals -1 with probability $1/(2d)$; and otherwise equals zero. In particular, $\mathbb{E}[X_j \cdot X_j] = 1$ and $\mathbb{E}[X_j \cdot X_k] = 0$ if $j \neq k$. Suppose $S_0 = 0$. Then $\mathbb{E}[S_n] = 0$, and a calculation as in the one-dimensional case gives

$$\mathbb{E}[|S_n|^2] = \mathbb{E}[S_n \cdot S_n] = \mathbb{E}\left[\left(\sum_{j=1}^{n} X_j\right) \cdot \left(\sum_{j=1}^{n} X_j\right)\right] = n.$$

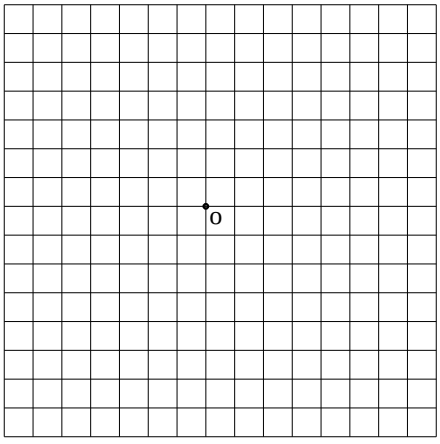

Figure 2. The integer lattice $\mathbb{Z}^2$

What is the probability that we are at the origin after n steps assuming $S_0 = 0$? This is zero if n is odd. If n is even, let us give a heuristic argument. The typical distance from the origin of S_n is of order $\sqrt{n}$. In d dimensions the number of lattice points within distance $\sqrt{n}$ grows like $(\sqrt{n})^d$. Hence the probability that we choose a particular point should decay like a constant times $n^{-d/2}$.

The combinatorics for justifying this is a little more complicated than in the one dimensional case so we will just wave our hands to get the right behavior. In $2n$ steps, we expect that approximately $2n/d$ of them will be taken in each of the d possible directions (e.g., if $d = 2$ we expect about n horizontal and n vertical steps). In order to be at the origin, we need to take an even number of steps in each of the d-directions. The probability of this (Exercise 1.17) is $2^{-(d-1)}$. Given that each of these numbers is even, the probability that each individual component is at the origin is the probability that a one dimensional walk is at the origin at time $2n/d$ (or, more precisely, an even integer very close to $2n/d$). Using this idea we get the asymptotics

$$\mathbb{P}\{S_{2n} = 0\} \sim \frac{c_d}{n^{d/2}}, \qquad c_d = \frac{d^{d/2}}{\pi^{d/2}\, 2^{d-1}}.$$

The particular value of c_d will not be important to us but the fact that the exponent of n is $d/2$ is very important.

Consider the expected number of returns to the origin. If V is the number of visits to the origin, then just as in the $d = 1$ case,

$$\mathbb{E}[V] = \sum_{n=0}^{\infty} \mathbb{P}\{S_{2n} = 0\}.$$

Also,

$$\mathbb{E}[V] = \frac{1}{1-q},$$

where $q = q_d$ is the probability that the d-dimensional walk returns to the origin. Since $\mathbb{P}\{S_{2n} = 0\} \sim c/n^{d/2}$,

$$\mathbb{E}[V] = \sum_{n=0}^{\infty} \mathbb{P}\{S_{2n} = 0\} = \begin{cases} < \infty, & d \geq 3 \\ = \infty, & d = 2 \end{cases}.$$

Theorem 1.2. *Suppose S_n is simple random walk in $\mathbb{Z}^d$ with $S_0 = 0$. If $d = 1, 2$, the random walk is* recurrent*, i.e., with probability one it*

returns to the origin infinitely often. If $d \geq 3$, the random walk is transient*, i.e., with probability one it returns to the origin only finitely often. Also,*

$$\mathbb{P}\{S_n \neq 0 \text{ for all } n > 0\} > 0 \text{ if } d \geq 3.$$

1.1.6. Notes about probability. We have already implicitly used some facts about probability. Let us be more explicit about some of the rules of probability. A *sample space* or *probability space* is a set Ω and *events* are a collection of subsets of Ω including $\emptyset$ and Ω. A probability $\mathbb{P}$ is a function from events to $[0,1]$ satisfying $\mathbb{P}(\Omega) = 1$ and the following countable additivity rule:

- If $E_1, E_2, \ldots$ are disjoint (mutually exclusive) events, then

$$\mathbb{P}\left(\bigcup_{n=1}^{\infty} E_n\right) = \sum_{n=1}^{\infty} \mathbb{P}(E_n).$$

We do not assume that $\mathbb{P}$ is defined for every subset of Ω but we do assume that the collection of events is closed under countable unions and "complementation", i.e., if $E_1, E_2, \ldots$ are events so are $\cup E_j$ and $\Omega \setminus E_j$.

$\diamondsuit$ The assumptions about probability are exactly the assumptions used in measure theory to define a measure. We will not discuss the difficulties involved in proving such a probability exists. In order to do many things in probability rigorously, one needs to use the theory of Lebesgue integration. We will not worry about this in this book.

We do want to discuss one important lemma that probabilists use all the time. It is very easy but it has a name. (It is very common for mathematicians to assign names to lemmas that are used frequently even if they are very simple—this way one can refer to them easily.)

Lemma 1.3 (Borel-Cantelli Lemma)**.** *Suppose $E_1, E_2, \ldots$ is a collection of events such that*

$$\sum_{n=1}^{\infty} \mathbb{P}(E_n) < \infty.$$

Then with probability one at most finitely many of the events occur.

Proof. Let A be the event that infinitely many of $E_1, E_2, \ldots$ occur. For each integer N, $A \subset A_N$ where A_N is the event that at least one of the events $E_N, E_{N+1}, \ldots$ occurs. Then,

$$\mathbb{P}(A) \leq \mathbb{P}(A_N) = \mathbb{P}\left(\bigcup_{n=N}^{\infty} E_n\right) \leq \sum_{n=N}^{\infty} \mathbb{P}(E_n).$$

But $\sum \mathbb{P}(E_n) < \infty$ implies

$$\lim_{N\to\infty} \sum_{n=N}^{\infty} \mathbb{P}(E_n) = 0.$$

Hence $\mathbb{P}(A) = 0$. □

As an example, consider the simple random walk in $\mathbb{Z}^d, d \geq 3$ and let E_n be the event that $S_n = 0$. Then, the estimates of the previous section show that

$$\sum_{n=1}^{\infty} \mathbb{P}(E_n) < \infty,$$

and hence with probability one, only finitely many of the events E_n occur. This says that with probability one, the random walk visits the origin only finitely often.

1.2. Boundary value problems

1.2.1. One dimension: gambler's ruin. Suppose N is a positive integer and a random walker starts at $x \in \{0, 1, \ldots, N\}$. Let S_n denote the position of the walker at time n. Suppose the walker stops when the walker reaches 0 or N. To be more precise, let

$$T = \min\{n : S_n = 0 \text{ or } N\}.$$

Then the position of the walker at time n is given by $\hat{S}_n = S_{n\wedge T}$ where $n \wedge T$ means the minimum of n and T. It is not hard to see that with probability one $T < \infty$, i.e., eventually the walker will reach 0 or N and then stop. Our goal is to try to figure out which point it stops at. Define the function $F : \{0, \ldots, N\} \to [0, 1]$ by

$$F(x) = \mathbb{P}\{S_T = N \mid S_0 = x\}.$$

♢ Recall that if V_1, V_2 are events, then $\mathbb{P}(V_1 \mid V_2)$ denotes the conditional probability of V_1 given V_2. It is defined by

$$\mathbb{P}(V_1 \mid V_2) = \frac{\mathbb{P}(V_1 \cap V_2)}{\mathbb{P}(V_2)},$$

assuming $\mathbb{P}(V_2) > 0$.

We can give a gambling interpretation to this by viewing S_n as the number of chips currently held by a gambler who is playing a fair game where at each time duration the player wins or loses one chip. The gambler starts with x chips and plays until he or she has N chips or has gone bankrupt. The chance that the gambler does not go bankrupt before attaining N is $F(x)$. Clearly, $F(0) = 0$ and $F(N) = 1$. Suppose $0 < x < N$. After the first game, the gambler has either $x - 1$ or $x + 1$ chips, and each of these outcomes is equally likely. Therefore,

$$F(x) = \frac{1}{2}\,F(x+1) + \frac{1}{2}\,F(x-1), \quad x = 1, \ldots, N-1. \tag{1.8}$$

One function F that satisfies (1.8) with the boundary conditions $F(0) = 0, F(N) = 1$ is the linear function $F(x) = x/N$. In fact, this is the only solution as we show now see.

Theorem 1.4. *Suppose a, b are real numbers and N is a positive integer. Then the only function $F : \{0, \ldots, N\} \to \mathbb{R}$ satisfying* (1.8) *with $F(0) = a$ and $F(N) = b$ is the linear function*

$$F_0(x) = a + \frac{x(b-a)}{N}.$$

This is a fairly easy theorem to prove. In fact, we will give several proofs. This is not just to show off how many proofs we can give! It is often useful to give different proofs to the same theorem because it gives us a number of different approaches to trying to prove generalizations. It is immediate that F_0 satisfies the conditions; the real question is one of uniqueness. We must show that F_0 is the *only* such function.

Proof 1. Consider the set $\mathcal{V}$ of all functions $F : \{0, \ldots, N\} \to \mathbb{R}$ that satisfy (1.8). It is easy to check that $\mathcal{V}$ is a vector space, i.e., if

$f, g \in \mathcal{V}$ and c_1, c_2 are real numbers, then $c_1 f + c_2 g \in \mathcal{V}$. In fact, we claim that this vector space has dimension two. To see this, we will give a basis. Let f_1 be the function defined by $f_1(0) = 0, f_1(1) = 1$ and then extended in the unique way to satisfy (1.8). In other words, we define $f_1(x)$ for $x > 1$ by

$$f_1(x) = 2f_1(x-1) - f_1(x-2).$$

It is easy to see that f_1 is the only solution to (1.8) satisfying $f_1(0) = 0, f_1(1) = 1$. We define f_2 similarly with initial conditions $f_2(0) = 1, f_2(1) = 0$. Then $c_1 f_1 + c_2 f_2$ is the unique solution to (1.8) satisfying $f_1(0) = c_2, f_1(1) = c_1$. The set of functions of the form F_0 as a, b vary form a two dimensional subspace of $\mathcal{V}$ and hence must be all of $\mathcal{V}$.

$\diamondsuit$ The set of all functions $f : \{0, \ldots, N\} \to \mathbb{R}$ is essentially the same as $\mathbb{R}^{N+1}$. One can see this by associating to the function f the vector $(f(0), f(1), \ldots, f(N))$. The set $\mathcal{V}$ is a subspace of this vector space. Recall to show that a subspace has dimension k, it suffices to find a basis for the subspace with k elements $v_1, \ldots, v_k$. To show they form a basis, we need to show that they are linearly independent and that every vector in the subspace is a linear combination of them.

Proof 2. Suppose F is a solution to (1.8). Then for each $0 < x < N$,

$$F(x) \leq \max\{F(x-1), F(x+1)\}.$$

Using this we can see that the maximum of F is obtained either at 0 or at N. Similarly, the minimum of F is obtained on $\{0, N\}$. Suppose $F(0) = 0, F(N) = 0$. Then the minimum and the maximum of the function are both 0 which means that $F \equiv 0$. Suppose $F(0) = a, F(N) = b$ and let F_0 be the linear function with these same boundary values. Then $F - F_0$ satisfies (1.8) with boundary value 0, and hence is identically zero. This implies that $F = F_0$.

Proof 3. Consider the equations (1.8) as $N - 1$ linear equations in $N - 1$ unknowns, $F(1), \ldots, F(N-1)$. We can write this as

$$\mathbf{A}\mathbf{v} = \mathbf{w},$$

where

$$\mathbf{A} = \begin{bmatrix} -1 & \frac{1}{2} & 0 & 0 & \cdots & 0 & 0 \\ \frac{1}{2} & -1 & \frac{1}{2} & 0 & \cdots & 0 & 0 \\ 0 & \frac{1}{2} & -1 & \frac{1}{2} & \cdots & 0 & 0 \\ & & & & \vdots & & \\ 0 & 0 & 0 & 0 & \cdots & -1 & \frac{1}{2} \\ 0 & 0 & 0 & 0 & \cdots & \frac{1}{2} & -1 \end{bmatrix}, \quad \mathbf{w} = \begin{bmatrix} -\frac{F(0)}{2} \\ 0 \\ 0 \\ \vdots \\ 0 \\ -\frac{F(N)}{2} \end{bmatrix}.$$

If we prove that $\mathbf{A}$ is invertible, then the unique solution is $\mathbf{v} = \mathbf{A}^{-1}\mathbf{w}$. To prove invertibility it suffices to show that $\mathbf{A}\mathbf{v} = 0$ has a unique solution and this can be done by an argument as in the previous proof.

Proof 4. Suppose F is a solution to (1.8). Let S_n be the random walk starting at x. We claim that for all n, $\mathbb{E}[F(S_{n\wedge T})] = F(x)$. We will show this by induction. For $n = 0$, $F(S_0) = F(x)$ and hence $\mathbb{E}[F(S_0)] = F(x)$. To do the inductive step, we use a rule for expectation in terms of conditional expectations:

$$\mathbb{E}[F(S_{(n+1)\wedge T})] = \sum_{y=0}^{N} \mathbb{P}\{S_{n\wedge T} = y\}\, \mathbb{E}[F(S_{(n+1)\wedge T}) \mid S_{n\wedge T} = y].$$

If $y = 0$ or $y = N$ and $S_{n\wedge T} = y$, then $S_{(n+1)\wedge T} = y$ and hence $\mathbb{E}[F(S_{(n+1)\wedge T}) \mid S_{n\wedge T} = y] = F(y)$. If $0 < y < x$ and $S_{n\wedge T} = y$, then

$$\mathbb{E}[F(S_{(n+1)\wedge T}) \mid S_{n\wedge T} = y] = \frac{1}{2}\, F(y+1) + \frac{1}{2}\, F(y-1) = F(y).$$

Therefore,

$$\mathbb{E}[F(S_{(n+1)\wedge T})] = \sum_{y=0}^{N} \mathbb{P}\{S_{n\wedge T} = y\}\, F(y) = \mathbb{E}[F(S_{n\wedge T})] = F(x),$$

with the last equality holding by the inductive hypothesis. Therefore,

$$\begin{aligned} F(x) &= \lim_{n\to\infty} \mathbb{E}[F(S_{n\wedge T})] \\ &= \lim_{n\to\infty} \sum_{y=0}^{N} \mathbb{P}\{S_{n\wedge T} = y\}\, F(y) \\ &= \mathbb{P}\{S_T = 0\}\, F(0) + \mathbb{P}\{S_T = N\}\, F(N) \\ &= [1 - \mathbb{P}\{S_T = N\}]\, F(0) + \mathbb{P}\{S_T = N\}\, F(N). \end{aligned}$$

Considering the case $F(0) = 0, F(N) = 1$ gives $\mathbb{P}\{S_T = N \mid S_0 = x\} = x/N$ and for more general boundary conditions,

$$F(x) = F(0) + \frac{x}{N}\,[F(N) - F(0)].$$

One nice thing about the last proof is that it was not necessary to have already guessed the linear functions as solutions. The proof produces these solutions.

1.2.2. Higher dimensions. We will generalize this result to higher dimensions. We replace the interval $\{1, \dots, N\}$ with an arbitrary finite subset A of $\mathbb{Z}^d$. We let ∂A be the *(outer) boundary* of A defined by

$$\partial A = \{z \in \mathbb{Z}^d \setminus A : \operatorname{dist}(z, A) = 1\},$$

and we let $\overline{A} = A \cup \partial A$ be the *"closure"* of A.

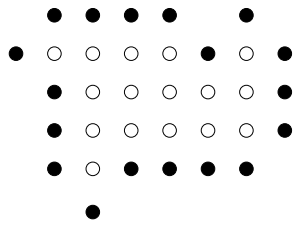

Figure 3. The white dots are A and the black dots are ∂A

♢ The term closure may seem strange, but in the continuous analogue, A will be an open set, ∂A its topological boundary and $\overline{A} = A \cup \partial A$ its topological closure.

We define the linear operators $\mathbf{Q}, \mathcal{L}$ on functions by

$$\mathbf{Q}F(x) = \frac{1}{2d} \sum_{y \in \mathbb{Z}^d, |x-y|=1} F(y),$$

$$\mathcal{L}F(x) = (\mathbf{Q} - \mathbf{I})F(x) = \frac{1}{2d} \sum_{y \in \mathbb{Z}^d, |x-y|=1} [F(y) - F(x)]$$

The operator $\mathcal{L}$ is often called the *(discrete) Laplacian*. We let S_n be a simple random walk in $\mathbb{Z}^d$. Then we can write

$$\mathcal{L}F(x) = \mathbb{E}[F(S_1) - F(S_0) \mid S_0 = x].$$

We say that F is *(discrete) harmonic* at x if $\mathcal{L}F(x) = 0$; this is an example of a *mean-value property*. The corresponding boundary value problem we will state is sometimes called the *Dirichlet problem* for harmonic functions.

♢ The term linear operator is often used for a linear function whose domain is a space of functions. In our case, the domain is the space of functions on the finite set A which is isomorphic to $\mathbb{R}^K$ where $K = \#(A)$. In this case a linear operator is the same as a linear transformation from linear algebra. We can think of $\mathbf{Q}$ and $\mathcal{L}$ as $K \times K$ matrices. We can write $\mathbf{Q} = [Q(x,y)]_{x,y\in A}$ where $Q(x,y) = 1/(2d)$ if $|x-y| = 1$ and otherwise $Q(x,y) = 0$. Define $Q_n(x,y)$ by $\mathbf{Q}^n = [Q_n(x,y)]$. Then $Q_n(x,y)$ is the probability that the random walk starts at x, is at site y at time n, and and has not left the set A by time n.

Dirichlet problem for harmonic functions. Given a set $A \subset \mathbb{Z}^d$ and a function $F : \partial A \to \mathbb{R}$ find an extension of F to $\overline{A}$ such that F is harmonic in A, i.e.,

$$\mathcal{L}F(x) = 0 \text{ for all } x \in A. \tag{1.9}$$

For the case $d = 1$ and $A = \{1, \ldots, N-1\}$, we were able to guess the solution and then verify that it is correct. In higher dimensions, it is not so obvious how to give a formula for the solution. We will show that the last proof for $d = 1$ generalizes in a natural way to $d > 1$. We let $T_A = \min\{n \geq 0 : S_n \notin A\}$.

Theorem 1.5. *If $A \subset \mathbb{Z}^d$ is finite, then for every $F : \partial A \to \mathbb{R}$, there is a unique extension of F to $\overline{A}$ that satisfies* (1.9)*. It is given by*

$$F_0(x) = \mathbb{E}[F(S_{T_A}) \mid S_0 = x] = \sum_{y\in\partial A} \mathbb{P}\{S_{T_A} = y \mid S_0 = x\}\, F(y).$$

It is not difficult to verify that F_0 as defined above is a solution to the Dirichlet problem. The problem is to show that it is unique. Suppose F is harmonic on A; $S_0 = x \in \overline{A}$; and let

$$M_n = F(S_{n\wedge T_A}).$$

Then (1.9) can be rewritten as

$$\mathbb{E}[M_{n+1} \mid S_0, \ldots, S_n] = F(S_{n\wedge T_A}) = M_n. \tag{1.10}$$

A process that satisfies $\mathbb{E}[M_{n+1} \mid S_0, \ldots, S_n] = M_n$ is called a *martingale (with respect to the random walk)*. It is easy to see that $F(S_{n\wedge T_A})$ being a martingale is essentially equivalent to F being harmonic on A. It is easy to check that martingales satisfy $\mathbb{E}[M_n] = \mathbb{E}[M_0]$, and hence if $S_0 = x$,

$$\sum_{y\in\overline{A}} \mathbb{P}\{S_{n\wedge T_A} = y\}\, F(y) = \mathbb{E}[M_n] = M_0 = F(x).$$

An easy argument shows that with probability one $T_A < \infty$. We can take limits and get

$$F(x) = \lim_{n\to\infty} \sum_{y\in\overline{A}} \mathbb{P}\{S_{n\wedge T_A} = y\}\, F(y) = \sum_{y\in\partial A} \mathbb{P}\{S_{T_A} = y\}\, F(y). \tag{1.11}$$

♢ There is no problem interchanging the limit and the sum because it is a finite sum. If A is infinite, one needs more assumptions to justify the exchange of the limit and the sum.

Let us consider this from the perspective of linear algebra. Suppose that A has N elements and ∂A has K elements. The solution of the Dirichlet problem assigns to each function on ∂A (a vector in $\mathbb{R}^K$) a function on A (a vector in $\mathbb{R}^N$). Hence the solution can be considered as a linear function from $\mathbb{R}^K$ to $\mathbb{R}^N$ (the reader should check that this is a linear transformation). Any linear transformation is given by an $N \times K$ matrix. Let us write the matrix for the solution as

$$\mathbf{H}_A = [H_A(x,y)]_{x\in A, y\in\partial A}.$$

Another way of stating (1.11) is to say that

$$H_A(x,y) = \mathbb{P}\{S_{T_A} = y \mid S_0 = x\}.$$

This matrix is often called the *Poisson kernel*. For a given set A, we can solve the Dirichlet problem for any boundary function in terms of the Poisson kernel.

$\diamondsuit$ Analysts who are not comfortable with probability[1] think of the Poisson kernel only as the matrix for the transformation which takes boundary data to values on the interior. Probabilists also have the interpretation of $H_A(x, y)$ as the probability that the random walk starting at x exits A at y.

What happens in Theorem 1.5 if we allow A to be an infinite set? In this case it is not always true that the solution is unique. Let us consider the one-dimensional example with $A = \{1, 2, 3, \ldots\}$ and $\partial A = \{0\}$. Then for every $c \in \mathbb{R}$, the function $F(x) = cx$ is harmonic in A with boundary value 0 at the origin. Where does our proof break down? This depends on which proof we consider (they all break down!), but let us consider the martingale version. Suppose F is harmonic on A with $F(0) = 0$ and suppose S_n is a simple random walk starting at positive integer x. As before, we let $T = \min\{n \geq 0 : S_n = 0\}$ and $M_n = F(S_{n\wedge T})$. The same argument shows that M_n is a martingale and

$$F(x) = \mathbb{E}[M_n] = \sum_{y=0}^{\infty} F(y)\, \mathbb{P}\{S_{n\wedge T} = y\}.$$

We have shown in a previous section that with probability one $T < \infty$. This implies that $\mathbb{P}\{S_{n\wedge T} = 0\}$ tends to 1, i.e.,

$$\lim_{n\to\infty} \sum_{y>0} \mathbb{P}\{S_{n\wedge T} = y\} = 0.$$

However, if F is unbounded, we cannot conclude from this that

$$\lim_{n\to\infty} \sum_{y>0} F(y)\, \mathbb{P}\{S_{n\wedge T} = y\} = 0.$$

However, we do see from this that there is only one *bounded* function that is harmonic on A with a given boundary value at 0. We state the theorem leaving the details as Exercise 1.7.

---

[1] The politically correct term is stochastically challenged.

Theorem 1.6. *Suppose A is a proper subset of $\mathbb{Z}^d$ such that for all $x \in \mathbb{Z}^d$,*

$$\lim_{n\to\infty} \mathbb{P}\{T_A > n \mid S_0 = x\} = 0.$$

Suppose $F : \partial A \to \mathbb{R}$ is a bounded function. Then there is a unique bounded extension of F to $\overline{A}$ that satisfies (1.9)*. It is given by*

$$F_0(x) = \mathbb{E}[F(S_{T_A}) \mid S_0 = x] = \sum_{y\in\partial A} \mathbb{P}\{S_{T_A} = y \mid S_0 = x\}\, F(y).$$

1.3. Heat equation

We will now introduce a mathematical model for heat flow. Let A be a finite subset of $\mathbb{Z}^d$ with boundary ∂A. We set the temperature at the boundary to be zero at all times and as an initial condition set the temperature at $x \in A$ to be $p_n(x)$. At each integer time unit n, the heat at x at time n is spread evenly among its $2d$ nearest neighbors. If one of those neighbors is a boundary point, then the heat that goes to that site is lost forever. A more probabilistic view of this is given by imagining that the temperature in A to be controlled by a very large number of "heat particles". These particles perform random walks on A until they leave A at which time they are killed. The temperature at x at time n, $p_n(x)$ is given by the density of particles at x. Either interpretation gives a difference equation for the temperature $p_n(x)$. For $x \in A$, the temperature at x is given by the amount of heat going in from neighboring sites,

$$p_{n+1}(x) = \frac{1}{2d} \sum_{|y-x|=1} p_n(y).$$

If we introduce the notation $\partial_n p_n(x) = p_{n+1}(x) - p_n(x)$, we get the *heat equation*

$$\partial_n p_n(x) = \mathcal{L} p_n(x), \quad x \in A, \tag{1.12}$$

where $\mathcal{L}$ denotes the discrete Laplacian as before. The initial temperature is given as an initial condition

$$p_0(x) = f(x), \quad x \in A. \tag{1.13}$$

We rewrite the boundary condition

$$p_n(x) = 0, \quad x \in \partial A. \tag{1.14}$$

If $x \in A$ and the initial condition is $f(x) = 1$ and $f(z) = 0$ for $z \neq x$, then

$$p_n(y) = \mathbb{P}\{S_{n \wedge T_A} = y \mid S_0 = x\}.$$

♢ The heat equation is a deterministic (i.e., without randomness) model for heat flow. It can be studied without probability. However, probability adds a layer of richness in terms of movements of individual random particles. This extra view is often useful for understanding the equation.

Given any initial condition f, it is easy to see that there is a unique function p_n satisfying (1.12)–(1.14). Indeed, we just set: $p_n(y) = 0$ for all $n \geq 0$ if $y \in \partial A$; $p_0(x) = f(x)$ if $x \in A$; and for $n > 0$, we define $p_n(x), x \in A$ recursively by (1.12). This tells us that set of functions satisfying (1.12) and (1.14) is a vector space of dimension $\#(A)$. In fact, $\{p_n(x) : x \in A\}$ is the vector $\mathbf{Q}^n f$.

Once we have existence and uniqueness, the problem remains to find the function. For a bounded set A, this is a problem in linear algebra and essentially becomes the question of diagonalizing the matrix $\mathbf{Q}$.

♢ Recall from linear algebra that if $\mathbf{A}$ is a $k \times k$ symmetric matrix with real entries, then we can find k (not necessarily distinct) real eigenvalues

$$\lambda_k \leq \lambda_{k-1} \leq \cdots \leq \lambda_1,$$

and k orthogonal vectors $\mathbf{v}_1, \ldots, \mathbf{v}_k$ that are eigenvectors,

$$\mathbf{A}\mathbf{v}_j = \lambda_j \, \mathbf{v}_j.$$

(If A is not symmetric, A might not have k linearly independent eigenvectors, some eigenvalues might not be real, and eigenvectors for different eigenvalues are not necessarily orthogonal.)

We will start by considering the case $d = 1$. Let us compute the function p_n for $A = \{1, \ldots, N-1\}$. We start by looking for functions satisfying (1.12) of the form

$$p_n(x) = \lambda^n \, \phi(x). \tag{1.15}$$

If p_n is of this form, then

$$\partial_n p_n(x) = \lambda^{n+1} \, \phi(x) - \lambda^n \, \phi(x) = (\lambda - 1) \, \lambda^n \, \phi(x).$$

This nice form leads us to try to find eigenvalues and eigenfunctions of $\mathbf{Q}$, i.e., to find λ, ϕ such that

$$\mathbf{Q}\phi(x) = \lambda\,\phi(x), \tag{1.16}$$

with $\phi \equiv 0$ on ∂A.

♢ The "algorithmic" way to find the eigenvalues and eigenvectors for a matrix Q is first to find the eigenvalues as the roots of the characteristic polynomial and then to find the corresponding eigenvector for each eigenvalue. Sometimes we can avoid this if we can make good guesses for the eigenvectors. This is what we will do here.

The sum rule for sine,

$$\sin((x \pm 1)\theta) = \sin(x\theta)\cos(\theta) \pm \cos(x\theta)\sin(\theta),$$

tells us that

$$\mathbf{Q}\{\sin(\theta x)\} = \lambda_\theta\,\{\sin(\theta x)\}, \qquad \lambda_\theta = \cos\theta,$$

where $\{\sin(\theta x)\}$ denotes the vector whose component associated to $x \in A$ is $\sin(\theta x)$. If we choose $\theta_j = \pi j/N$, then $\phi_j(x) = \sin(\pi j x/N)$ which satisfies the boundary condition $\phi_j(0) = \phi_j(N) = 0$. Since these are eigenvectors with different eigenvalues for a symmetric matrix $\mathbf{Q}$, we know that they are orthogonal, and hence linearly independent. Hence every function f on A can be written in a unique way as

$$f(x) = \sum_{j=1}^{N-1} c_j\,\sin\left(\frac{\pi j x}{N}\right). \tag{1.17}$$

This sum in terms of trigonometric functions is called a finite *Fourier series*. The solution to the heat equation with initial condition f is

$$p_n(y) = \sum_{j=1}^{N-1} c_j\,\left[\cos\left(\frac{j\pi}{N}\right)\right]^n\,\phi_j(y).$$

Orthogonality of eigenvectors tells us that

$$\sum_{x=1}^{N-1} \sin\left(\frac{\pi j x}{N}\right)\,\sin\left(\frac{\pi k x}{N}\right) = 0 \text{ if } j \neq k.$$

Also,

$$\sum_{x=1}^{N-1} \sin^2\left(\frac{\pi j x}{N}\right) = \frac{N}{2}. \tag{1.18}$$

♢ The Nth roots of unity, $\zeta_1, \ldots, \zeta_N$ are the N complex numbers ζ such that $\zeta^N = 1$. They are given by

$$\zeta_k = \cos\left(\frac{2k\pi}{N}\right) + i\,\sin\left(\frac{2k\pi}{N}\right), \quad j = 1, \ldots, N.$$

The roots of unity are spread evenly about the unit circle in $\mathbb{C}$; in particular,

$$\omega_1 + \omega_2 + \cdots + \omega_N = 0,$$

which implies that

$$\sum_{j=1}^{N} \cos\left(\frac{2k\pi}{N}\right) = \sum_{j=1}^{N} \sin\left(\frac{2k\pi}{N}\right) = 0.$$

The double angle formula for sine gives

$$\begin{aligned}
\sum_{j=1}^{N-1} \sin^2\left(\frac{jx\pi}{N}\right) &= \sum_{j=1}^{N} \sin^2\left(\frac{jx\pi}{N}\right) \\
&= \frac{1}{2} \sum_{j=0}^{N-1} \left[1 - \cos\left(\frac{2jx\pi}{N}\right)\right] \\
&= \frac{N}{2} - \frac{1}{2} \sum_{j=1}^{N} \cos\left(\frac{2jx\pi}{N}\right).
\end{aligned}$$

If x is an integer, the last sum is zero. This gives (1.18).

In particular, if we choose the solution with initial condition $f(x) = 1; f(z) = 0, z \neq x$ we can see that

$$\mathbb{P}\{S_{n\wedge T_A} = y \mid S_0 = x\} = \frac{2}{N} \sum_{j=1}^{N-1} \phi_j(x) \left[\cos\left(\frac{j\pi}{N}\right)\right]^n \phi_j(y).$$

It is interesting to see what happens as $n \to \infty$. For large n, the sum is very small but it is dominated by the $j = 1$ and $j = N - 1$ terms for which the eigenvalue has maximal absolute value. These two terms give

$$\frac{2}{N} \cos^n\left(\frac{\pi}{N}\right) \left[\sin\left(\frac{\pi x}{N}\right) \sin\left(\frac{\pi y}{N}\right) +\right.$$

$$(-1)^n \sin\left(\frac{x\pi(N-1)}{N}\right) \sin\left(\frac{y\pi(N-1)}{N}\right)\Big].$$

One can check that

$$\sin\left(\frac{x\pi(N-1)}{N}\right) = (-1)^{x+1} \sin\left(\frac{\pi x}{N}\right),$$

and hence if $x, y \in \{1, \ldots, N-1\}$, as $n \to \infty$,

$$\mathbb{P}\{S_{n\wedge T_A} = y \mid S_0 = x\} \sim$$
$$\frac{2}{N} \cos^n\left(\frac{\pi}{N}\right) [1 + (-1)^{n+x+y}] \sin\left(\frac{\pi x}{N}\right) \sin\left(\frac{\pi y}{N}\right).$$

For large n, conditioned that the walker has not left $\{1, \ldots, N-1\}$, the probability that the walker is at y is about $c \sin(\pi y/N)$ assuming that the "parity" is correct ($n + x + y$ is even). Other than the parity, there is no dependence on the starting point x for the limiting distribution. Note that the walker is more likely to be at points toward the "middle" of the interval.

The above example illustrates a technique for finding solutions of the form (1.15) called *separation of variables*. The same idea works for all d although it may not always be possible to give nice expressions for the eigenvalues and eigenvectors. For finite A this is essentially the same as computing powers of a matrix by diagonalization. We summarize here.

Theorem 1.7. *If A is a finite subset of $\mathbb{Z}^d$ with N elements, then we can find N linearly independent functions $\phi_1, \ldots, \phi_N$ that satisfy* (1.16) *with real eigenvalues $\lambda_1, \ldots, \lambda_N$. The solution to* (1.12)–(1.14) *is given by*

$$p_n(x) = \sum_{j=1}^{N} c_j \, \lambda_j^n \, \phi_j(x),$$

where c_j are chosen so that

$$f(x) = \sum_{j=1}^{N} c_j \, \phi_j(x).$$

In fact, the ϕ_j can be chosen to be orthonormal,

$$\langle \phi_j, \phi_k \rangle := \sum_{x \in A} \phi_j(x) \, \phi_k(x) = \delta(k - j).$$

♢ Here we have introduced the *delta function* notation, $\delta(z) = 1$ if $z = 0$ and $\delta(z) = 0$ if $z \neq 0$.

Since $p_n(x) \to 0$ as $n \to \infty$, we know that the eigenvalues have absolute value strictly less than one. We can order the eigenvalues

$$1 > \lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_N > -1.$$

We will write $p(x, y; A)$ to be the solution of the heat equation with initial condition equal to one at x and 0 otherwise. In other words,

$$p_n(x, y; A) = \mathbb{P}\{S_n = y, T_A > n \mid S_0 = x\}, \qquad x, y \in A.$$

Then if $\#(A) = N$,

$$p_n(x, y; A) = \sum_{j=1}^{N} c_j(x)\, \lambda_j^n\, \phi_j(y)$$

where $c_j(x)$ have been chosen so that

$$\sum_{j=1}^{N} c_j(x)\phi_j(y) = \delta(y - x).$$

In fact, this tells us that $c_j(x) = \phi_j(x)$. Hence

$$p_n(x, y; A) = \sum_{j=1}^{N} \lambda_j^n\, \phi_j(x)\, \phi_j(y).$$

Note that the quantity on the right is symmetric in x, y. One can check that the symmetry also follows from the definition of $p_n(x, y; A)$.

The largest eigenvalue λ_1 is often denoted λ_A. We can give a "variational" definition of λ_A as follows. This is really just a theorem about the largest eigenvalue of symmetric matrices.

Theorem 1.8. *If A is a finite subset of $\mathbb{Z}^d$, then λ_A is given by*

$$\lambda_A = \sup_f \frac{\langle \mathbf{Q}f, f\rangle}{\langle f, f\rangle},$$

where the supremum is over all functions f on A, and $\langle \cdot, \cdot \rangle$ denotes inner product

$$\langle f, g\rangle = \sum_{x \in A} f(x)\, g(x).$$

Proof. If ϕ is an eigenvector with eigenvalue λ_1, then $\mathbf{Q}\phi = \lambda_1\phi$ and setting $f = \phi$ shows that the supremum is at least as large as λ_1. Conversely, there is an orthogonal basis of eigenfunctions $\phi_1, \ldots, \phi_N$ and we can write any f as

$$f = \sum_{j=1}^{N} c_j \, \phi_j.$$

Then

$$\begin{aligned}
\langle \mathbf{Q}f, f\rangle &= \left\langle \mathbf{Q}\sum_{j=1}^{N} c_j\,\phi_j, \sum_{j=1}^{N} c_j\,\phi_j \right\rangle \\
&= \left\langle \sum_{j=1}^{N} c_j \mathbf{Q}\phi_j, \sum_{j=1}^{N} c_j\,\phi_j \right\rangle \\
&= \sum_{j=1}^{N} c_j^2\,\lambda_j\,\langle \phi_j, \phi_j\rangle \\
&\le \lambda_1 \sum_{j=1}^{N} c_j^2\,\langle \phi_j, \phi_j\rangle = \lambda_1\,\langle f, f\rangle.
\end{aligned}$$

The reader should check that the computation above uses the orthogonality of the eigenfunctions and also the fact that $\langle \phi_j, \phi_j\rangle \ge 0$. □

Using this variational formulation, we can see that the eigenfunction for λ_1 can be chosen so that $\phi_1(x) \ge 0$ for each x (since if ϕ_1 took on both positive and negative values, we would have $\langle \mathbf{Q}|\phi_1|, |\phi_1|\rangle > \langle \phi_1, \phi_1\rangle$). The eigenfunction is unique, i.e., $\lambda_2 < \lambda_1$, provided we put an additional condition on A. We say that a subset A on $\mathbb{Z}^d$ is *connected* if any two points in A are connected by a nearest neighbor path that stays entirely in A. Equivalently, A is connected if for each $x, y \in A$ there exists an n such that $p_n(x, y; A) > 0$. We leave it as Exercise 1.23 to show that this implies that $\lambda_1 > \lambda_2$.

Before stating the final theorem, we need to discuss some parity (even/odd) issues. If $x = (x_1, \ldots, x_d) \in \mathbb{Z}^d$ we let $\mathrm{par}(x) = (-1)^{x_1+\cdots+x_d}$. We call x *even* if $\mathrm{par}(x) = 1$ and otherwise x is odd. If n is a nonnegative integer, then

$$p_n(x, y; A) = 0 \quad \text{if} \quad (-1)^n \mathrm{par}(x + y) = -1.$$

If $\mathbf{Q}\phi = \lambda\phi$, then $\mathbf{Q}[\mathrm{par}\phi] = -\lambda \mathrm{par}\phi$.

Theorem 1.9. *Suppose A is a finite connected subset of $\mathbb{Z}^d$ with at least two points. Then $\lambda_1 > \lambda_2$, $\lambda_N = -\lambda_1 < \lambda_{N-1}$. The eigenfunction ϕ_1 can be chosen so that $\phi_1(x) > 0$ for all $x \in A$.*

$$\lim_{n\to\infty} \lambda_1^{-n}\, p_n(x, y; A) = [1 + (-1)^n \operatorname{par}(x + y)]\, \phi_1(x)\, \phi_1(y).$$

Example 1.10. One set in $\mathbb{Z}^d$ for which we can compute the eigenfunctions and eigenvalues exactly is a d-dimensional rectangle

$$A = \{(x_1, \ldots, x_d) \in \mathbb{Z}^d : 1 \le x_j \le N_j - 1\}.$$

The eigenfunctions are indexed by $\bar{k} = (k_1, \ldots, k_d) \in A$,

$$\phi_{\bar{k}}(x_1, \ldots, x_d) = \sin\left(\frac{k_1 \pi x_1}{N_1}\right) \sin\left(\frac{k_2 \pi x_2}{N_2}\right) \cdots \sin\left(\frac{k_d \pi x_d}{N_d}\right),$$

with eigenvalue

$$\lambda_{\bar{k}} = \frac{1}{d}\left[\cos\left(\frac{k_1 \pi}{N_1}\right) + \cdots + \cos\left(\frac{k_d \pi}{N_d}\right)\right].$$

1.4. Expected time to escape

1.4.1. One dimension. Let S_n denote a one-dimensional random walk starting at $x \in \{0, \ldots, N\}$ and let T be the first time that the walker reaches $\{0, N\}$. Here we study the expected time to reach 0 or N,

$$e(x) = \mathbb{E}[T \mid S_0 = x].$$

Clearly $e(0) = e(N) = 0$. Now suppose $x \in \{1, \ldots, N-1\}$. Then the walker takes one step which goes to either $x - 1$ or $x + 1$. Using this we get the relation

$$e(x) = 1 + \frac{1}{2}\left[e(x + 1) + e(x - 1)\right].$$

Hence e satisfies

$$e(0) = e(N) = 0, \quad \mathcal{L}e(x) = -1, \quad x = 1, \ldots, N - 1. \tag{1.19}$$

A simple calculation shows that if $f(x) = x^2$, then $\mathcal{L}f(x) = 1$ for all x. Also the linear function $g(x) = x$ is harmonic, $\mathcal{L}g \equiv 0$. Using this we can see that one solution to (1.19) is

$$e(x) = x\,(N - x).$$

In fact, as we will now show, it is the unique solution. Assume that e_1 is another solution. Then for $x = 1, \ldots, N-1$,

$$\mathcal{L}(e - e_1)(x) = \mathcal{L}e(x) - \mathcal{L}e_1(x) = -1 - (-1) = 0,$$

i.e., $e - e_1$ is harmonic on $\{1, \ldots, N-1\}$. Since this function also vanishes at 0 and N we know that $e - e_1 = 0$.

Suppose $N = 2m$ is even. Then we get

$$e(m) = N^2/4 = m^2.$$

In other words, the expected time for a random walker starting at m (or anywhere else, in fact) to go distance m is *exactly* m^2.

Suppose the random walker starts at $x = 1$. Then the expected time to leave the interval is $N - 1$. While this is an expected value, it is not necessarily a "typical" value. Most of the time the random walker will leave quickly. However, the gambler's ruin estimate tells us that there is a probability of $1/m$ that the random walker will reach m before leaving the interval. If that happens then the walker will still have on the order of N^2 steps before leaving.

One other interesting fact concerns the time until a walker starting at 1 reaches the origin. Let T_0 be the first n such that $S_n = 0$. If $S_0 = 1$, we know that $T_0 < \infty$ with probability one. However, the amount of time to reach 0 is at least as large as the amount of time to reach 0 or N. Therefore $\mathbb{E}[T_0] \geq N$. Since this is true for every N, we must have $\mathbb{E}[T_0] = \infty$. In other words, while it is guaranteed that a random walker will return to the origin *the expected amount of time until it happens is infinite!*

1.4.2. Several dimensions. Let A be a finite subset of $\mathbb{Z}^d$; S_n a simple random walker starting at $x \in \overline{A}$; and T_A the first time that the walker is not in A. Let

$$e_A(x) = \mathbb{E}[T_A \mid S_0 = x].$$

Then just as in the one-dimensional case we can see that $f(x) = e_A(x)$ satisfies

$$f(x) = 0, \quad x \in \partial A \tag{1.20}$$

$$\mathcal{L}f(x) = -1, \quad x \in A. \tag{1.21}$$

We can argue in the same as in the one-dimensional case that there is at most one function satisfying these equations. Indeed if f, g were two such functions, then $\mathcal{L}[f-g] \equiv 0$ with $f - g \equiv 0$ on ∂A, and only the zero function satisfies this.

Let $f(x) = |x|^2 = x_1^2 + \cdots + x_d^2$. Then a simple calculation shows that $\mathcal{L}f(x) = 1$. Let us consider the process

$$M_n = |S_{n \wedge T_A}|^2 - (n \wedge T_A).$$

Then, we can see that

$$\mathbb{E}[M_{n+1} \mid S_0, \ldots, S_n] = M_n,$$

and hence M_n is a martingale. This implies

$$\mathbb{E}[M_n] = \mathbb{E}[M_0] = |S_0|^2, \quad \mathbb{E}[n \wedge T_A] = \mathbb{E}[|S_{n \wedge T_A}|^2] - |S_0|^2.$$

In fact, we claim we can take the limit to assert

$$\mathbb{E}[T_A] = \mathbb{E}[|S_{T_A}|^2] - |S_0|^2.$$

To prove this we use the *monotone convergence theorem*, see Exercise 1.6. This justifies the step

$$\lim_{n \to \infty} \left[\mathbb{E}[|S_{T_A}|^2 \, 1\{T_A \le n\}\right] = \mathbb{E}[|S_{T_A}|^2].$$

Also,

$$\mathbb{E}\left[|S_{T_A}|^2 \, 1\{T_A > n\}\right] \le \mathbb{P}\{T_A > n\} \left[\max_{x \in A} |x|^2\right] \to 0.$$

This is a generalization of a formula we derived in the one-dimensional case. If $d = 1$ and $A = \{1, \ldots, N-1\}$, and

$$\mathbb{E}[|S_{T_A}|^2] = N^2 \, \mathbb{P}\{S_{T_A} = N \mid S_0 = x\} = N\, x,$$

$$\mathbb{E}[T_A] = \mathbb{E}[|S_{T_A}|^2] - x^2 = x\,(N - x).$$

Example 1.11. Suppose that A is the "discrete ball" of radius r about the origin,

$$A = \{x \in \mathbb{Z}^d : |x| < r\}.$$

Then every $y \in \partial A$ satisfies $r \le |y| < r + 1$. Suppose we start the random walk at the origin. Then,

$$r^2 \le \mathbb{E}[T_A] < (r+1)^2.$$

For any $y \in A$, let V_y denote the number of visits to y before leaving A,

$$V_y = \sum_{n=0}^{T_A-1} 1\{S_n = y\} = \sum_{n=0}^{\infty} 1\{S_n = y, T_A > n\}.$$

Here we again use the indicator function notation. Note that

$$\mathbb{E}[V_y \mid S_0 = x] = \sum_{n=0}^{\infty} \mathbb{P}\{S_n = y, T_A > n \mid S_0 = x\} = \sum_{n=0}^{\infty} p_n(x, y; A).$$

This quantity is of sufficient interest that it is given a name. The *Green's function* $G_A(x, y)$ is the function on $A \times A$ given by

$$G_A(x, y) = \mathbb{E}[V_y \mid S_0 = x] = \sum_{n=0}^{\infty} p_n(x, y; A).$$

We define $G_A(x, y) = 0$ if $x \notin A$ or $y \notin A$. The Green's function satisfies $G_A(x, y) = G_A(y, x)$. This is not immediately obvious from the first equality but follows from the symmetry of $p_n(x, y; A)$. If we fix $y \in A$, then the function $f(x) = G_A(x, y)$ satisfies the following:

$$\begin{aligned} \mathcal{L}f(y) &= -1, \\ \mathcal{L}f(x) &= 0, \quad x \in A \setminus \{y\}, \\ f(x) &= 0, \quad x \in \partial A. \end{aligned}$$

Note that

$$T_A = \sum_{y \in A} V_y,$$

and hence

$$\mathbb{E}[T_A \mid S_0 = x] = \sum_{y \in A} G_A(x, y).$$

Theorem 1.12. *Suppose A is a bounded subset of $\mathbb{Z}^d$, and $g : A \to \mathbb{R}$ is a given function. Then the unique function $F : \overline{A} \to \mathbb{R}$ satisfying*

$$\begin{aligned} F(x) &= 0, \quad x \in \partial A, \\ \mathcal{L}F(x) &= -g(x), \quad x \in A, \end{aligned}$$

is

$$F(x) = \mathbb{E}\left[\sum_{j=0}^{T_A-1} g(S_j) \mid S_0 = x\right] = \sum_{y \in A} g(y)\, G_A(x, y). \tag{1.22}$$

We have essentially already proved this. Uniqueness follows from the fact that if F, F_1 are both solutions, then $F - F_1$ is harmonic in A with boundary value 0 and hence equals 0 everywhere. Linearity of $\mathcal{L}$ shows that

$$\mathcal{L}\left[\sum_{y\in A} g(y)\, G_A(x,y)\right] = \sum_{y\in A} g(y)\, \mathcal{L}G_A(x,y) = -g(x). \tag{1.23}$$

The second equality in (1.22) follows by writing

$$\begin{aligned}
\sum_{j=0}^{T_A-1} g(S_j) &= \sum_{j=0}^{T_A-1} \sum_{y\in A} g(y)\, 1\{S_j = y\} \\
&= \sum_{y\in A} g(y) \sum_{j=0}^{T_A-1} 1\{S_j = y\} \\
&= \sum_{y\in A} g(y)\, V_y.
\end{aligned}$$

We can consider the Green's function as a matrix or operator,

$$G_A g(x) = \sum_{x\in A} G_A(x,y)\, g(y).$$

Then (1.23) can be written as

$$-\mathcal{L}G_A g(x) = g(x),$$

or $G_A = [-\mathcal{L}]^{-1}$ For this reason the Green's function is often referred to as the *inverse of (the negative of) the Laplacian.*

If $d \geq 3$, then the expected number of visits to a point is finite and we can define the *(whole space) Green's function*

$$\begin{aligned}
G(x,y) = \lim_{A\uparrow\mathbb{Z}^d} G_A(x,y) &= \mathbb{E}\left[\sum_{n=0}^{\infty} 1\{S_n = y\} \mid S_0 = x\right] \\
&= \sum_{n=0}^{\infty} \mathbb{P}\{S_n = y \mid S_0 = x\}.
\end{aligned}$$

It is a bounded function. In fact, if τ_y denotes the smallest $n \geq 0$ such that $S_n = y$, then

$$\begin{aligned}
G(x,y) &= \mathbb{P}\{\tau_y < \infty \mid S_0 = x\}\, G(y,y) \\
&= \mathbb{P}\{\tau_y < \infty \mid S_0 = x\}\, G(0,0) \leq G(0,0) < \infty.
\end{aligned}$$

The function G is symmetric and satisfies a translation invariance property: $G(x,y) = G(0, y-x)$. For fixed y, $f(x) = G(x,y)$ satisfies

$$\mathcal{L}f(y) = -1, \quad \mathcal{L}f(x) = 0,\ x \neq y, \quad f(x) \to 0 \text{ as } x \to \infty.$$

1.5. Space of harmonic functions

If $d = 1$, the only harmonic functions $f : \mathbb{Z} \to \mathbb{R}$ are the linear functions $f(x) = ax + b$. This follows since $\mathcal{L}f(x) = 0$ implies

$$f(x+1) = 2f(x) - f(x-1), \qquad f(x-1) = 2f(x) - f(x+1).$$

If $f(0), f(1)$ are given, then the value of $f(x)$ for all other x is determined uniquely by the equations above. In other words, the space of harmonic functions is a vector space of dimension 2.

For $d > 1$, the space of harmonic functions on $\mathbb{Z}^d$ is still a vector space, but it is infinite dimensional. Let us consider the case $d = 2$. For every positive number t and real r let

$$f(x_1, x_2) = f_{t,r}(x_1, x_2) = e^{rx_1} \sin(tx_2), \tag{1.24}$$

Using the sum rule for sine, we get

$$\mathcal{L}f(x_1, x_2) = \frac{1}{2} f(x_1, x_2) \left[\cosh(r) + \cos(t) - 2\right].$$

If we choose r such that

$$\cosh(r) + \cos(t) = 2,$$

then f is harmonic. So is $e^{-rx_1} \sin(tx_2)$, and hence, since linear combinations of harmonic functions are harmonic, so is

$$\sinh(rx_1) \sin(tx_2).$$

If A is a finite subset of $\mathbb{Z}^d$, then the space of functions on $\overline{A}$ that are harmonic on A has dimension $\#(\partial A)$. In fact, as we have seen, there is a linear isomorphism between this space and the set of all functions on ∂A. In Section 1.2.2, we discussed one basis for the space of harmonic functions, the Poisson kernel,

$$H_{A,y}(x) = H_A(x,y) = \mathbb{P}\{S_{T_A} = y \mid S_0 = x\}.$$

Every harmonic function f can be written as

$$f(x) = \sum_{y \in \partial A} f(y)\, H_{A,y}(x).$$

The Poisson kernel is often hard to find explicitly. For some sets A, we can find other bases that are more explicit. We will illustrate this for the square, where we use the functions (1.24) which have "separated variables", i.e., are products of functions of x_1 and functions of x_2.

Example 1.13. Let A be the square in $\mathbb{Z}^2$,

$$A = \{(x_1, x_2) : x_j = 1, \ldots, N-1\}.$$

We write $\partial A = \partial_{1,0} \cup \partial_{1,N} \cup \partial_{2,0} \cup \partial_{2,N}$ where $\partial_{1,0} = \{(0, x_2) : x_2 = 1, \ldots, N-1\}$, etc. Consider the function

$$h_j(x) = h_{j,1,N}(x) = \sinh\left(\frac{\beta_j\, x_1}{N}\right) \sin\left(\frac{j\pi x_2}{N}\right).$$

Since $\cosh(0) = 1$ and $\cosh(x)$ increases to infinity for $0 \le x < \infty$, there is a unique positive number which we call β_j such that

$$\cosh\left(\frac{\beta_j}{N}\right) + \cos\left(\frac{j\pi}{N}\right) = 2,$$

When we choose this β_j, h_j is a harmonic function. Note that h_j vanishes on three of the four parts of the boundary and

$$h_j(N, y) = \sinh(\beta_j) \sin\left(\frac{j\pi y}{N}\right).$$

If we choose $y \in \{1, \ldots, N-1\}$ and find constants $c_1, \ldots, c_{N-1}$ such that

$$\sum_{j=1}^{N-1} c_j \sinh(\beta_j) \sin\left(\frac{j\pi k}{N}\right) = \delta(y-k),$$

Then,

$$H_{(N,y)}(x) = H_{A,(N,y)}(N, x) \sum_{j=1}^{N-1} c_j\, h_j(x).$$

But we have already seen that the correct choice is

$$c_j = \frac{2}{(N-1)\sinh(\beta_j)} \sin\left(\frac{j\pi y}{N}\right).$$

Therefore,

$$H_{(N,y)}(x_1, x_2) = \tag{1.25}$$

$$\frac{2}{N-1} \sum_{j=1}^{N-1} \frac{1}{\sinh(\beta_j)} \sin\left(\frac{j\pi y}{N}\right) \sinh\left(\frac{\beta_j\, x_1}{N}\right) \sin\left(\frac{j\pi x_2}{N}\right).$$

The formula (1.25) is somewhat complicated, but there are some nice things that can be proved using this formula. Let A_N denote the square and let

$$\hat{A}_N = \left\{ (x_1, x_2) \in A : \frac{N}{4} \le x_j \le \frac{3N}{4} \right\}.$$

Note that $\hat{A}_N$ is a cube of (about) half the side length of A_N in the middle of A_N. Let $y \in \{1, \ldots, N-1\}$ and consider $H_{(N,y)}$. In Exercise 1.13 you are asked to show the following: there exist $c, c_1 < \infty$ such that the following is true for every N and every y and every $x, \tilde{x} \in \hat{A}_N$:

-

$$c^{-1}\, N^{-1} \sin(\pi y/N) \le H_{N,y}(x) \le c\, N^{-1} \sin(\pi y/N). \tag{1.26}$$

In particular,

$$H_{N,y}(x) \le c^2\, H_{N,y}(\tilde{x}). \tag{1.27}$$

-

$$|H_{N,y}(x) - H_{N,y}(\tilde{x})| \le \tag{1.28}$$

$$c_1\, \frac{|x-\tilde{x}|}{N}\, H_{n,y}(x) \le c_1\, c\, \frac{|x-\tilde{x}|}{N}\, N^{-1} \sin(\pi y/N).$$

The argument uses the explicit formula that we derive for the rectangle. Although we cannot get such a nice formula in general, we can derive two important facts. Suppose A is a finite subset of $\mathbb{Z}^2$ containing A_N. Then for $x \in A_N, z \in \partial A$,

$$H_A(x,z) = \sum_{y \in \partial A_N} H_{A_N}(x,y)\, H_A(y,z).$$

Using this and (1.27) we get for $x, \tilde{x} \in \hat{A}_N$,

$$H_A(x,z) \le c^2\, H_A(\tilde{x}, z),$$

$$|H_A(x,z) - H_A(\tilde{x},z)| \le c_1\, \frac{|x-\tilde{x}|}{N}\, H_A(x,z).$$

We can extend this to harmonic functions.

Theorem 1.14 (Difference estimates)**.** *There is a $c < \infty$ such that if A is a finite subset of $\mathbb{Z}^d$ and $F : \overline{A} \to [-M, M]$ is harmonic on A, then if $x, z \in A$ with $|z - x| = 1$,*

$$|F(z) - F(x)| \le \frac{c\,M}{\operatorname{dist}(x, \partial A)}. \tag{1.29}$$

Theorem 1.15 (Harnack principle)**.** *Suppose K is a compact subset of $\mathbb{R}^d$ and U is an open set containing K. There is a $c = c(K, U) < \infty$ such that the following holds. Suppose N is a positive integer; A is a finite subset of $\mathbb{Z}^d$ contained in $NU = \{z \in \mathbb{R}^d : z/N \in U\}$; and $\hat{A}$ is a subset of A contained in NK. Suppose $F : \overline{A} \to [0, \infty)$ is a harmonic function. Then for all $x, z \in \hat{A}$,*

$$F(x) \le c\,F(z).$$

As an application of this, let us show that: *the only bounded functions on $\mathbb{Z}^d$ that are harmonic everywhere are constants.* For $d = 1$, this is immediate from the fact that the only harmonic functions are the linear functions. For $d \ge 2$, we suppose that F is a harmonic function on $\mathbb{Z}^d$ with $|F(z)| \le M$ for all z. If $x \in \mathbb{Z}^d$ and A_R is a bounded subset of $\mathbb{Z}^d$ containing all the points within distance R of the origin, then (1.29) shows that

$$|F(x) - F(0)| \le c\,M\,\frac{|x|}{R - |x|}.$$

(Although (1.29) gives this only for $|x| = 1$, we can apply the estimate $O(|x|)$ times to get this estimate.) By letting $R \to \infty$ we see that $F(x) = F(0)$. Since this is true for every x, F must be constant.

1.5.1. Exterior Dirichlet problem. Consider the following problem. Suppose A is a cofinite subset of $\mathbb{Z}^d$, i.e., a subset such that $\mathbb{Z}^d \setminus A$ is finite. Suppose $F : \mathbb{Z}^d \setminus A \to \mathbb{R}$ is given. Find all bounded functions on $\mathbb{Z}^d$ that are harmonic on A and take on the boundary value F on $\mathbb{Z}^d \setminus A$. If $A = \mathbb{Z}^d$, then this was answered at the end of the last section; the only possible functions are constants. For the remainder of this section we assume that A is nontrivial, i.e., $\mathbb{Z}^d \setminus A$ is nonempty.

For $d = 1, 2$, there is, in fact, only a single solution. Suppose F is such a function with $L = \sup |F(x)| < \infty$. Let S_n be a simple

random walk starting at $x \in \mathbb{Z}^d$, and let $T = T_A$ be the first time n with $S_n \notin A$. If $d \le 2$, we know that the random walk is recurrent and hence $T < \infty$ with probability one. As done before, we can see that $M_n = F(S_{n\wedge T})$ is a martingale and hence

$$F(x) = M_0 = \mathbb{E}[M_n] = \mathbb{E}[F(S_T)\,1\{T \le n\}] + \mathbb{E}\left[F(S_n)\,1\{T > n\}\right].$$

The monotone convergence theorem tells us that

$$\lim_{n\to\infty} \mathbb{E}[F(S_T)\,1\{T \le n\}] = \mathbb{E}[F(S_T)].$$

Also

$$\lim_{n\to\infty} \left|\mathbb{E}\left[F(S_n)\,1\{T > n\}\right]\right| \le \lim_{n\to\infty} L\,\mathbb{P}\{T > n\} = 0.$$

Therefore,

$$F(x) = \mathbb{E}[F(S_T) \mid S_0 = x],$$

which is exactly the same solution as we had for bounded A.

If $d \ge 3$, there is more than one solution. In fact,

$$f(x) = \mathbb{P}\{T_A = \infty \mid S_0 = x\},$$

is a bounded function that is harmonic in A and equals zero on $\mathbb{Z}^d \setminus A$. The next theorem shows that this is essentially the only new function that we get. We can interpret the theorem as saying that the boundary value determines the function *if* we include ∞ as a boundary point.

Theorem 1.16. *If A is a proper cofinite subset of $\mathbb{Z}^d$ ($d \ge 3$), then the only bounded functions on $\mathbb{Z}^d$ that vanish on $\mathbb{Z}^d \setminus A$ and are harmonic on A are of the form*

$$F(x) = r\,\mathbb{P}\{T_A = \infty \mid S_0 = x\}, \quad r \in \mathbb{R}. \tag{1.30}$$

We will first consider the case $A = \mathbb{Z}^d \setminus \{0\}$ and assume that $F : \mathbb{Z}^d \to [-M, M]$ is a function satisfying $F(0) = 0$ and $\mathcal{L}F(x) = 0$ for $x \ne 0$. Let $\alpha = \mathcal{L}F(0)$ and let

$$f(x) = F(x) + \alpha\, G(x, 0).$$

Then f is a bounded harmonic function and hence must be equal to a constant. Since $G(x, 0) \to 0$ as $x \to \infty$, the constant must be r and hence

$$F(x) = r - \alpha\, G(x, 0) =$$

$$r\,\mathbb{P}\{\tau_0 = \infty \mid S_0 = x\} + \mathbb{P}\{\tau_0 < \infty \mid S_0 = x\}[r - \alpha G(0,0)].$$

Since $F(0) = 0$ and $\mathbb{P}\{\tau_0 = \infty \mid S_0 = 0\} = 0$, we know that $r - \alpha G(0,0) = 0$ and hence F is of the form (1.30).

For other cofinite A, assume F is such a function with $|F| \le 1$. Then F satisfies

$$\mathcal{L}F(x) = -g(x), \quad x \in A$$

for some function g that vanishes on A. In particular,

$$f(x) = F(x) + \sum_{y \in \mathbb{Z}^d \setminus A} G(x,y)\, g(x),$$

is a bounded harmonic function (why is it bounded?) and hence constant. This tells us that there is an r such that

$$F(x) = r - \sum_{y \in \mathbb{Z}^d \setminus A} G(x,y)\, g(x),$$

which implies, in particular, that $F(x) \to r$ as $r \to \infty$. Also, if $x \in \mathbb{Z}^d \setminus A$, $F(x) = 0$ which implies

$$\sum_{y \in \mathbb{Z}^d \setminus A} G(x,y)\, g(x) = r.$$

If we show that $G(x,y)$ is invertible on $\mathbb{Z}^d \setminus A$, then we know there is a unique solution to this equation, which would determine g, and hence F.

To do this, assume $\#(\mathbb{Z}^d \setminus A) = K$; let $\tilde{T}_A = \min\{n \ge 1 : S_n \notin A\}$; and for $x, y \in \mathbb{Z}^d \setminus A$, we define

$$J(x,y) = \mathbb{P}\{\tilde{T}_A < \infty, S_{\tilde{T}_A} = y \mid S_0 = x\}.$$

Then J is a $K \times K$ matrix. In fact (Exercise 1.25),

$$(J - I)\, G = -I. \tag{1.31}$$

In particular, G is invertible.

1.6. Exercises

Exercise 1.1. Suppose that $X_1, X_2, \ldots$ are independent, identically distributed random variables such that

$$\mathbb{E}[X_j] = 0, \quad \mathbb{P}\{|X_j| > K\} = 0,$$

for some $K < \infty$.

- Let $M(t) = \mathbb{E}[e^{tX_j}]$ denote the moment generating function of X_j. Show that for every $t > 0, \epsilon > 0$,
$$\mathbb{P}\{X_1 + \cdots + X_n \geq \epsilon n\} \leq [M(t)\, e^{-t\epsilon}]^n.$$
- Show that for each $\epsilon > 0$, there is a $t > 0$ such that $M(t)\, e^{-\epsilon t} < 1$. Conclude the following: for every $\epsilon > 0$, there is a $\rho = \rho(\epsilon) < 1$ such that for all n
$$\mathbb{P}\{|X_1 + \cdots + X_n| \geq \epsilon n\} \leq 2\, \rho^n.$$
- Show that we can prove the last result with the boundedness assumption replaced by the following: there exists a $\delta > 0$ such that for all $|t| < \delta$, $\mathbb{E}[e^{tX_j}] < \infty$.

Exercise 1.2. Prove the following: there is a constant γ (called Euler's constant) and a $c < \infty$ such that for all positive integers n,
$$\left| \left(\sum_{j=1}^{n} \frac{1}{j} \right) - \gamma - \log n \right| \leq \frac{c}{n}.$$
Hint: write
$$\log\left(n + \frac{1}{2}\right) - \log\left(\frac{1}{2}\right) = \int_{\frac{1}{2}}^{n+\frac{1}{2}} \frac{1}{x}\, dx,$$
and estimate
$$\left| \frac{1}{j} - \int_{j-\frac{1}{2}}^{j+\frac{1}{2}} \frac{dx}{x} \right|.$$

Exercise 1.3. Show that there is a $c > 0$ such that the following is true. For every real number r and every integer n,
$$e^{-cr^2/n} \leq e^r \left(1 - \frac{r}{n}\right)^n \leq e^{cr^2/n}. \tag{1.32}$$

Exercise 1.4. Find constants a_1, a_2 such that the following is true as $n \to \infty$,
$$\left(1 - \frac{1}{n}\right)^n = e^{-1} \left[1 + \frac{a_1}{n} + \frac{a_2}{n^2} + O\left(n^{-3}\right)\right].$$

Exercise 1.5. Let S_n be a one-dimensional simple random walk and let
$$p_n = \mathbb{P}\{S_{2n} = 0 \mid S_0 = 0\}.$$

- Show that

$$p_{n+1} = p_n \frac{2n+1}{2n+2}, \tag{1.33}$$

and hence

$$p_n = \frac{1 \cdot 3 \cdot 5 \cdots (2n-1)}{2 \cdot 4 \cdot 6 \cdots (2n)}.$$

- Use the relation (1.33) to give another proof that there is a c such that as $n \to \infty$

$$p_n \sim \frac{c}{\sqrt{n}}.$$

(Our work in this chapter shows in fact that $c = 1/\sqrt{\pi}$, but you do not need to prove this here.)

Exercise 1.6.

- Show that if X is a nonnegative random variable, then

$$\lim_{n\to\infty} \mathbb{E}[X\, 1\{X \le n\}] = \lim_{n\to\infty} \mathbb{E}[X \wedge n] = \mathbb{E}[X].$$

- (Monotone Convergence Theorem) Show that if $0 \le X_1 \le X_2 \le \cdots$, then

$$\mathbb{E}\left[\lim_{n\to\infty} X_n\right] = \lim_{n\to\infty} \mathbb{E}[X_n].$$

In both parts, the limits and the expectations are allowed to take on the value infinity.

Exercise 1.7. Prove Theorem 1.6.

Exercise 1.8. Suppose $X_1, X_2, \ldots$ are independent random variables each of whose distribution is symmetric about 0. Show that for every $a > 0$,

$$\mathbb{P}\left\{\max_{1\le j\le n} X_1 + \cdots + X_j \ge a\right\} \le 2\,\mathbb{P}\{X_1 + \cdots + X_n \ge a\}.$$

(Hint: Let K be the smallest j with $X_1 + \cdots + X_j \ge a$ and consider

$$\mathbb{P}\{X_1 + \cdots + X_n \ge a \mid K = j\}. \ )$$

Exercise 1.9. Suppose X is a random variable taking values in $\mathbb{Z}$. Let

$$\phi(t) = \mathbb{E}[e^{itX}] = \mathbb{E}[\cos(tX)] + i\,\mathbb{E}[\sin(tX)] = \sum_{x\in\mathbb{Z}} e^{itx}\,\mathbb{P}\{X = x\},$$

be its *characteristic function*. Prove the following facts.

- $\phi(0) = 1$, $|\phi(t)| \le 1$ for all t and $\phi(t + 2\pi) = \phi(t)$.
- If the distribution of X is symmetric about the origin, then $\phi(t) \in \mathbb{R}$ for all t.
- For all integers x,
$$\mathbb{P}\{X = x\} = \frac{1}{2\pi} \int_{-\pi}^{\pi} \phi(t)\, e^{-ixt}\, dt.$$
- Let k be the greatest common divisor of the set of integers n with $\mathbb{P}\{|X| = n\} > 0$. Show that $\phi(t + (2\pi/k)) = \phi(t)$ and $|\phi(t)| < 1$ for $0 < t < (2\pi/k)$.
- Show that ϕ is a continuous (in fact, uniformly continuous) function of t.

Exercise 1.10. Suppose $X_1, X_2, \ldots$ are independent, identically distributed random variables taking values in the integers with characteristic function ϕ. Let $S_n = X_1 + \cdots + X_n$. Suppose that the distribution of X_j is symmetric about the origin, $\mathrm{Var}[X_j] = \mathbb{E}[X_j^2] = \sigma^2, \mathbb{E}[|X_j|^3] < \infty$. Also assume,

$$\mathbb{P}\{X_j = 0\} > 0, \quad \mathbb{P}\{X_j = 1\} > 0.$$

The goal of this exercise is to prove

$$\lim_{n\to\infty} \sqrt{2\pi\sigma^2 n}\, \mathbb{P}\{S_n = 0\} = 1.$$

Prove the following facts.

- The characteristic function of $X_1 + \cdots + X_n$ is ϕ^n.
- For every $0 < \epsilon \le \pi$ there is a $\rho < 1$ such that $|\phi(t)| < \rho$ for $\epsilon \le |t| \le \pi$.
- $$\mathbb{P}\{S_n = 0\} = \frac{1}{2\pi} \int_{-\pi}^{\pi} \phi(t)^n\, dt = \frac{1}{2\pi\sqrt{n}} \int_{-\pi\sqrt{n}}^{\pi\sqrt{n}} \phi(t/\sqrt{n})^n\, dt.$$
- There is a c such that for $|t| \le \pi$,
$$\left| \phi(t) - 1 - \frac{\sigma^2 t^2}{2} \right| \le c\, t^3.$$

- $$\lim_{n\to\infty} \int_{-\pi\sqrt{n}}^{\pi\sqrt{n}} \phi(t/\sqrt{n})^n \, dt = \int_{-\infty}^{\infty} e^{-\sigma^2 t^2/2} \, dt = \frac{\sqrt{2\pi}}{\sigma}.$$

  Hint: you will probably want to use (1.32).

Exercise 1.11. Suppose A is a finite subset of $\mathbb{Z}^d$ and

$$F : \partial A \to \mathbb{R}, \qquad g : A \to \mathbb{R}$$

are given functions. Show that there is a unique extension of F to $\overline{A}$ such that

$$\mathcal{L}F(x) = -g(x), \quad x \in A.$$

Give a formula for F.

Exercise 1.12. Suppose A is a finite subset of $\mathbb{Z}^d$ and

$$F : \partial A \to \mathbb{R}, \qquad f : A \to \mathbb{R}$$

are given functions. Show that there is a unique function $p_n(x), n = 0, 1, 2 \ldots, x \in \overline{A}$ satisfying the following:

$$p_n(x) = F(x), \quad x \in \partial A,$$
$$\partial p_n(x) = \mathcal{L}F(x), \quad x \in A.$$

Show that $p(x) = \lim_{n\to\infty} p_n(x)$ exists and describe the limit function p.

Exercise 1.13. Prove (1.26) and (1.28).

Exercise 1.14. Find the analogue of the formula (1.25) for the d-dimensional cube

$$A = \{(x_1, \ldots, x_d) \in \mathbb{Z}^d : x_j = 1, \ldots, N-1\}$$

Exercise 1.15. Suppose F is a harmonic function on $\mathbb{Z}^d$ such that

$$\lim_{|x|\to\infty} \frac{|F(x)|}{|x|} = 0.$$

Show that F is constant.

Exercise 1.16. The *relaxation method* for solving the Dirichlet problem is the following. Suppose A is a bounded subset of $\mathbb{Z}^d$ and $F : \partial A \to \mathbb{R}$ is a given function. Define the functions $F_n(x), x \in \overline{A}$ as follows.

$$F_n(x) = F(x) \text{ for all } n \text{ if } x \in \partial A.$$

$$F_0(x), x \in A, \text{ is defined arbitrarily ,}$$

and for $n \geq 0$,

$$F_{n+1}(x) = \frac{1}{2d} \sum_{|x-y|=1} F_n(y), \qquad x \in A.$$

Show that for any choice of initial function F_0 on A,

$$\lim_{n\to\infty} F_n(x) = F(x), \qquad x \in A,$$

where F is the solution to the Dirichlet problem with the given boundary value. (Hint: compare this to Exercise 1.12.)

Exercise 1.17. Let S_n denote a d-dimensional simple random walk and let $R_n^1, \ldots, R_n^d$ denote the number of steps taken in each of the d components. Show that for all $n > 0$, the probability that $R_{2n}^1, \ldots, R_{2n}^d$ are all even is $2^{-(d-1)}$.

Exercise 1.18. Suppose that S_n is a biased one-dimensional random walk. To be more specific, let $p > 1/2$ and

$$S_n = X_1 + \cdots + X_n,$$

where $X_1, \ldots, X_n$ are independent with

$$\mathbb{P}\{X_j = 1\} = 1 - \mathbb{P}\{X_j = -1\} = p.$$

Show that there is a $\rho < 1$ such that as $n \to \infty$,

$$\mathbb{P}\{S_{2n} = 0\} \sim \rho^n \frac{1}{\sqrt{\pi n}}.$$

Find ρ explicitly. Use this to show that with probability one the random walk does not return to the origin infinitely often.

Exercise 1.19. Suppose δ_n is a sequence of real numbers with $|\delta_n| < 1$ and such that

$$\sum_{j=1}^{\infty} |\delta_n| < \infty.$$

Let

$$s_n = \prod_{j=1}^{n} (1 + \delta_j).$$

Show that the limit $s_\infty = \lim_{n\to\infty} s_n$ exists and is strictly positive. Moreover, there exists an N such that for all $n \geq N$,

$$\left|1 - \frac{s_n}{s_\infty}\right| \leq 2 \sum_{j=n+1}^{\infty} |\delta_j|.$$

Exercise 1.20. Find the number t such that

$$n! = \sqrt{2\pi}\, n^{n+\frac{1}{2}}\, e^{-n} \left[1 + \frac{t}{n} + O(n^{-2})\right].$$

Exercise 1.21. Prove that

$$\int_{-\infty}^{\infty} e^{-x^2/2}\, dx = \sqrt{2\pi}.$$

Hint: there are many ways to do this but direct antidifferentiation is not one of them. One approach is to consider the square of the left hand side, write it as a double (iterated) integral, and then use polar coordinates.

Exercise 1.22. Suppose S_n is a simple random walk in $\mathbb{Z}^d$ and $A \subset \mathbb{Z}^d$ is finite with N points. Let T_A be the smallest n such that $S_n \notin A$. Show that

$$\mathbb{P}\{T_A > kN\} \leq \left(1 - \frac{1}{2d}\right)^k.$$

Exercise 1.23. Finish the proof of Theorem 1.9 by doing the following.

- Use connectedness of A to show that any nonzero eigenfunction ϕ with every component nonnegative must actually have every component strictly positive.
- Give an example of a disconnected A such that λ_1 has multiplicity greater than one.
- Given an example of a disconnected A such that λ_1 has multiplicity one. Does Theorem 1.9 hold in this case?

Exercise 1.24. Suppose A is a bounded subset of $\mathbb{Z}^d$. We call $\{x, y\}$ an *edge* of $\overline{A}$ if $x, y \in \overline{A}$, $|x - y| = 1$ and at least one of x, y is in A. If $F : \overline{A} \to \mathbb{R}$ is a function, we define its energy by

$$\mathcal{E}(f) = \sum [f(x) - f(y)]^2,$$

where the sum is over the edges of $\overline{A}$. For any $F : \partial A \to \mathbb{R}$, define $\mathcal{E}(F)$ to be the infimum of $\mathcal{E}(f)$ where the infimum is over all f on $\overline{A}$ that agree with F on ∂A. Show that if f agrees with F on ∂A, then $\mathcal{E}(f) = \mathcal{E}(F)$ if and only if f is harmonic in A.

Exercise 1.25. Verify (1.31).

Exercise 1.26. We will construct a "tree" each of whose vertices has three neighbors. We start by constructing $\mathcal{T}_1$ as follows: the vertices of $\mathcal{T}_1$ are the "empty word", denoted by o, and all finite sequences of the letters a, b, i.e., "words" $x_1 \cdots x_n$ where $x_1, x_2, \ldots, x_n \in \{a, b\}$. Both words of one letter are adjacent to o. We say that a word of length $n-1$ and of length n are adjacent if they have the exact same letters, in order, in the first $n-1$ positions. Note that each word of positive length is adjacent to three words and the root is adjacent to only two words. We construct another tree $\mathcal{T}_2$ similarly, calling the root $\tilde{o}$ and using the letters $\tilde{a}, \tilde{b}$. Finally we make a tree $\mathcal{T}$ by taking the union of $\mathcal{T}_1$ and $\mathcal{T}_2$ and adding one more connection: we say that o and $\tilde{o}$ are adjacent.

- Convince yourself that $\mathcal{T}$ is a connected tree, i.e., between any two points of $\mathcal{T}$ there is a unique path in the tree that does not go through any point more than once.
- Let S_n denote simple random walk on the tree, i.e., the process that at each step chooses one of the three nearest neighbors at random, each with probability 1/3, with the choice being independent of all the previous moves. Show that S_n is transient, i.e., with probability one S_n visits the origin only finitely often. (Hint: Exercise 1.18 could be helpful.)
- Show that with probability one the random walk does one of the two following things: either the random walk visits $\mathcal{T}_1$ only finitely often or it visits $\mathcal{T}_2$ only finitely often. Let $f(x)$ be the probability that the walk visits $\mathcal{T}_1$ finitely often. Show that f is a nonconstant bounded harmonic function. (A function f on $\mathcal{T}$ is harmonic if for every $x \in \mathcal{T}$, $f(x)$ equals the average value of f on the nearest neighbors of x.)

- Consider the space of bounded harmonic functions on $\mathcal{T}$. Show that this is an infinite dimensional vector space.

Exercise 1.27. Show that if $A \subset A_1$ are two subsets of $\mathbb{Z}^d$, then $\lambda_A \leq \lambda_{A_1}$. Show that if A_1 is connected and $A \neq A_1$, then $\lambda_A < \lambda_{A_1}$. Give an example with A_1 disconnected and A a strict subset of A_1 for which $\lambda_A = \lambda_{A_1}$.

Exercise 1.28. Consider $\beta(j, N), j = 1, \ldots, N-1$ where $\beta(j, N)$ is the unique positive number satisfying

$$\cosh\left(\frac{\beta(j,N)}{N}\right) + \cos\left(\frac{j\pi}{N}\right) = 2.$$

Prove the following estimates. The constants c_1, c_2, c_3 are positive constants independent of N and the estimates should hold for all N and all $j = 1, \ldots, N-1$.

- $$\beta(1,N) < \beta(2,N) < \cdots < \beta(N-1,N) \leq N \cosh^{-1}(2).$$
- There is a c_1 such that
$$\left|\cosh\left(\frac{j\pi}{N}\right) + \cos\left(\frac{j\pi}{N}\right) - 2\right| \leq \frac{c_1 j^4}{N^4}.$$
- There is a c_2 such that
$$|\beta(j,N) - \pi j| \leq \frac{c_2 j^4}{N^3}.$$
- There is a c_3 such that
$$\beta(j,N) \geq c_3 j.$$

Chapter 2

Brownian Motion and the Heat Equation

2.1. Brownian motion

In this section we introduce *Brownian motion*. Brownian motion can be considered as the limit of random walk as the time and space increments go to zero. It is not obvious how to take such a limit or even what is meant by the word limit. Rather than worry about these details for the moment, we will instead assume some kind of limit of simple random walk exists and list some properties that the limit should have. We will start with the one-dimensional case. The process will be defined in continuous time and continuous space — we let W_t be the position of the Brownian motion (continuous random walker) at time t.

$\diamondsuit$ Brownian motion is also called the Wiener process. For mathematicians the terms Brownian motion and Wiener process are synonymous. In scientific literature, the term Brownian motion is often used for a physical process for which there are several different mathematical models one of which is the one we discuss here. The term Wiener process always refers to the process we describe here.

Brownian motion is an example of a continuous *stochastic process*. A stochastic process is a collection of random variables $\{W_t\}$ indexed by time. In our case time runs over the nonnegative reals. (The simple random walk is a stochastic process with time indexed by the nonnegative integers). This collection of random variables can also be viewed as a random function

$$t \longmapsto W_t.$$

We will work up to a definition of Brownian motion by considering a discrete approximation. Suppose we have small steps in time and space so that in time Δt the typical change in space is of order Δx. Simple random walk S_n on $\mathbb{Z}$ is a process with $\Delta t = 1$ and $\Delta x = 1$; in each integer time step the process moves distance one. Let us take the random walk path and change the time and space increments. Suppose that the time increments are $\Delta t = \delta = 1/N$ where N is a large integer. Let us see how we should change the spatial steps Δx.

For large N the limit process should look like

$$W_{k\delta} \approx \Delta x \, S_k,$$

where Δx denotes the spatial scaling factor. We need to figure out what Δx should be in terms of δ so that the process scales correctly. Let us normalize the limit process so that $\mathbb{E}[W_1^2] = 1$. Since

$$\mathbb{E}\left[(\Delta x S_N)^2\right] = (\Delta x)^2 \, \mathbb{E}[S_N^2] = (\Delta x)^2 \, N$$

we see that the right scaling is $\Delta x = 1/\sqrt{N} = \sqrt{\delta}$. Therefore, if $t = j\delta = j/N$, we write

$$W_t = W_{j/N} \approx \frac{S_j}{\sqrt{N}}.$$

$\diamondsuit$ The above may seem a little strange at first. In N time steps of size $1/N$, the process has taken N spatial steps of size $1/\sqrt{N}$. It may seem that the process will have gone distance $\sqrt{N}$. However, about half of the steps are in the positive direction and about half in the negative direction so that the net distance turns out to be of order one.

♢ If $X_1, X_2, \ldots$ are independent random variables with mean μ and variance σ^2, the *central limit theorem* states that for large n the distribution of

$$Z_n = \frac{(X_1 + \cdots + X_n) - n\mu}{\sqrt{\sigma^2 n}}$$

is approximately that of a normal distribution with mean 0 and variance 1. More precisely, for every $a < b$,

$$\lim_{n\to\infty} \mathbb{P}\{a \le Z_n \le b\} = \int_a^b \frac{1}{\sqrt{2\pi}}\, e^{-x^2/2}\, dx.$$

Let us write

$$W_t = W_{j\delta} \approx \frac{S_j}{\sqrt{N}} = \frac{S_j}{\sqrt{j}} \sqrt{\frac{j}{N}} = \frac{S_j}{\sqrt{j}}\, \sqrt{t}.$$

The central limit theorem tells us that as $j \to \infty$, the distribution of $S_j/\sqrt{j}$ approaches a normal distribution with mean zero and variance one. Therefore $\sqrt{t}\, S_j/\sqrt{j}$ approaches a normal distribution with mean zero and variance t. Recall such a random variable has density

$$\frac{1}{\sqrt{2\pi t}}\, e^{-\frac{x^2}{2t}}, \quad -\infty < x < \infty.$$

(It is easy to check that if Y has a normal distribution with variance 1, then $\sigma\, Y$ has a normal distribution with mean 0 and variance σ^2.) Using this as a motivation, we have the first property for our definition of Brownian motion.

- For each t, the random variable W_t has a normal distribution with mean zero and variance t.

In fact, we can do the same argument for $W_t - W_s$ to derive the following sometimes referred to as *identically distributed normal increments*.

- For $0 \le s < t < \infty$, the random variable $W_t - W_s$ has a normal distribution with mean 0 and variance $t - s$.

The next property is called *independent increments*. If S_n is a simple random walk and $n < m$, then the random variable $S_m - S_n$ is independent of the random walk up to time n. We expect this property to hold in the limit.

- For all $s < t$, the random variable $W_t - W_s$ is independent of all the random variables $\{W_r : r \le s\}$.

♢ Note that we are not saying that the positions W_s and W_t are independent for $s < t$. In fact, if s and t are close, we expect W_s and W_t to be close. It is the increments, W_s and $W_t - W_s$ that are independent.

A major technical problem in defining a stochastic process such as W_t is that there are uncountably many positive real numbers t. It is easier to consider a process defined on only a countable set of times. If we choose a countable set that is dense in the set of all times, then we would hope this is good enough. We can restrict ourselves at the moment to rational times, and we might as well restrict ourselves even more to a subset of these, the *dyadic rationals*. Let $\mathcal{D}_n$ denote the set of nonnegative rational numbers with denominator 2^n (not necessarily in reduced form):

$$\mathcal{D}_n = \left\{ \frac{k}{2^n} : k = 0, 1, 2, \ldots \right\}.$$

Then $\mathcal{D}_0 \subset \mathcal{D}_1 \subset \mathcal{D}_2 \subset \cdots$. The set of (nonnegative) dyadic rationals is $\mathcal{D} = \cup_n \mathcal{D}_n$.

We combine our assumptions so far into a definition.

Definition 2.1. A *(standard one-dimensional) Brownian motion* on the dyadic rationals is a collection of random variables $\{W_t : t \in \mathcal{D}\}$ satisfying:

- For each n, the random variables

  $$W_{k/2^n} - W_{(k-1)/2^n}, \quad k = 1, 2, \ldots$$

  are independent normal random variables with mean zero and variance 2^{-n}.

If we were only interested in the position of the Brownian motion at the times $1/2^n$, $2/2^n$, $3/2^n$, ..., then we could consider this as a random walk in time increments of size 2^{-n}. The spatial increments would not be two-valued as in the case for simple random walk but rather would have normal distributions.

Let us write

$$J(k,n) = 2^{n/2}\left[W_{k/2^n} - W_{(k-1)/2^n}\right]. \tag{2.1}$$

Then another way of phrasing the condition is to say that for each n, the random variables

$$J(1,n), J(2,n), J(3,n), \ldots$$

are independent normal random variables each with mean zero and variance one.

We have stated the requirements on $W_t, t \in \mathcal{D}$ to be a Brownian motion. In order to guarantee that the definition is at all useful, we need to show that such a process exists. We prove this in Section 2.7.1. This section can be skipped if one is content to believe that such random variables can be found. Using the definition one can show

- If $s, t \in \mathcal{D}$ with $0 \le s \le t$, then $W_t - W_s$ is independent of $\{W_r : r \le s\}$ and has a normal distribution with mean zero and variance $t - s$.

The dyadic rationals $\mathcal{D}$ have the property that if j is an integer, then

$$2^j\,\mathcal{D} := \{2^j q : q \in \mathcal{D}\} = \mathcal{D}.$$

If W_t is a Brownian motion and we scale time by 2^j, then we get another Brownian motion provided that we scale space by $1/\sqrt{2^j}$. To be more precise: If $W_q, q \in \mathcal{D}$ is a standard Brownian motion, $j \in \mathbb{Z}$, and

$$\tilde{W}_q = 2^{-j/2}\, W_{2^j q},$$

then $\tilde{W}_q$ is also a standard Brownian motion. This is checked by verifying that $\tilde{W}_q$ satisfies the properties that define a Brownian motion (Exercise 2.1).

The times $\mathcal{D}$ are dense in the positive reals. If we know the value of a function W_t at all times t in a dense subset, do we know W_t at all times? The answer is yes, *if* W_t is a continuous function of t. In order to show that W_t is a continuous function it suffices to show that W_t restricted to $t \in \mathcal{D}$ is *uniformly continuous* on each compact interval.

♢ Suppose $f : \mathcal{D} \to \mathbb{R}$ is a function and suppose that on every bounded interval $[a, b]$, f is uniformly continuous. This means that for every $\epsilon > 0$, there is a $\delta > 0$ such that if $|s - t| < \delta$ and $s, t \in \mathcal{D}$, then $|f(s) - f(t)| < \epsilon$. Since $\mathcal{D}$ is a dense subset of $[0, \infty)$, if $a \leq t \leq b$ we can find a sequence of $t_n \in \mathcal{D} \cap [a, b]$ with $t_n \to t$. Since f is uniformly continuous on $[a, b]$, it is easy to check that $\{f(t_n)\}$ is a Cauchy sequence and hence has a unique limit. If we define $f(t), t \in \mathbb{R}$ as the limit, then it is also easy to see that f is uniformly continuous on $[a, b]$. We summarize: *if f is a function on $\mathcal{D}$ that is uniformly continuous on each bounded interval $[a, b]$, then there is a unique extension of f to $\mathbb{R}$ that is continuous.*

For Brownian motion on the dyadics, uniform continuity holds with probability one.

Theorem 2.2. *If $W_q, q \in \mathcal{D}$ is a standard one-dimensional Brownian motion, then with probability one, for every interval $[a, b]$, the function W_q is uniformly continuous. In particular, $W_t, t \in [0, \infty)$ can be defined by continuity so it has the following properties:*

- $W_0 = 0$.
- *For every $0 \leq s \leq t$, $W_t - W_s$ has a normal distribution with mean zero and variance $t - s$. Moreover, $W_t - W_s$ is independent of $\{W_r : r \leq s\}$.*
- *With probability one, the function $t \mapsto W_t$ is a continuous function.*

♢ There is a simple heuristic reason why W_t should be continuous. Consider $W_{t+\delta} - W_t$. Since this random variable has mean zero and variance δ, $\mathbb{E}[(W_{t+\delta} - W_t)^2] = \delta$. In other words, one expects that

$$(W_{t+\delta} - W_t)^2 \approx \delta, \quad |W_{t+\delta} - W_t| \approx \sqrt{\delta}.$$

As $\delta \to 0$, we have $\sqrt{\delta} \to 0$, and so we expect continuity. This heuristic argument is nice but in order to make it rigorous we need to show that not only is the average value of $(W_{t+\delta} - W_t)^2$ of order δ but in some sense it is always of that order (or not too much bigger).

The hardest part of the proof of this theorem is establishing the uniform continuity. This will require making some estimates, or as mathematicians sometimes say, getting our hands dirty. We will do the proof for the interval $[0,1]$; other intervals can be handled similarly. Let

$$K_n^* = \sup\left\{|W_s - W_t| : 0 \le s,t \le 1, |s-t| \le 2^{-n}, s,t \in \mathcal{D}\right\}. \tag{2.2}$$

Then, uniform continuity of the function $t \mapsto W_t$ is equivalent to the statement that $K_n^* \to 0$ as $n \to \infty$. (Verify this if this is not immediate to you.) A slightly different quantity is easier to estimate,

$$K_n = \max_{k=1,\ldots,2^n} \sup\left\{|W_q - W_{(k-1)/2^n}| : q \in \mathcal{D}, \frac{k-1}{2^n} \le q \le \frac{k}{2^n}\right\}. \tag{2.3}$$

The difference is that we require one (but not both) of the times to be in $\mathcal{D}_n$. Using the triangle inequality, we can see that

$$K_n \le K_n^* \le 3K_n,$$

and hence it is equivalent to show that $K_n \to 0$. In Section 2.7.2 we give some sharp estimates for the probability that K_n is large. In particular, we show that

$$\sum_{n=1}^{\infty} \mathbb{P}\{K_n \ge 2\sqrt{n}\,2^{-n/2}\} < \infty. \tag{2.4}$$

The Borel-Cantelli lemma (Lemma 1.3) then implies that with probability one, the estimate

$$K_n < 2\sqrt{n}\,2^{-n/2}$$

holds for all n sufficiently large. In particular, $K_n \to 0$.

Brownian motion defines a random function $t \mapsto W_t$. We can ask: how smooth is such a function? Suppose we try to take a derivative at a point t. The definition of the derivative is

$$\frac{dW_t}{dt} = \lim_{\delta \to 0} \frac{W_{t+\delta} - W_t}{\delta},$$

provided that the limit exists. The typical magnitude of the numerator on the right hand side is $|W_{t+\delta} - W_t| \approx \sqrt{\delta}$ which is much larger than δ if δ is small. From this we see that we do not expect the derivative to exist very often, and, in fact, it never exists.

Theorem 2.3. *With probability one, the function $t \mapsto W_t$ is nowhere differentiable.*

It is not easy to write down even one function that is continuous but nowhere differentiable, but this tells us that this is always true for the Brownian motion. We have given the intuition for this theorem already. We leave the details of the proof as Exercise 2.26.

♢ While it may appear surprising that we are getting functions that are not differentiable, one can actually see that our initial assumptions imply that we have nondifferentiable functions. Suppose W_t were differentiable at t_0 with derivative m. Then one could determine m by looking at W_t for $t \leq t_0$. We would know that for small $t > t_0$, $W_t - W_{t_0} \approx m\,(t - t_0)$. But one of our assumptions is that the increment $W_t - W_{t_0}$ is independent of the values of the Brownian motion before time t_0.

To define a Brownian motion in $\mathbb{R}^d$, we take d independent Brownian motions $W_t^1, \ldots, W_t^d$ and let $W_t = (W_t^1, \ldots, W_t^d)$. The density of W_t is

$$(2\pi t)^{-d/2} \exp\left\{\frac{x_1^2 + \cdots + x_d^2}{2t}\right\}.$$

Note that the density of W_t is radially symmetric. (The fact that independent variables in each coordinate give something that is radially symmetric may be surprising. In fact, the normal distribution is in some sense the only distribution with this property, see Exercise 2.4.) We will use the fact that the d-dimensional Brownian motion is invariant under rotations.

Suppose W_t is a Brownian motion starting with $W_0 = x$ and U is an open subset of $\mathbb{R}^d$. Then $\mathbb{E}[W_t] = x$. We need a "stopped" version of this result. Let

$$T = T_U = \inf\{t \geq 0 : W_t \notin U\}.$$

The random variable T is an example of a *stopping time*; in this case, we stop when we leave U. The term stopping time implies that the decision whether or not to stop at a particular time is made using only the information available at that time without looking into the future. Recall that $t \wedge T = \min\{t, T\}$.

Proposition 2.4. *For every $t \geq 0$,*

$$\mathbb{E}[W_{t\wedge T} \mid W_0 = x] = x. \tag{2.5}$$

Proof. We will prove this for $t = 1$ and $x = 0$; the proof for other values of t, x is similar. For each positive integer n, let

$$T_n = \min\{q \in \mathcal{D}_n : q \geq T\}.$$

In other words, $T_n = k/2^n$ if $(k-1)/2^n \leq T < k/2^n$. We will first show that for each n,

$$\mathbb{E}[W_{1\wedge T_n}] = 0. \tag{2.6}$$

Indeed, we can write

$$W_{1\wedge T_n} = \sum_{k=1}^{2^n} 1\left\{T > \frac{k-1}{2^n}\right\} [W_{k/2^n} - W_{(k-1)/2^n}],$$

where again we use the indicator function notation. Hence,

$$\mathbb{E}[W_{1\wedge T_n}] = \sum_{k=1}^{2^n} \mathbb{E}\left[1\left\{T > \frac{k-1}{2^n}\right\} [W_{k/2^n} - W_{(k-1)/2^n}]\right].$$

The event $T > (k-1)/2^n$ can be determined by observing W_t for $t \leq (k-1)/2^n$. Therefore, $W_{k/2^n} - W_{(k-1)/2^n}$ is independent of this event and

$$\mathbb{E}\left[1\left\{T > \frac{k-1}{2^n}\right\} [W_{k/2^n} - W_{(k-1)/2^n}]\right] =$$

$$\mathbb{P}\left\{T > \frac{k-1}{2^n}\right\} \mathbb{E}\left[W_{k/2^n} - W_{(k-1)/2^n}\right] = 0.$$

By summing, we get (2.6). Finally, note that

$$|W_{1\wedge T_n} - W_{1\wedge T}| \leq K_n \to 0,$$

where K_n is as defined in (2.3). Using this we get the proposition. □

2.2. Harmonic functions

Recall that a function f on $\mathbb{Z}^d$ is harmonic at x if $f(x)$ equals the average of f on its nearest neighbors. If U is an open subset of $\mathbb{R}^d$, we will say that f is *harmonic* in U if and only if it continuous and satisfies the following *mean value property*: for every $x \in U$, and every $0 < \epsilon < \text{dist}(x, \partial U)$,

$$f(x) = MV(f; x, \epsilon) = \int_{|y-x|=\epsilon} f(y)\, ds(y). \tag{2.7}$$

This definition includes a number of undefined quantities so we will now define them. First, $\text{dist}(x, \partial U)$ denotes the distance from x to the boundary of U which can be defined as $\inf\{|x - y| : y \notin U\}$. The s in the integral refers to surface measure on the sphere of radius ϵ about x normalized so that

$$\int_{|y-x|=\epsilon} 1\, ds(y) = 1.$$

Here $MV(f; x, \epsilon)$ stands for the mean value of f on the sphere of radius ϵ about x.

♢ If $d = 3$, s is a constant times the usual surface area. For $d > 3$, it is the analogous $(d-1)$-dimensional "volume".

In the case $d = 1$, f is harmonic in (a, b) if it is continuous and for each x, ϵ with $a < x - \epsilon < x < x + \epsilon < b$,

$$f(x) = \frac{1}{2}\left[f(x + \epsilon) + f(x - \epsilon)\right]. \tag{2.8}$$

Linear functions $f(x) = mx + r$ satisfy this equation. The next proposition shows that these are the only harmonic functions in $\mathbb{R}$.

Proposition 2.5. *If $f : (a, b) \to \mathbb{R}$ is harmonic, then*

$$f(x) = mx + r$$

for some m, r.

Proof. We will assume $a < 0, b > 1$ and $f(0) = 0, f(1) = 1$. We will show $f(x) = x$ for $0 \le x \le 1$. The general proof works similarly. Using (2.8) with $\epsilon = 1/2$ gives

$$f\left(\frac{1}{2}\right) = \frac{f(0) + f(1)}{2} = \frac{1}{2}.$$

Similarly, by iterating (2.8), letting ϵ range over the dyadic rationals $\mathcal{D} \cap [0, 1]$, we can see that for every dyadic rational $f(j/2^n) = j/2^n$. Since f is continuous, we must have $f(t) = t$ for all t. □

Even though the last proof was easy, we will give another proof of the last proposition. Suppose that f has two continuous derivatives in a neighborhood of x. Then Taylor's theorem gives

$$f(x + \epsilon) = f(x) + \epsilon\, f'(x) + \frac{\epsilon^2}{2} f''(x) + o(\epsilon^2),$$

$$f(x - \epsilon) = f(x) - \epsilon\, f'(x) + \frac{\epsilon^2}{2} f''(x) + o(\epsilon^2),$$

where $o(\epsilon^2)$ denotes a function (depending on x) such that $o(\epsilon^2)/\epsilon^2 \to 0$ as $\epsilon \to 0$. If we add the two equations and let $\epsilon \to 0$, we get

$$f''(x) = \lim_{\epsilon \to 0} \frac{f(x + \epsilon) + f(x - \epsilon) - 2f(x)}{\epsilon^2}. \tag{2.9}$$

If f is harmonic, then the right hand side equals zero for all x and hence $f'' \equiv 0$. From calculus, we know that this implies that f is a linear function.

We can rewrite the right hand side of (2.9) as

$$\lim_{\epsilon \to 0} \frac{1}{\epsilon} \left[\frac{f(x + \epsilon) - f(x)}{\epsilon} - \frac{f(x) - f(x - \epsilon)}{\epsilon} \right].$$

The fractions inside the square brackets are approximations of f' and from this we see get an expression that looks like the derivative of f'.

We can extend (2.9) to d dimensions. Define

$$\Delta f(x) = \lim_{\epsilon \to 0} \frac{1}{\epsilon^2} \sum_{y \in \mathbb{Z}^d, |y| = 1} [f(x + \epsilon y) - f(x)].$$

♢ A function $f : \mathbb{R}^d \to \mathbb{R}$ is C^k if all of its partial derivatives of order k exist and are continuous functions.

Proposition 2.6. *Suppose f is a C^2 function in a neighborhood of x in $\mathbb{R}^d$. Then $\Delta f(x)$ exists at x and*

$$\Delta f(x) = \sum_{j=1}^{d} \partial_{jj} f(x). \tag{2.10}$$

Proof. If y_j is the unit vector in $\mathbb{Z}^d$ (or $\mathbb{R}^d$) whose jth component equals 1, we can use the one-dimensional Taylor theorem in the direction of y_j to give

$$f(x \pm \epsilon\, y_j) = f(x) \pm \epsilon\, \partial_j f(x) + \frac{\epsilon^2}{2}\, \partial_{jj} f(x) + o(\epsilon^2).$$

Therefore,

$$\sum_{y \in \mathbb{Z}^d, |y|=1} [f(x+\epsilon y) - f(x)] = \epsilon^2 \sum_{j=1}^{d} \partial_{jj} f(x) + o(\epsilon^2).$$

□

We can rewrite the definition of Δ as

$$\frac{1}{2d}\, \Delta f(x) = \lim_{\epsilon \to 0} \frac{1}{\epsilon^2} \left[\left(\frac{1}{2d} \sum_{y \in \mathbb{Z}^d, |y|=1} f(x+\epsilon y) \right) - f(x) \right].$$

The term in the inner brackets can be considered a kind of mean value of f where the mean value is taken only in the coordinate directions. This mean value depends on our choice of coordinate axes. The next proposition shows that we can average over spheres as well and hence the definition does not depend on the choice of axes.

Proposition 2.7. *If f is C^2 in a neighborhood of x, then*

$$\frac{1}{2d}\, \Delta f(x) = \lim_{\epsilon \to 0} \frac{MV(f; x, \epsilon) - f(x)}{\epsilon^2}. \tag{2.11}$$

Here $MV(f; x, \epsilon)$ denotes the mean value of f on the sphere of radius ϵ about x as in (2.7).

Proof. Assume for notational ease that $x = 0$ and $f(x) = f(0) = 0$. If f is C^2 in a neighborhood of 0, we can write

$$f(y) = P_2(y) + o(|y|^2), \tag{2.12}$$

where P_2 denotes the second order Taylor polynomial of f about 0,

$$P_2(y) = y \cdot \nabla f(0) + \frac{1}{2} \sum_{1 \le k,l \le d} a_{kl}\, y_k\, y_l.$$

Here $y = (y_1, \ldots, y_d)$ and $a_{kl} = \partial_{kl} f(0)$. Note that $MV(y_k; 0, \epsilon) = 0$ by symmetry. Similarly, if $k \neq l$, $MV(y_k y_l; 0, \epsilon) = 0$. Therefore,

$$MV(P_2; 0, \epsilon) = \frac{1}{2} \sum_{k=1}^{d} a_{kk} MV(y_k^2; 0, \epsilon).$$

We could compute $\beta_k := MV(y_k^2; 0, \epsilon)$ by doing an integral, but we will use a trick to avoid this computation. By symmetry β_k is the same for all k and

$$\beta_1 + \cdots + \beta_d = MV(y_1^2 + \cdots + y_d^2; 0, \epsilon) = MV(\epsilon^2; 0, \epsilon) = \epsilon^2.$$

Therefore, $MV(y_k^2; 0, \epsilon) = \epsilon^2/d$. Finally, since the error term in (2.12) is $o(|y|^2)$,

$$\begin{aligned} \lim_{\epsilon \to 0+} \epsilon^{-2}\, MV(f; 0, \epsilon) &= \lim_{\epsilon \to 0+} \epsilon^{-2}\, MV(P_2; 0, \epsilon) \\ &= \frac{1}{2d} \sum_{k=1}^{d} a_{kk} = \frac{1}{2d} \Delta f(0). \end{aligned}$$

□

The operator Δ is called the *Laplacian*. In hindsight it might have been more convenient to call $\frac{1}{2}\Delta$ the Laplacian, but the terminology is fixed. Most books *define* the Laplacian by (2.10) (which is why the 1/2 does not appear), but it is more natural to think of the Laplacian as being defined by the mean value property. Note that $\Delta f(x) = \operatorname{div}[\nabla f(x)]$. Another standard notation for the Laplacian is ∇^2; one should think of this as

$$\nabla^2 = \nabla \cdot \nabla = \left(\frac{\partial}{\partial x_1}, \ldots, \frac{\partial}{\partial x_d} \right) \cdot \left(\frac{\partial}{\partial x_1}, \ldots, \frac{\partial}{\partial x_d} \right).$$

To add to the confusion, many analysts choose to define the Laplacian to be $-\Delta$. There are advantages in this in that this makes it a "positive operator", see (2.46). Whether one multiplies by 1/2 or puts in a minus sign, the condition $\Delta f(x) = 0$ means the same thing.

Theorem 2.8. *A function in a domain U is harmonic if and only if f is C^2 with $\Delta f(x) = 0$ for all $x \in U$.*

We will discuss the proof of this in the remainder of the section. This theorem will not be found in most books because they will define f to be harmonic in U if $\Delta f(x) = 0$. However, they will show that such functions satisfy the mean value property so in either case one needs to prove a theorem.

$\diamondsuit$ If $z = (z_1, \ldots, z_d) \in \mathbb{R}^d$, then we write $d^d z$ for $dz_1 \cdots dz_d$. Analysts generally favor n for the dimension of the space while probabilists tend to use d, reserving n as an index for sequences. One unfortunate consequence of using d for dimension is that one has to write $d^d z$ where the two ds have different meanings.

In fact, we have *almost* shown that f harmonic implies $\Delta f(x) = 0$ for all x. What we have shown is that if a function f satisfies the mean value property *and it is* C^2, then it satisfies $\Delta f = 0$. To finish the proof we will show that if f is continuous and satisfies the mean value property then it is automatically C^∞! To show this we need the fact (see Exercise 2.6) that there is a C^∞ function ϕ on $\mathbb{R}^d$ satisfying the following: ϕ is radially symmetric; $\phi(x) = 0$ for $|x| \geq 1$; $\phi(x) > 0$ for $|x| < 1$; and

$$\int_{\mathbb{R}^d} \phi(z)\, d^d z = 1.$$

Let ϕ be such a function and let $\phi_\epsilon(x) = \epsilon^{-d}\, \phi(x/\epsilon)$. Then ϕ_ϵ is positive if and only if $|x| < \epsilon$ and

$$\int_{\mathbb{R}^d} \phi_\epsilon(z)\, d^d z = 1.$$

Assume $\operatorname{dist}(x, \partial U) \geq 2\epsilon$. Then if f is continuous and satisfies the mean value property in U, we can use spherical coordinates and the radial symmetry of ϕ_ϵ to see that for $|y - x| < \epsilon$,

$$f(y) = \int_{\mathbb{R}^d} \phi_\epsilon(z)\, f(y+z)\, d^d z = \int_{\mathbb{R}^d} \phi_\epsilon(z-y)\, f(z)\, d^d z.$$

Derivatives of f with respect to y can now be taken by differentiating the right hand side. The continuity of f is used to justify the interchange of the integral and the derivative.

$\diamondsuit$ Often we want to interchange integrals and derivatives so let us discuss when this can be justified. Suppose $g(t,x), t \in \mathbb{R}, x \in \mathbb{R}^d$ is a continuous function on $\mathbb{R}^{d+1}$ for which the partial derivative $\partial_t g(t,x)$ exists and is a continuous function on $\mathbb{R}^{d+1}$. If V is a subset of $\mathbb{R}^d$, we would like to write

$$\partial_t \int_V g(t,x)\, d^d x = \int_V \partial_t g(t,x)\, d^d x. \tag{2.13}$$

Let

$$g_\epsilon(t,x) = \epsilon^{-1}\left[g(t+\epsilon,x) - g(t,x)\right],$$

so that

$$\partial_t g(t,x) = \lim_{\epsilon \to 0} g_\epsilon(t,x).$$

Using the definition of the derivative, we see that

$$\begin{aligned} \partial_t \int_V g(t,x)\, d^d x &= \lim_{\epsilon \to 0} \frac{1}{\epsilon}\left[\int_V g(t+\epsilon,x)\, d^d x - \int_V g(t,x)\, d^d x\right] \\ &= \lim_{\epsilon \to 0} \int_V g_\epsilon(t,x)\, d^d x, \end{aligned}$$

with the derivative existing if and only if the limit on the right exists. In order to show

$$\lim_{\epsilon \to 0} \int_V g_\epsilon(t,x)\, d^d x = \int_V \partial_t g(t,x)\, d^d x,$$

it suffices to show that

$$\lim_{\epsilon \to 0} \int_V \left[g_\epsilon(t,x) - \partial_t g(t,x)\right] d^d x = 0,$$

which in turn will follow if we can show that

$$\lim_{\epsilon \to 0} \int_V \left|g_\epsilon(t,x) - \partial_t g(t,x)\right|\, d^d x = 0. \tag{2.14}$$

Therefore (2.14) gives a sufficient condition to justify (2.13). If g has two continuous derivatives in t, the Taylor theorem with remainder tells us that

$$\left|g_\epsilon(t,x) - \partial_t g(t,x)\right| \le \frac{\epsilon}{2} K_\epsilon(t,x),$$

where

$$K_\epsilon(t,x) = \max\left\{\left|\partial_{tt} g(s,x)\right| : |s-t| \le \epsilon\right\}.$$

Therefore, a sufficient condition to justify (2.13) is the existence of an $\epsilon > 0$ such that

$$\int_V K_\epsilon(t,x)\, d^d x < \infty. \tag{2.15}$$

In all of the cases where we need to justify an interchange of an integral and a derivative, we can prove (2.15). However, we do not always include the details.

To finish the proof, we need to show that if f is C^2 in a domain U with $\Delta f(x) = 0$ at every x, and $x \in U$ with $\text{dist}(x, \partial U) > \epsilon$, then

$$MV(f; x, \epsilon) = f(x).$$

In other words, f satisfies the mean value property. (We know by (2.11) that f satisfies the mean value property "in the limit as ϵ tends to zero", but we want to show the actual mean value property.) For ease, we will assume that f has compact support; to derive the general case, use Exercise 2.7. Without loss of generality, we may assume $x = 0, f(0) = 0$.

♢ A function $f : \mathbb{R}^d \to \mathbb{R}$ has compact support if there is a $K < \infty$ such that $f(x) = 0$ for $|x| > K$.

The easiest way to establish the computation is by a direct calculation. Let us write

$$MV(\epsilon) = MV(f; 0, \epsilon) = \int_{|x|=1} f(\epsilon x)\, ds(x),$$

where now s represents normalized surface measure on the sphere of radius 1. We know that $MV(0) = 0$. To show that $MV(\epsilon) = 0$ for all ϵ, it suffices to show that $MV'(\epsilon) = 0$. Differentiation gives

$$MV'(\epsilon) = \int_{|x|=1} \frac{d}{d\epsilon} f(\epsilon x)\, ds(x) = \int_{|x|=1} \partial_r f(\epsilon x)\, ds(x),$$

where ∂_r denotes differentiation in the radial direction. Using Green's theorem (Exercise 2.11), one can show that if f is harmonic, then

$$\int_{|x|=1} \partial_r f(\epsilon x)\, ds(x) = 0. \tag{2.16}$$

♢ There are various versions of the fundamental theorem of calculus in d-dimensions that go under the name of Stokes' or Green's theorem. Here we need the following version. If $U \subset \mathbb{R}^d$ is a bounded, connected open set with smooth boundary and F is a smooth vector field, then

$$\int_{\partial U} (F \cdot \mathbf{n})(y)\, ds(y) = \int_U (\text{div} F)(x)\, d^d x.$$

Here s denotes (unnormalized) surface measure on ∂U and $\mathbf{n}$ denotes the outward unit normal. If f is a function, then its (outward) normal derivative on ∂U is given by $\nabla f \cdot \mathbf{n}$.

We now discuss a probabilistic way of seeing this relation which has the advantage of generalizing to mean values on sets other than spheres. Let W_t be a Brownian motion starting at the origin and let

$$T = T_\epsilon = \inf\{t \geq 0 : |W_t| = \epsilon\}$$

be the first time that W_t hits the sphere of radius ϵ about x. (Note that since W_t is a continuous function, we can replace the word infimum with the word minimum.) Then the radial symmetry of Brownian motion implies that the distribution of W_T is uniformly distributed on the sphere of radius ϵ, and hence

$$MV(f; 0, \epsilon) = \mathbb{E}\left[f(W_T)\right].$$

We need to show

$$\mathbb{E}\left[f(W_T)\right] = 0 \tag{2.17}$$

This will follow from the following martingale property: for every $t < \infty$,

$$\mathbb{E}\left[f(W_{T\wedge t})\right] = 0. \tag{2.18}$$

The argument to go from (2.18) to (2.17) is essentially the same as in the discrete Dirichlet problem so we skip the details. We will concentrate on deriving (2.18). We will start by deriving the easier relation $\mathbb{E}[f(W_t)] = 0$ and for ease we assume $f(W_0) = 0$ and choose $t = 1$.

Let $P_2(y; x)$ denote the second order Taylor polynomial of f about x and let

$$e(x, y) = f(y) - P_2(y; x).$$

Since f is C^2 with compact support, there is a function $\epsilon(\delta)$ with $\epsilon(0+) = 0$ such that for all x, y,

$$|e(x, y)| \leq \epsilon(|x - y|)\,|x - y|^2.$$

(See Exercise 2.5.)

For each n we write

$$f(W_1) = \sum_{m=1}^{n} \left[f(W_{m/n}) - f(W_{(m-1)/n}) \right].$$

We can write

$$P_2(y;x) = f(x) + (y-x)\cdot\nabla f(x) + \sum_{1\le j\le k\le d} a_{jk}(x)\,(y_j - x_j)\,(y_k - x_k),$$

where $a_{jk}(x) = \partial_{jk} f(x) = \partial_{kj} f(x)$. Since $\Delta f \equiv 0$, and f is C^2, we can see that

$$\sum_{j=1}^{d} a_{jj}(x)\,(y_j - x_j)\,(y_k - x_k) = o(|y|^2),$$

Hence, up to an error that is $o(|y|^2)$, $f(W_1)$ is the sum of the following three terms:

$$\sum_{m=1}^{n} \nabla f(W_{(m-1)/n}) \cdot \left[W_{m/n} - W_{(m-1)/n} \right],$$

$$\sum_{m=1}^{n} \sum_{1\le j<k\le d} a_{jk}(W_{(m-1)/n})\,[W^j_{m/n} - W^j_{(m-1)/n}\,]\,[W^k_{m/n} - W^k_{(m-1)/n}],$$

$$\sum_{m=1}^{n} e(W_{(m-1)/n}, W_{m/n}).$$

Here we write the d-dimensional Brownian motion $W_t = (W_t^1, \ldots, W_t^d)$. Since $\nabla f(W_{(m-1)/n})$ depends on the Brownian motion only up to time $(m-1)/n$ and $W_{m/n} - W_{(m-1)/n}$ is independent of this with mean zero, we can see that

$$\mathbb{E}\left(\nabla f(W_{(m-1)/n}) \cdot \left[W_{m/n} - W_{(m-1)/n} \right] \right) = 0.$$

Similarly, if $j < k$, $a_{jk}(W_{(m-1)/n})$, $W^j_{m/n} - W^j_{(m-1)/n}$, and $W^k_{m/n} - W^k_{(m-1)/n}$ are independent with the last two having mean zero. Therefore,

$$\mathbb{E}\left(a_{jk}(W_{(m-1)/n})\,[W^j_{m/n} - W^j_{(m-1)/n}\,]\,[W^k_{m/n} - W^k_{(m-1)/n}] \right) = 0,$$

For the final term, we note that

$$\left| \sum_{m=1}^{n} e(W_{(m-1)/n}, W_{m/n}) \right| \le \epsilon(R_n)\, Q_n$$

where

$$Q_n = \sum_{j=1}^{d} \sum_{m=1}^{n} [W^j_{m/n} - W^j_{(m-1)/n}\ ]^2,$$

and $R_n = \max\{|W_{m/n} - W_{(m-1)/n}| : m = 1, \ldots, n\}$. We claim that with probability one the right hand side converges to zero. Continuity of the Brownian motion implies that $R_n \to 0$. Also $\epsilon(R_n)$ is bounded (why?) and hence $\mathbb{E}[\epsilon(R_n)^2) \to 0$. In the exercises (see Exercise 2.23) we study Q_n; in particular, $\mathbb{E}[Q_n^2]$ is bounded in n. Therefore, using the Cauchy-Schwarz inequality,

$$\mathbb{E}[\epsilon(R_n)\, Q_n]^2 \le \mathbb{E}[\epsilon(R_n)^2]\, \mathbb{E}[Q_n^2] \to 0.$$

In particular, $\mathbb{E}[f(W_1)] = 0$.

This argument might tempt the reader to write

$$\begin{aligned} f(W_1) - f(W_0) &= \lim_{n\to\infty} \sum_{k=1}^{n} \nabla f(W_{T_{k-1}}) \cdot [W_{T_k} - W_{T_{k-1}}] \\ &= \int_0^1 \nabla f(W_t) \cdot dW_t. \end{aligned}$$

In fact, this can be made precise. This is an example of an Itô stochastic integral.

2.3. Dirichlet problem

We will consider the problem of finding harmonic functions with prescribed boundary values. We will restrict our discussion to bounded domains and continuous boundary values.

Dirichlet problem for harmonic functions. Given a bounded domain $U \subset \mathbb{R}^d$ and a continuous function $F : \partial U \to \mathbb{R}$, find an extension of F to all of $\overline{U}$ such that:

$$F : \overline{U} \to \mathbb{R} \text{ is continuous;} \tag{2.19}$$

$$\Delta F(x) = 0, \quad x \in U. \tag{2.20}$$

♢ If F represents temperature, then this gives the equilibrium temperature distribution on the interior if the temperature is fixed and known at the boundary.

We first show that the solution to the Dirichlet problem, if it exists, is unique. We do this using the *maximum principle* which states the following: *if U is a bounded domain and F satisfies* (2.19) *and* (2.20)*, then the maximum value of F is obtained somewhere on the boundary.* This follows from continuity and the mean value property as we now demonstrate. Since F is a continuous function on a compact set $\overline{U}$, F obtains its maximum somewhere. Suppose that the maximum were obtained at an interior point x. Since the average value about every sphere surrounding x in U is $F(x)$, continuity tells us that the function must take on the constant value $F(x)$ on each of these spheres. By letting the spheres grow, we can find a point on the boundary whose value is $F(x)$. Given the maximum principle (and the corresponding minimum principle), we can see that if F_1, F_2 are two solutions to (2.19) and (2.20) such that $F_1 \equiv F_2$ on ∂U, then $F = F_1 - F_2$ is a solution with $F \equiv 0$ on ∂U and hence $F \equiv 0$ on U.

♢ Here we use the linearity property of harmonic functions: if f, g are harmonic and a, b are constants, then $af+bg$ is harmonic. Many of the classical equations of mathematical physics such as (2.20) and the heat equation which we discuss below are linear partial differential equations. Most research in differential equations today involves *nonlinear* equations.

♢ The above argument proves the stronger fact that if U is connected and F obtains its maximum at an interior point, then it is constant.

To show existence, we make a good guess based on the discrete analogue. Suppose W_t is a d-dimensional Brownian motion starting at $x \in \overline{U}$ and let

$$T_U = \inf\{t \geq 0 : W_t \notin U\} = \min\{t \geq 0 : W_t \in \partial U\}.$$

Suppose $F : \partial U \to \mathbb{R}$ is given. For $x \in U$, we define

$$F(x) = \mathbb{E}\left[F(W_{T_U}) \mid W_0 = x\right]. \tag{2.21}$$

In other words, we start a Brownian motion at x and let it run until it hits the boundary and then observe the temperature. The temperature at x is the average value of this temperature, averaged over all Brownian paths. The rotational invariance of the Brownian motion shows that F as defined in (2.21) satisfies the mean value property. It is not difficult to show that F is continuous in U (continuity in $\overline{U}$ is trickier — we discuss this below) and hence $\Delta F(x) = 0$ for $x \in U$.

♢ Actually, a subtle fact about Brownian motion called the *strong Markov property* is being used here. It is sufficiently subtle that we will ignore this issue and let the reader read an advanced book on Brownian motion to find out what this means.

Example 2.9. Let $d = 1$ and $U = (0, R)$. Then $\partial U = \{0, R\}$. Let

$$F(x) = \mathbb{P}\{W_{T_U} = R \mid W_0 = x\}$$

be the probability that the Brownian motion starting at x reaches R before reaching 0. Then F satisfies

$$F(0) = 0, \quad F(R) = 1, \quad F''(x) = 0, 0 < x < U.$$

The unique solution to this is $F(x) = x/R$. More generally, the harmonic function on U with boundary values $F(0), F(R)$ is

$$F(x) = F(0) + \frac{x}{R}\left[F(R) - F(0)\right].$$

Example 2.10. Let $d \geq 2$, $0 < r < R < \infty$ and let U be the annular region

$$U = \{x \in \mathbb{R}^d : r < |x| < R\}, \quad \partial U = \{|x| = r\} \cup \{|x| = R\}.$$

Let $F(x) = \mathbb{P}\{|W_{T_U}| = R \mid W_0 = x\}$. By rotational symmetry we can see that $F(x) = \phi(|x|)$ for a one-variable function ϕ satisfying $\phi(r) = 0, \phi(R) = 1$. Also,

$$\Delta F(x) = \sum_{j=1}^{d} \partial_{jj} F(x) = 0, \quad x \in U.$$

If we write, $F(x_1, \ldots, x_d) = \phi(|x|) = \phi(\sqrt{x_1^2 + \cdots + x_d^2})$, then a chain rule computation (Exercise 2.16) gives

$$\Delta F(x) = \sum_{j=1}^{d} \partial_{jj} \phi(|x|) = \phi''(|x|) + \frac{(d-1)}{|x|}\, \phi'(|x|). \tag{2.22}$$

Therefore, we need to find the solutions to the one-variable equation

$$s\, \phi''(s) + (d-1)\, \phi'(s) = 0.$$

This is a first-order linear differential equation in ϕ' and standard methods give $\phi'(s) = c\, s^{1-d}$, which can be integrated again to yield

$$\phi(s) = c_1\, \log s + c_2, \quad d = 2,$$

$$\phi(s) = c_1\, s^{2-d} + c_2, \quad d \geq 3.$$

Plugging in the boundary conditions $\phi(r) = 0, \phi(R) = 1$ gives

$$\phi(|x|) = \frac{\log |x| - \log r}{\log R - \log r}, \quad d = 2,$$

$$\phi(|x|) = \frac{r^{2-d} - |x|^{2-d}}{r^{2-d} - R^{2-d}}, \quad d \geq 3.$$

The last example allows us to conclude some interesting facts about the d-dimensional Brownian motion. Let us first consider $d \geq 3$. If $r < |x|$ and we start a Brownian motion at $x \in \mathbb{R}^d$, then the probability that the Brownian motion ever reaches the ball of radius r about the origin is given by

$$\lim_{R \to \infty} \mathbb{P}\{|W_{T_U}| = r \mid W_0 = x\}$$

$$= \lim_{R \to \infty} \frac{|x|^{2-d} - R^{2-d}}{r^{2-d} - R^{2-d}} = \left(\frac{r}{|x|}\right)^{d-2} < 1.$$

In particular, there is a positive probability that the Brownian motion never returns to the ball. From this one can see (Exercise 2.8) that with probability one

$$\lim_{t \to \infty} |W_t| = \infty, \quad d \geq 3. \tag{2.23}$$

We say that Brownian motion for $d \geq 3$ is *transient*.

Now let us consider $d = 2$. We first ask the same question: if we start at x with $|x| > r$, what is the probability that we ever reach the sphere of radius r about the origin? This is given by

$$\lim_{R\to\infty} \mathbb{P}\{|W_{T_U}| = r \mid W_0 = x\} = \lim_{R\to\infty} \frac{\log R - \log |x|}{\log R - \log r} = 1.$$

In other words, for every fixed positive r, the Brownian motion keeps returning to the ball of radius r about the origin. Consider a second question: what is the probability that the Brownian motion starting at $x \neq 0$ ever reaches the origin? Assume for a moment that the probability were positive. Then there would be an R such that the probability of reaching the origin before getting distance R from the origin is positive. But,

$$\begin{aligned} \mathbb{P}\{\text{reach 0 before distance } R\} &\leq \lim_{r\to 0} \mathbb{P}\{|W_{T_U}| = r \mid W_0 = x\} \\ &= \lim_{r\to 0} \frac{\log R - \log |x|}{\log R - \log r} = 0. \end{aligned}$$

Therefore with probability one, the Brownian motion never reaches the origin. We say that Brownian motion for $d = 2$ is not *point recurrent* but is *neighborhood recurrent*.

There is nothing special about the point zero in the argument. The same argument shows that for all $x \in \mathbb{R}^2$, with probability one the Brownian motion never visits x after time zero. (On the other hand, it is obviously not true that with probability one, for every $x \in \mathbb{R}^2$, x is never visited. The order of quantifiers is important!)

♢ The next example concerns harmonic functions in $\mathbb{R}^2$. We identify $\mathbb{R}^2$ with $\mathbb{C}$, the set of complex numbers. This is not just for notational convenience. The theory of complex functions is very important in the study of real-valued harmonic functions in $\mathbb{R}^2$.

Example 2.11. Let $U = \{x \in \mathbb{R}^2 : |x| < 1\} = \{re^{i\theta} \in \mathbb{C} : 0 \leq r < 1, 0 \leq \theta < 2\pi\}$ be the two-dimensional unit disk whose boundary is the unit circle $\partial U = \{e^{i\theta} : 0 \leq \theta < 2\pi\}$. A continuous function $F : \partial U \to \mathbb{R}$ can be considered as a continuous function on $\mathbb{R}$ satisfying

$F(\theta) = F(\theta + 2\pi)$ for all θ. The harmonic function with boundary value F is

$$F(x) = \mathbb{E}\left[F(W_{\tau_U}) \mid W_0 = x\right].$$

For each $x \in U$, there is a probability distribution on ∂U that corresponds to the distribution of the first visit to ∂U by a Brownian motion starting at x. This distribution turns out to have a density with respect to length, $H(x, e^{i\theta})$, i.e., if $\theta_1 < \theta_2 < \theta_1 + 2\pi$, the probability that the Brownian motion starting at x hits ∂U with angle between θ_1 and θ_2 is given by

$$\int_{\theta_1}^{\theta_2} H(x, e^{i\theta})\, d\theta.$$

The function $H(x, e^{i\theta})$ is known explicitly and is called the *Poisson kernel*,

$$H(x, z) = \frac{1 - |x|^2}{2\pi\, |z - x|^2}, \quad |x| < 1, |z| = 1.$$

Therefore,

$$F(x) = \int_0^{2\pi} F(e^{i\theta})\, \frac{1 - |x|^2}{2\pi\, |e^{i\theta} - x|^2}\, d\theta. \tag{2.24}$$

We have pulled the kernel $H(x, z)$ out of a hat but given the formula one can check directly that F defined as in (2.24) is harmonic in U and F is continuous on $\overline{U}$ (see Exercise 2.9).

Example 2.12. There is a similar Poisson kernel if $d \geq 3$, $U = \{x \in \mathbb{R}^d : |x| < 1\}$. Let s denote the surface measure of the sphere $\{|x| = 1\}$. Then with respect to this measure, the Poisson kernel is

$$H(x, z) = \frac{1 - |x|^2}{C_d\, |z - x|^d}, \quad |x| < 1, |z| = 1, \tag{2.25}$$

where C_d denotes the $(d-1)$-dimensional surface measure of ∂U. In other words, the solution to the Dirichlet problem with given function F is

$$F(x) = \int_{|z|=1} F(z)\, H(x, z)\, ds(z).$$

It is a calculus exercise (Exercise 2.18) to verify that F defined as above is harmonic in U and is continuous in ∂U. If $V \subset \partial U$, then the

probability that a Brownian motion starting at x exits U at a point in V is given by

$$\int_V H(x,z)\,ds(z).$$

We end this section by asking the question: can the Dirichlet problem be solved for any bounded domain U? Suppose U is a bounded domain, and F is a continuous function on the boundary. Ifwe define F in U by

$$F(x) = \mathbb{E}\left[F(W_{T_U}) \mid W_0 = x\right], \tag{2.26}$$

then this is a continuous function in U satisfying the mean value property and hence is harmonic. This is the only candidate for the solution, but it is not clear if F is continuous in $\overline{U}$. In fact, this is not always the case. For example suppose that U is the "punctured unit disk"

$$U = \{x \in \mathbb{R}^2 : 0 < |x| < 1\}, \quad \partial U = \{0\} \cup \{x \in \mathbb{R}^2 : |x| = 1\}.$$

Suppose that we set $F(0) = 0$ and $F(x) = 1$ for $|x| = 1$. Then this is a continuous function on ∂U. We have seen that with probability one, a Brownian motion starting at $x \neq 0$, never hits the origin. Therefore, if we define F in U by (2.26), we get $F(x) = 1$ for all $x \neq 0$, and F is not continuous at 0. Of course, this example is bad because the boundary is not connected. One might ask then, what if we force ∂U to be connected? In this case, in two dimensions the Dirichlet problem always has a solution, but in more than two dimensions we can still have problems (see Exercise 2.14).

We will not prove this, but we just mention that in order for the Dirichlet problem to have a solution the domain U has to have a certain property which can be stated roughly as "if $y \in \partial U$ and the Brownian motion starts in U near y, then with very high probability the Brownian motion exits U near y". Such a domain is said to have a *regular boundary*.

2.4. Heat equation

We now consider the continuous analogue of the heat equation. We start by considering the equation in all of $\mathbb{R}^d$. Suppose that an initial

temperature is given by $f(x), x \in \mathbb{R}^d$, which we assume is a bounded, continuous function. Similarly to the discrete case, we imagine the temperature as being determined by a very large number of heat particles, all doing Brownian motions. Let $u(t,x), t \geq 0, x \in \mathbb{R}^d$ denote the temperature at x at time t which can be thought of (roughly) as the density of heat particles at x at time t. On a very heuristic level, we imagine that there are $f(y)$ points starting at site y and the fraction of them that are at x at time t is the probability that a particle at y has moved to x. We would expect that this probability would be the same as the probability that a particle moves from x to y in time t. If we average over all possible y, we get

$$u(t,x) = \mathbb{E}\left[f(W_t) \mid W_0 = x\right]. \tag{2.27}$$

Let us make this a little more precise. If W_t is a Brownian motion in $\mathbb{R}^d$ starting at x, then for fixed t, W_t is a random variable with (probability) density (function)

$$p(t,x,y) = \frac{1}{(2\pi t)^{d/2}}\, e^{-\frac{|y-x|^2}{2t}}.$$

Here t, x are fixed and $p(t,x,y)$ is considered as a function of y. In other words, the components of $W_t - x$ are independent normal random variables with mean zero and variance t. Symmetry is seen by noting that $p(t,x,y) = p(t,y,x)$. We can write (2.27) as

$$u(t,x) = \int_{\mathbb{R}^d} f(y)\, p(t,x,y)\, d^d y = \int_{\mathbb{R}^d} f(y)\, p(t,y,x)\, d^d y. \tag{2.28}$$

The first equality is a restatement of (2.27) and the second equality more closely reflects our interpretation of the heat flow — the density of heat particles that started at y and are at x at time t is $f(y)\, p(t,y,x)$.

We define $u(0,x) = f(x)$ and $u(t,x)$ for $t > 0$ by (2.27). It is not difficult to show (Exercise 2.13) that

$$u(0,x) = u(0+,x) := \lim_{t \to 0+} u(t,x). \tag{2.29}$$

$\diamondsuit$ Mathematicians sometimes say that $\{p(t,0,y) : t > 0\}$ is an approximate δ-function. Heuristically, a delta function in $\mathbb{R}^d$ is a function δ satisfying

the following:

$$\delta(0) = \infty, \quad \delta(x) = 0, \;\; x > 0, \quad \int_{\mathbb{R}^d} \delta(y)\, d^d y = 1.$$

In particular, if f is a bounded continuous function

$$\int_{R^d} f(y)\, \delta(y - x)\, d^d y = f(x). \tag{2.30}$$

As stated, this does not make mathematical sense, but there are a number of ways to make this precise. One way is to think of the delta function as $p(0+, 0, y)$ and then (2.30) becomes (2.29).

We will now find the partial differential equation that $u(t, x)$ satisfies. First assume $d = 1$, and consider the right derivative with respect to time. (To see that computing the right derivative suffices, see Exercise 2.15.) For ease, assume that $t = 0, x = 0, f(0) = 0$. Then,

$$\lim_{s \to 0+} \frac{u(s, 0) - f(0)}{s} = \lim_{s \to 0+} \frac{\mathbb{E}[f(W_s) \mid W_0 = 0]}{s}.$$

Assume that f is C^2 and write the approximation by the second order Taylor polynomial,

$$f(x) = f'(0)\, x + \frac{1}{2}\, f''(0)\, x^2 + o(x^2), \quad x \to 0.$$

Then

$$\mathbb{E}[f(W_s)] = f'(0)\, \mathbb{E}[W_s] + \frac{1}{2}\, f''(0) \mathbb{E}[W_s^2] + o(W_s^2).$$

But $\mathbb{E}[W_s] = 0, \mathbb{E}[W_s^2] = s$ and $o(W_s^2) = o(s)$. Hence, if we divide by s and let $s \to 0$ we expect the limit to be $f''(0)/2$. A similar argument for $t > 0$ gives the prediction

$$\partial_t\, u(t, x) = \frac{1}{2}\, \partial_{xx} u(t, x). \tag{2.31}$$

Some might not be happy with the level of rigor in this argument, but that is not a problem because once we guess the equation we can verify it directly by checking that

$$u(t, x) = \int_{-\infty}^{\infty} f(y)\, \frac{1}{\sqrt{2\pi t}}\, e^{-\frac{(y-x)^2}{2t}}\, dy \tag{2.32}$$

satisfies (2.31) at least for $t > 0$. This is a straightforward computation provided that one justifies an interchange of a derivative and an

integral (see the remark in Section 2.2). This will also hold at $t = 0$ for the right derivative if f is C^2, but if f is only continuous we must be content with (2.29).

The computation for d dimensions is similar so we just state it. Suppose f is a bounded continuous function in $\mathbb{R}^d$. Then $u(t, x)$ as defined in (2.28) satisfies the *heat equation*

$$\partial_t u(t,x) = \frac{1}{2}\Delta_x u(t,x), \quad t > 0, \tag{2.33}$$

with initial condition

$$u(0,x) = u(0+,x) = f(x). \tag{2.34}$$

Here we write Δ_x to indicate that the Laplacian is in the x variable only. One can verify by direct differentiation that u satisfies (2.33). In fact, although we will not prove it here, this is the unique solution to (2.33) with initial condition (2.34).

2.5. Bounded domain

The solution of the heat equation in all of $\mathbb{R}^d$ is easy; in fact, we could have just written down the solution (2.32) and verified that it works. However, if we restrict to a bounded domain, it can be harder to give a solution and for this it is useful to have the probabilistic interpretation.

Suppose $U \subset \mathbb{R}^d$ is a bounded domain. Assume an initial temperature $f(x), x \in U$ is given, and let us fix the temperature to be 0 at the boundary at all times. We will assume that the boundary is not too bad — at least that it is regular as described in Section 2.3. If we let $u(t, x) = u(t, x; U)$ denote the temperature at point x at time t, then $u(t, x)$ satisfies the following:

$$u(0,x) = u(0+,x) = f(x), \quad x \in U,$$

$$u(t,x) = 0, \quad x \in \partial U, \tag{2.35}$$

$$x \mapsto u(t,x) \text{ is continuous on } \overline{U} \text{ for } t > 0,$$

$$\partial_t u(t,x) = \frac{1}{2}\,\Delta_x u(t,x), \quad t > 0, x \in U. \tag{2.36}$$

The derivation of the heat equation (2.36) is similar to that for the case $U = \mathbb{R}^d$. Essentially, if $x \in U$ and t is very small, the effect of the boundary on heat flow about x is minimal (and the effect goes to zero as t goes to 0). Hence we get the same differential equation as in the unbounded case. We note that the set of functions satisfying (2.35) and (2.36) is a vector space.

We still have the interpretation of heat as being given by heat particles doing Brownian motions, but these particles are destroyed when they reach the boundary. Let W_t be one such Brownian motion, let $T = T_U = \inf\{t : W_t \notin U\}$, and let $p(t, x, y, U)$ denote the density at y at time t assuming that $W_0 = x$ and that the particle has not been killed by time t. To be more precise, if $V \subset U$, then

$$\mathbb{P}\{W_t \in V, T > t \mid W_0 = x\} = \int_V p(t, x, y, U)\, d^d y.$$

The expression (2.27) becomes

$$\begin{aligned} u(t, x; U) &= \mathbb{E}\left[f(W_t)\, 1\{T > t\} \mid W_0 = x\right] \\ &= \int_U f(y)\, p(t, x, y, U)\, d^d y. \end{aligned}$$

The derivation of this equation uses $p(t, x, y, U) = p(t, y, x, U)$. This equality may not be as obvious as in the case of the entire plane, but for any "path" from x to y staying in U, there is a corresponding path from y to x staying in U obtained by reversal.

2.5.1. One dimension. Suppose $d = 1, U = (0, \pi)$. We can solve the heat equation exactly in terms of an infinite series of functions. We start by looking for solutions of (2.35) and (2.36) of the form

$$u(t, x) = e^{-\lambda t}\, \phi(x).$$

Such a function satisfies (2.36) if and only if

$$\phi''(x) = -2\lambda\, \phi(x),$$

which has general solution $\phi(x) = c_1 \sin(\sqrt{2\lambda}\, x) + c_2 \cos(\sqrt{2\lambda}\, x)$. Imposing the boundary condition $u(t, 0) = u(t, \pi) = 0$ gives us the following solutions

$$\phi_k(x) = \sin(kx), \qquad \lambda_k = \frac{k^2}{2}, \quad k = 1, 2, \ldots$$

Since linear combinations of solutions are solutions, we get a family of solutions of the form

$$u(t,x) = \sum_{k=1}^{\infty} a_k \, e^{-k^2 t/2} \, \sin(kx). \tag{2.37}$$

At this point, we need to take some care. If all but a finite number of the a_k are zero, this is a finite sum and this gives a solution. Otherwise we have to worry about convergence of the sum and whether the infinite sum satisfies the heat equation. Let us ignore that problem at the moment and do some formal calculations. If we plug in $t = 0$ we get

$$u(0,x) = \sum_{k=1}^{\infty} a_k \, \sin(kx).$$

If we want this to equal f, then we need to find coefficients a_k such that this holds. This is an example of a *Fourier series*. Continuing the formal calculations, let us suppose that

$$f(x) = \sum_{k=1}^{\infty} a_k \, \sin(kx) \tag{2.38}$$

and compute the coefficients a_k. A simple calculation gives

$$\int_0^{\pi} \sin(jx) \, \sin(kx) \, dx = \begin{cases} 0, & j \neq k \\ \pi/2, & j = k \end{cases}.$$

If we use the sum rule for integrals (and do not worry about the fact that it is an infinite sum!), we see that

$$\int_0^{\pi} f(x) \sin(kx) \, dx = \sum_{j=1}^{\infty} \int_0^{\pi} a_j \, \sin(jx) \, \sin(kx) \, dx = \frac{\pi}{2} \, a_k,$$

which gives

$$a_k = \frac{2}{\pi} \int_0^{\pi} f(x) \, \sin(kx) \, dx. \tag{2.39}$$

We now return to the convergence issue. Suppose we start with a continuous function f and define a_k as above. In what sense is (2.38) true? In other words, if

$$f_n(x) = \sum_{k=1}^{n} a_k \, \sin(kx), \tag{2.40}$$

do the functions f_n converge to f? Unfortunately, it is not true that $f_n(x) \to f(x)$ for every continuous f and every x. However, if we change our definition of convergence, we do always have convergence.

♢ There are many different nonequivalent definitions of convergence of functions. This is not just because mathematicians like to have many definitions and like to prove theorems about them! Here, the notion of pointwise convergence of functions would be very natural but convergence does not hold for all functions. Choosing another notion of convergence allows all functions to converge.

For every continuous f, f_n as defined in (2.40) converges to f *in mean-square* or *in* L^2; this means

$$\lim_{n\to\infty} \int_0^\pi [f(x) - f_n(x)]^2 \, dx = 0.$$

We will not give the full proof of this theorem, but in the proof one derives *Parseval's identity* (which is another form of the Pythagorean theorem!) by justifying this calculation:

$$\begin{aligned} \int_0^\pi f(x)^2 \, dx &= \int_0^\pi \left[\sum_{k=1}^\infty a_k \, \sin(kx) \right]^2 dx \\ &= \sum_{j=1}^\infty \sum_{k=1}^\infty \int_0^\pi a_j \, a_k \, \sin(jx) \, \sin(kx) \, dx \\ &= \sum_{k=1}^\infty \int_0^\pi a_k^2 \, \sin^2(kx) \, dx = \frac{\pi}{2} \sum_{k=1}^\infty a_k^2. \end{aligned}$$

In particular, $\sum a_k^2 < \infty$. This does not imply that $\sum |a_k| < \infty$. However, it does imply that $a_k \to 0$, from which we can see that the sum in (2.37) converges absolutely for each x if $t > 0$.

We can use intuition from Brownian motion to derive the heat equation. Conversely, we can use solutions to the heat equation to study the Brownian motion. Consider Brownian motion in $U = (0, \pi)$ killed when it reaches a boundary point. We can compute the density of a Brownian particle starting at y assuming that it has not died.

This corresponds to the formal initial condition

$$f(x) = \delta(y - x).$$

The solution of the heat equation with this initial condition should be the density $p(t, y, x, U)$. Using the formal property of the delta function, we get

$$\int_0^\pi \delta(y - x)\, \sin(kx)\, dx = \sin(ky).$$

Plugging into (2.39) gives $a_k = (2/\pi)\, \sin(ky)$, and hence we get

$$p(t, y, x, U) = \frac{2}{\pi} \sum_{k=1}^{\infty} e^{-k^2 t/2}\, \sin(kx)\, \sin(ky).$$

Note that

$$\left| \sum_{k=2}^{\infty} e^{-k^2 t/2}\, \sin(kx)\, \sin(ky) \right| \le \sum_{k=2}^{\infty} e^{-kt} = \frac{e^{-2t}}{1 - e^{-t}}.$$

Hence as $t \to \infty$, the sum on the right hand side is dominated by the $k = 1$ term,

$$p(t, y, x, U) \sim \frac{2}{\pi}\, e^{-t/2}\, \sin(x)\, \sin(y), \quad t \to \infty.$$

Note that

$$\int_0^\pi \frac{2}{\pi}\, e^{-t/2}\, \sin(x)\, \sin(y)\, dx = \frac{4}{\pi}\, e^{-t/2}\, \sin y.$$

Let us interpret this. If we start a Brownian motion at y, then the probability at a very large time t, that the particle has not left $(0, \pi)$ is about $(4/\pi)\, e^{-t/2}\, \sin y$. Given that it has not left the domain, the probability density for where we expect the particle to be is $(1/2)\, \sin x$. Note that this last density does not depend on y — at a very large time, the position of the particle given that it stays in the domain is independent of the starting point. In other words, the particle forgets its starting point. In the random walk case, we had a similar result except there is something that the random walker does not forget — whether its initial point is even or odd.

2.5.2. Many dimensions. The same idea, separation of variables, can be used to solve the heat equation in bounded domains U in $\mathbb{R}^d$. Again, one looks for solutions in a product form $u_j(t,x) = e^{-\lambda_j t}\,\phi_j(x)$. This will give a solution satisfying the boundary condition if

$$\frac{1}{2}\,\Delta\phi_j(x) = -\lambda_j\phi_j(x), \qquad \phi_j \equiv 0 \text{ on } \partial U. \tag{2.41}$$

The function ϕ_j is called an *eigenfunction* for Δ with Dirichlet boundary conditions with *eigenvalue* $-2\lambda_j$. This leads to the following problem for each domain: can we find a *complete* family of eigenfunctions satisfying (2.41)? In other words, can we find sufficiently many such functions so that every initial condition f can be written as

$$f(x) = \sum_{k=1}^{\infty} a_k\,\phi_k(x), \tag{2.42}$$

for appropriate constants a_k?

Since heat is lost at the boundary, we do not expect to have any solutions that grow with time — hence, we expect $\lambda_j > 0$. Since we have put in a minus sign, λ_j is an eigenvalue for $-\frac{1}{2}\Delta$. It turns out that under very general conditions one can find such functions and the eigenvalues can be ordered

$$\lambda_1 < \lambda_2 \le \lambda_3 \le \cdots,$$

and the eigenfunctions ϕ_j can be chosen to be orthonormal

$$\langle \phi_j, \phi_k\rangle := \int_U \phi_j(x)\,\phi_k(x)\,d^dx = 0, \qquad j \neq k,$$

$$\langle \phi_j, \phi_j\rangle = \int_U \phi_j(x)^2\,d^dx = 1.$$

(In the one-dimensional case, the orthonormal eigenfunctions were $\phi_j(x) = \sqrt{2/\pi}\,\sin(jx)$.) Every continuous function f on $\overline{U}$ can be written as a generalized Fourier series (2.42) with

$$\sum_{k=1}^{\infty} a_k^2 < \infty,$$

where the convergence in the sum is in the mean-square or L^2 sense. The coefficients are given by

$$a_k = \langle f, \phi_k \rangle = \int_U f(x)\, \phi_k(x)\, d^d x.$$

The solution to the heat equation with initial condition f is

$$u(t,x) = \sum_{k=1}^{\infty} a_k\, e^{-\lambda_k t}\, \phi_k(x).$$

If we want a solution whose initial condition is the "delta function" at y, then we choose

$$a_k = \int_U \delta(y-x)\, \phi_k(x)\, d^d x = \phi_k(y),$$

and get

$$p(t,y,x;U) = p(t,x,y;U) = \sum_{k=1}^{\infty} e^{-\lambda_k t}\, \phi_k(x)\, \phi_k(y).$$

The first equality uses the fact that for each path from x to y staying in U there is a corresponding path from y to x staying in U obtained by traversing backwards. If we start a Brownian motion at y, the probability that it is still in U is asymptotic to $c\, e^{-\lambda_1 t}\, \phi_1(y)$ and the probability density for the particle conditioned that it stays in U is about $c^{-1}\, \phi_1(x)$. Here $c = \int_U \phi_1(x)\, d^d x$.

Example 2.13. Many special functions arising in physics involve the eigenfunctions of the Laplacian on a domain. Let $U = \{x \in \mathbb{R}^2 : |x| < 1\}$. We will look for solutions of the equation

$$\frac{1}{2}\Delta\phi(x) = -\lambda\, \phi(x), \qquad \phi \equiv 0 \text{ on } \partial U.$$

We use separation of variables to look for solutions of the form

$$\phi(r,\theta) = h(r)\, g(\theta).$$

Then (see Exercise 2.10),

$$\Delta\phi(r,\theta) = \left[h''(r) + r^{-1}\, h'(r)\right] g(\theta) + r^{-2}\, h(r)\, g''(\theta).$$

If we want this to equal $-2\lambda\, h(r)\, g(\theta)$, then

$$\frac{r^2\, h''(r) + r\, h'(r) + 2\, r^2\, \lambda h(r)}{h(r)} = -\frac{g''(\theta)}{g(\theta)}. \tag{2.43}$$

Note that the left hand side is a function of r and the right hand side is a function of θ. In order for these to be equal, they must be equal to a constant, say β. The function g is periodic with period 2π so we can see that the only possible choices are $\beta_j = j^2$ and

$$\phi_j(\theta) = \sin(j\theta), \quad \psi_j(\theta) = \cos(j\theta).$$

(For $j = 0$, only the cosine function is nonzero.). Then h satisfies

$$h''(r) + \frac{1}{r}\, h'(r) + \left[2\lambda - \frac{j^2}{r^2}\right] h(r) = 0.$$

We need to solve this equation with the boundary value $h(1) = 0$. If we write $v(s) = h(s/\sqrt{2\lambda})$, this equation becomes

$$v''(s) + \frac{1}{s}\, v'(s) + \left(1 - \frac{j^2}{s^2}\right) v(s) = 0.$$

There is only one solution of this equation (up to multiplicative constant) that stays bounded as $s \to 0+$. It is called the jth order *Bessel function*.

We consider a special case where we assume that the initial function f is radially symmetric. We look for functions of the form $\phi(r, \theta) = h(r)$. This requires the constant $\beta = j^2$ in (2.43) to be 0, and hence $v(s) = h(s/\sqrt{2\lambda})$ satisfies the zero order Bessel equation

$$v''(s) + \frac{1}{s}\, v'(s) + v(s) = 0. \tag{2.44}$$

This is a second-order differential equation that has two linearly independent solutions. The solutions cannot be given in closed form. There exists only one solution (up to multiplicative constant) that is bounded and continuous at 0; it is unique if we specify $v(0) = 1$ and can be given by (Exercise 2.20)

$$v(x) = J_0(x) := \frac{2}{\pi} \int_0^{\pi/2} \cos(x \cos\theta)\, d\theta. \tag{2.45}$$

This is the zeroth order Bessel function of the first kind.

By analyzing (2.45) (Exercise 2.21), one can show that the roots of J_0 form an increasing sequence

$$0 < r_1 < r_2 < r_3 < \cdots$$

We therefore have the functions $h_k(x) = J_0(r_k x)$ satisfy

$$h_k''(x) + \frac{1}{x}\, h_k'(x) + r_k^2\, h_k(x) = 0, \quad h_k(1) = 0.$$

In fact, the functions are orthogonal,

$$\int_0^1 h_j(x)\, h_k(x)\, dx = 0, \quad j \neq k.$$

If we set $\phi_k(x) = h_k(|x|)$, we have

$$\Delta\phi_k(x) = -\, r_k^2\, \phi_k(x), \quad \phi_k \equiv 0 \text{ on } \partial U.$$

This gives a complete set of radially symmetric eigenfunctions.

♢ The Bessel function $J_0(x)$ plays a similar role to that played by $J(x) = \sin x$ for the usual Fourier series. The zeroes of J are the positive integers and hence $h_k(x) = J(kx) = \sin(kx)$.

♢ The eigenfunctions and eigenvalues for the Laplacian on a domain are sometimes called the harmonics of the domain. The last example shows that finding harmonics for domains leads to studying differential equations. The Bessel functions are just some of the many "special functions" of mathematics and physics that have arisen in studying the Laplacian and other operators on domains. Much is known about such functions — see any book on special functions to learn more.

To solve the heat equation on the domain U we need to find the eigenfunctions and eigenvalues. Of particular importance is $-\lambda_1$, the eigenvalue of smallest absolute value. In fact, λ_1 is always positive (and hence all the eigenvalues are negative with other eigenvalues having greater absolute value). As $t \to \infty$,

$$p(t, x, y; U) \sim e^{-\lambda_1 t} \phi_1(x)\, \phi_1(y).$$

Since the left hand side is positive, it had better be the case that the eigenfunction ϕ_1 can be chosen to be strictly positive (or strictly negative). Let us look at this from a different perspective; for ease assume that U has a smooth boundary. Then using Green's identities

(d-dimensional analogues of integration by parts), one can see for any function f satisfying $f \equiv 0$ on ∂U,

$$\langle -\Delta f, f\rangle = -\int_U f(x)\,\Delta f(x)\,dx = \int_U |\nabla f(x)|^2\,dx > 0. \tag{2.46}$$

(For this reason, $-\Delta$ is sometimes called a *positive operator*). We then get

$$2\,\lambda_1 = \min \frac{\langle -\Delta f, f\rangle}{\langle f, f\rangle},$$

where the minimum is over all smooth functions f that equal zero on ∂U. This is the continuous analogue of Theorem 1.8. To show this, one notes that by plugging in ϕ_1, the eigenfunction associated to λ_1, we can see that the minimum on the right hand side is no more than λ_1. But for general f, we can write

$$f(x) = \sum_{j=1}^{\infty} a_k\,\phi_k$$

and we can see that we actually get equality.

♢ The Green's identity used is

$$\int_U f\Delta g(x)\,d^d x = \int_{\partial U} f(\nabla g\cdot \mathbf{n})(y)\,ds(y) - \int_U (\nabla f\cdot\nabla g)(x)\,d^d x.$$

2.6. More on harmonic functions

If $d \geq 2$ and U is an open subset of $\mathbb{R}^d$, then the set of harmonic functions F on U is a an infinite dimensional vector space. The set of such functions F that can be extended to $\overline{U}$ in a continuous way is also an infinite dimensional subspace. Harmonic functions have some nice properties. For example, the next proposition shows that the derivatives can be bounded in terms of the maximum of the function.

Proposition 2.14. *Suppose U is an open subset of $\mathbb{R}^d$, and f is a harmonic function on U. For $x \in U$, let*

$$\rho(x) = \operatorname{dist}(x, \partial U) = \inf\{|x-y| : y \in \partial U\}.$$

Then

$$|\nabla f(x)| \le \frac{d}{\rho} \sup_{x \in U} |f(x)|.$$

Proof. Let $M = \sup_{x \in U} |f(x)|$ and assume $M < \infty$ (the result is trivial if $M = \infty$). Let us first consider the case where U is the open unit ball, $x = 0$, and f extends to a continuous function on $\overline{U}$. Then $M = \max_{x \in \overline{U}} |f(x)|$. We know that

$$f(x) = \int_{\partial U} f(z)\, H(x,z)\, ds(z).$$

A calculation (Exercise 2.18) shows that $|\nabla H(0,z)| = d/C_d$. (Here ∇ refers to the gradient in the first component.) Therefore,

$$\begin{aligned} |\nabla f(0)| &= \left| \nabla \int_{\partial U} f(y)\, H(x,z)\, ds(z) \,\big|_{x=o} \right| \\ &= \left| \int_{\partial U} f(z)\, \nabla H(0,z)\, ds(z) \right| \\ &\le \int_{\partial U} |f(z)|\, |\nabla H(0,z)|\, ds(z) \\ &\le \int_{\partial U} M\, \frac{d}{C_d}\, ds(z) = Md. \end{aligned}$$

More generally, let $r < \rho(x)$ and let

$$g(y) = f(x + ry).$$

Then g is a continuous function on $\overline{U}$ that is harmonic in U. Also $\max_{y \in \overline{U}} |g(y)| \le M$. Therefore,

$$|\nabla g(0)| \le d\, M.$$

But $\nabla f(x) = r \nabla g(0)$. Therefore,

$$r\, |\nabla f(x)| \le M.$$

Since this holds for all $r < \rho(x)$, we have proved the proposition. □

We use this proposition to establish a continuous analogue of a theorem we proved for discrete harmonic functions.

Proposition 2.15. *The only bounded harmonic functions on $\mathbb{R}^d$ are the constant functions.*

Proof. Suppose $f : \mathbb{R}^d \to \mathbb{R}$ satisfies $\sup_x |f(x)| = M < \infty$. Then applying the previous proposition to the domain $U = \{x : |x| < 2R\}$ we see that for every $R < \infty$ and every $|x| < R$,

$$|\nabla f(x)| \leq \frac{Md}{R}.$$

Therefore $\nabla f(x) = 0$ for all x and hence f is constant. □

Because the Poisson kernel for the unit ball is given explicitly we can do many computations. For most domains, it is impossible to give an explicit form for the kernel. For some domains, however, separation of variables can be used effectively.

Example 2.16. Let U denote the rectangle

$$U = \{(x, y) \in \mathbb{R}^2 : 0 < x < 1, 0 < y < \pi\}.$$

The boundary of U consists of four line segments. We will consider harmonic functions F whose boundary values are zero on three of those segments but may be nonzero on $\partial^* = \{(1, y) : 0 \leq y \leq \pi\}$. An easy calculation shows that if

$$\phi_j(x, y) = \sinh(jx)\, \sin(jy),$$

then ϕ_j is harmonic in U (in fact, ϕ_j is harmonic in $\mathbb{R}^2$). Moreover, if j is a positive integer, $\phi_j \equiv 0$ on the three boundary segments other than ∂^*. Suppose F is defined on ∂^* by $F(1, y) = g(y)$ where $g : [0, \pi] \to \mathbb{R}$. Then we want a function of the form

$$F(x, y) = \sum_{j=1}^{\infty} a_j\, \sinh(jx)\, \sin(jy),$$

where the constants a_j have been chosen so that

$$g(y) = \sum_{j=1}^{\infty} a_j\, \sinh(j)\, \sin(jy).$$

This is the Fourier series for g and we have seen that we should choose

$$a_j\, \sinh(j) = \frac{2}{\pi} \int_0^{\pi} g(z)\, \sin(jz)\, dz.$$

To find the Poisson kernel, we choose the boundary value equal to the "delta function" at y', i.e.,

$$a_j \sinh(j) = \frac{2}{\pi} \sin(jy').$$

Hence,

$$H((x,y),(\pi,y')) = \frac{2}{\pi} \sum_{j=1}^{\infty} \frac{\sinh(jx)\, \sin(jy)\, \sin(j\, y')}{\sinh(j)}.$$

This is a continuous analogue of (1.25).

2.7. Constructing Brownian motion

In this section we discuss the construction of Brownian motion. It is basically a two step process. Brownian motion is first constructed on the countable dense subset $\mathcal{D}$ and then it is proved that the process is uniformly continuous and hence can be extended to all times.

2.7.1. Existence of Brownian motion on $\mathcal{D}$. We start by establishing a fact about sums of independent normal random variables. We write $X \sim N(\mu, \sigma^2)$ if X has a normal distribution with mean μ and variance σ^2. It is well known that if X, Y are independent with $X \sim N(\mu_X, \sigma_X^2)$ and $Y \sim N(\mu_Y, \sigma_Y^2)$, then $X + Y \sim N(\mu_X + \mu_Y, \sigma_X^2 + \sigma_Y^2)$. Let X, Y be independent random variables each $N(0, 1/2)$, so that if $Z = X + Y$, then $Z \sim N(0, 1)$. Suppose the value of Z is known, say $Z = z$. What can we say about X and Y?

The joint density for (X, Y) is

$$\left(\frac{1}{\sqrt{2\pi(1/2)}} e^{-x^2} \right) \left(\frac{1}{\sqrt{2\pi(1/2)}} e^{-y^2} \right) = \frac{1}{\pi} e^{-(x^2+y^2)}.$$

The joint density for (X, Z) is

$$\frac{1}{\pi} e^{-(x^2+(z-x)^2)},$$

and the density for Z is

$$\frac{1}{\sqrt{2\pi}}\, e^{-z^2/2}.$$

The conditional density of X given $Z = z$ is

$$\frac{(1/\pi)e^{-(x^2+(z-x)^2)}}{(1/\sqrt{2\pi})\,e^{-z^2/2}} = \frac{1}{\sqrt{\pi/2}}\,e^{-2(x-\frac{z}{2})^2}.$$

This is the density for a normal random variable with mean $z/2$ and variance $1/4$. In other words, conditioned on the value of Z, $X \sim N(Z/2, 1/4)$. We can write

$$X = \frac{Z}{2} + \frac{\tilde{Z}}{2},$$

where $\tilde{Z} \sim N(0,1)$ and is independent of Z. Similarly conditioned on $Z = z$, $Y \sim N(Z/2, 1/4)$ and

$$Y = \frac{Z}{2} - \frac{\tilde{Z}}{2}.$$

(Note that conditioned on $Z = z$, the random variables X and Y are not conditionally independent!) We have essentially proved the following proposition.

Proposition 2.17. *Suppose X, Y are independent normal random variables, each $N(0,1)$. If*

$$Z = \frac{1}{\sqrt{2}}\,X + \frac{1}{\sqrt{2}}\,Y,$$

$$\tilde{Z} = \frac{1}{\sqrt{2}}\,X - \frac{1}{\sqrt{2}}\,Y,$$

then $Z, \tilde{Z}$ are independent random variables, each $N(0,1)$.

We will construct W_q for $q \in \mathcal{D}$. It suffices to define the random variables $J(k,n)$ as in (2.1). We assume that we have at our disposal a countable number of independent normal random variables $Z_1, Z_2, \ldots$. Since the dyadics $\mathcal{D}$ are a countable set, we may assume that the random variables Z_q are actually indexed by $q \in \mathcal{D}$. Our definition of $J(k,n)$ will be recursive. We start by defining

$$J(k,0) = Z_k, \quad k = 1, 2, \ldots.$$

We now assume that

$$J(k,n), \quad k = 1, 2, \ldots$$

have been defined using only $\{Z_q : q \in \mathcal{D}_n\}$ so that they are independent $N(0,1)$ random variables. We then define

$$J(2k-1, n+1) = \frac{1}{\sqrt{2}} J(k,n) + \frac{1}{\sqrt{2}} Z_{(2k+1)/2^{n+1}},$$

$$J(2k, n+1) = \frac{1}{\sqrt{2}} J(k,n) - \frac{1}{\sqrt{2}} Z_{(2k+1)/2^{n+1}}.$$

By repeated use of the proposition we see that

$$J(k, n+1), \quad k = 1, 2, \ldots,$$

are independent $N(0,1)$ random variables. We define $W_{k/2^n}$ by

$$W_{k/2^n} = 2^{-n/2} \sum_{j=1}^{k} J(j,n),$$

so that (2.1) holds.

2.7.2. Continuity of Brownian motion. Let $W_q, q \in \mathcal{D}$ be a standard one dimensional Brownian motion and let K_n be as defined in (2.3). In this section we prove the following.

Theorem 2.18. *If n is a positive integer and $a > 0$,*

$$\mathbb{P}\left\{K_n \geq a\, 2^{-n/2}\right\} \leq \frac{4 \cdot 2^n}{a}\, e^{-a^2/2}. \tag{2.47}$$

In particular, by setting $a = 2\sqrt{n}$, we see that

$$\mathbb{P}\left\{K_n \geq 2\sqrt{n}\, 2^{-n/2}\right\} \leq \frac{2}{\sqrt{n}}\, (2/e^2)^n,$$

which gives (2.4).

Note that $2^{n/2} K_n$ is the maximum of 2^n random variables all with the same distribution as

$$\hat{K} := \sup\{|W_q| : q \in \mathcal{D} \cap [0,1]\}.$$

The probability that the maximum of a collection of random variables is greater than a number r is no more than the sum of the probabilities that the individual random variables are greater than r. Hence to prove (2.47), it suffices to show that

$$\mathbb{P}\left\{\hat{K} \geq a\right\} \leq \frac{4}{a}\, e^{-a^2/2}.$$

Proposition 2.19. *Suppose $W_q, q \in \mathcal{D}$ is a standard Brownian motion. Then for every $a > 0$,*

$$\mathbb{P}\left\{\hat{K} > a\right\} \le 4\mathbb{P}\{W_1 \ge a\} = \frac{4}{\sqrt{2\pi}} \int_a^\infty e^{-x^2/2}\, dx \le \frac{4}{a}\, e^{-a^2/2}.$$

Proof. The equality comes from the fact that W_1 has a normal distribution with mean zero and variance one. The last inequality follows from

$$\int_a^\infty e^{-x^2/2}\, dx \le \int_a^\infty e^{-ax/2}\, dx = \frac{2}{a}\, e^{-a^2/2},$$

and $2 < \sqrt{2\pi}$. Therefore, we only need to prove the first inequality. By symmetry it suffices to show that

$$\mathbb{P}\left\{\sup\{W_q : q \in \mathcal{D} \cap [0,1]\} > a\right\} \le 2\,\mathbb{P}\{W_1 \ge a\}.$$

Also, if $\sup\{W_q : q \in \mathcal{D} \cap [0,1]\} > a$, then $W_q > a$ for some $q \in \mathcal{D} \cap [0,1]$. Therefore,

$$\mathbb{P}\left\{\sup\{W_q : q \in \mathcal{D} \cap [0,1]\} > a\right\} \le$$

$$\lim_{n\to\infty} \mathbb{P}\left\{\max\{W_q : q \in \mathcal{D}_n \cap [0,1]\} \ge a\right\},$$

and it suffices to show for each n,

$$\mathbb{P}\left\{\max\{W_{k/2^n} : k = 1, \ldots, 2^n\} \ge a\right\} \le 2\mathbb{P}\{W_1 \ge a\}. \tag{2.48}$$

This looks complicated because we are taking the maximum of many random variables. However, we can use the fact that if we are greater than a at some time t then there is at least a 50% chance that we are above a at the final time. We need to take a little care in the argument; in order to avoid ambiguity, we consider the *first* time that the value is at least a. Fix n and let $E_k = E_{k,n}$ denote the event that $k/2^n$ is the first such time, i.e.,

$$W_{k/2^n} \ge a, \quad W_{j/2^n} < a, \;\; j = 1, \ldots, k-1.$$

The events $E_1, E_2, \ldots, E_{2^n}$ are mutually exclusive (i.e., $E_j \cap E_k = \emptyset$ for $j \ne k$) and their union is the event on the left hand side of (2.48). The event E_k depends on $W_{j/2^n}$ for $j = 1, \ldots, k$. In particular,

the random variable $W_1 - W_{k/2^n}$ is independent of the event E_k. Therefore, for each $k = 1, \ldots, 2^n$,

$$\begin{aligned} \mathbb{P}[E_k \cap \{W_1 \geq a\}] &\geq \mathbb{P}[E_k \cap \{W_1 - W_{k/2^n} \geq 0\}] \\ &= \mathbb{P}(E_k)\, \mathbb{P}\{W_1 - W_{k/2^n} \geq 0\} \\ &\geq \frac{1}{2}\, \mathbb{P}(E_k). \end{aligned}$$

(The last inequality is an equality if $k < 2^n$.) Therefore,

$$\begin{aligned} \mathbb{P}\{W_1 \geq a\} &= \sum_{k=1}^{2^n} \mathbb{P}[E_k \cap \{W_1 \geq a\}] \\ &\geq \frac{1}{2} \sum_{k=1}^{2^n} \mathbb{P}(E_k) \\ &= \frac{1}{2}\, \mathbb{P}\left\{\max\{W_{k/2^n} : k = 1, \ldots, 2^n\} \geq a\right\}. \end{aligned}$$

This proves (2.48). □

It is often useful to have a similar result for d-dimensional Brownian motion. Suppose W_t is a standard d-dimensional Brownian motion and K_n, K_n^* are defined as before. The triangle inequality again gives $K_n \leq K_n^* \leq 3\, K_n$. If $K_n \geq a$, then the corresponding quantity for at least one of the components must be at least $a/\sqrt{d}$. Hence (2.47) implies the following.

Theorem 2.20. *For a d-dimensional Brownian motion, if n is a positive integer and $a > 0$,*

$$\begin{aligned} \mathbb{P}\left\{2^{n/2}\, K_n^* \geq 3a\right\} &\leq \mathbb{P}\left\{2^{n/2}\, K_n \geq a\right\} \\ &\leq d\, \frac{4 \cdot 2^n}{(a/\sqrt{d})}\, e^{-\frac{(a/\sqrt{d})^2}{2}} \\ &= \frac{4 \cdot 2^n\, d^{3/2}}{a}\, e^{-\frac{a^2}{2d}}. \end{aligned} \tag{2.49}$$

2.8. Exercises

Exercise 2.1.

- Suppose $W_t, t \in \mathcal{D}$ is a standard Brownian motion on the dyadics. If n is an integer show that $\hat{W}_t = 2^{-n/2} W_{t2^n}$ is a standard Brownian motion on the dyadics.
- Suppose $W_t, t \in [0, \infty)$ is a standard Brownian motion, $a \neq 0$ and $\hat{W}_t = a\, W_{t/a^2}$. Show that $\hat{W}_t$ is a standard Brownian motion.

(In both cases you need to show that $\hat{W}_t$ satisfies the conditions to be a Brownian motion.)

Exercise 2.2. Suppose W_t is a standard Brownian motion.

- Show that with probability one for every $N < \infty$ there is a $t > N$ with $W_t = 0$.
- Show that with probability one for every $\epsilon > 0$ there is a $t \in (0, \epsilon)$ with $W_t = 0$.

Exercise 2.3. Suppose $U_r = \{x \in \mathbb{R}^d : |x| < r\}$. Suppose that F is a continuous function on $\overline{U_r}$ that is harmonic in U_r. Let $\phi(x) = F(rx)$. Show that ϕ is a harmonic function on U_1.

Exercise 2.4. Suppose $X = (X^1, \ldots, X^d)$ is a d-dimensional random variable with density $\phi(x_1, \ldots, x_d)$. Suppose that:

- ϕ is a radially symmetric function; in other words.
$$\phi(x_1, \ldots, x_d) = f(x_1^2 + \cdots + x_d^2)$$
for some $f : [0, \infty) \to (0, \infty)$.
- $X^1, \ldots, X^d$ are independent random variables.

Show that this implies that ϕ is of the form
$$\phi(x_1, \ldots, x_d) = \frac{1}{(2\pi\sigma^2)^{d/2}} \exp\left\{\frac{x_1^2 + \cdots + x_d^2}{2\sigma^2}\right\}$$
for some $\sigma^2 > 0$.

Exercise 2.5. Suppose $f : \mathbb{R}^d \to \mathbb{R}$ is a C^2 function and that there is a K such that $f(x) = 0$ for $|x| \geq K$. Let $P_2(x; y)$ denote the 2nd order Taylor polynomial about y. Show that for every $\epsilon > 0$ there is a $\delta > 0$ such that if $|x - y| < \delta$,
$$|f(x) - P_2(x; y)| \leq \epsilon\, |x - y|^2.$$

(Note: δ may depend on f but it cannot depend on x, y. It will be useful to recall that continuous functions on compact sets are uniformly continuous.)

Exercise 2.6.

- Let
$$f(x) = \begin{cases} e^{-1/x}, & x > 0 \\ 0, & x = 0. \end{cases}$$
Show that f has derivatives of all orders (this is trivial except at $x = 0$) and $f^{(k)}(0) = 0$ for all k.
- Use Taylor's Theorem with remainder to conclude that for all $x > 0$,
$$\sup_{0 \le t \le x} |f^{(k)}(t)| \ge \frac{k!\, e^{-1/x}}{x^k}.$$
- Define $\phi : \mathbb{R}^d \to [0, \infty)$ as follows:
$$\phi(x) = \begin{cases} 0, & |x| \ge 1 \\ e^{-1/(1-|x|^2)}, & |x| < 1 \end{cases}.$$
Show that ϕ is C^∞.

Exercise 2.7. Suppose U is an open set and V is a compact subset of U. Suppose f is a C^2 function on U. Show that there is a C^2 function g on $\mathbb{R}^d$ such that g has compact support (for some K, $g(x) = 0$ for $|x| \ge K$) and $g(x) = f(x)$ for $x \in V$.

Exercise 2.8. Verify (2.23).

Exercise 2.9. Suppose F is a continuous function on the unit circle $\{|x| = 1\}$ in $\mathbb{R}^2$ and let F be defined on the open disk $U = \{|x| < 1\}$ by (2.24).

- Show that F is harmonic in U.
- Show that F is continuous on ∂U.

Exercise 2.10. Suppose $\phi(r, \theta)$ is a function on $\mathbb{R}^2$ written in polar coordinates. Show that
$$\Delta\phi = \partial_{rr}\phi + \frac{1}{r}\,\partial_r\phi + \frac{1}{r^2}\,\partial_{\theta\theta}\phi.$$

Exercise 2.11. Use Green's theorem to establish (2.16).

Exercise 2.12. Prove the following *Harnack inequality*. Suppose $U = \{x \in \mathbb{R}^d : |x| < 1\}$. For every $r < 1$, there is a $C = C(r, d) < \infty$ such that if $F : U \to (0, \infty)$ is harmonic in U, then

$$C^{-1} F(0) \le F(x) \le C F(0), \quad |x| \le r.$$

(The constant C may depend on r, d but does not depend on x or F.)

Exercise 2.13.

- Verify (2.29) if f is a bounded, continuous function.
- The boundedness assumption on f is more than what is needed. Show that there is a $\beta > 0$ such that if f is continuous at x and

$$\lim_{|y| \to \infty} e^{-\beta |y|^2} f(y) = 0,$$

then (2.29) holds.

Exercise 2.14. Let W_t be a three-dimensional Brownian motion and let

$$U = \{x \in \mathbb{R}^3 : |x| < 1\} \setminus \{(s, 0, 0); 0 \le s < 1\}.$$

- Show that U is a connected domain and ∂U is connected.
- Show that if $x \in U$, then with probability one a Brownian motion starting at x exits U on $\{|y| = 1\}$.
- Find a continuous function on ∂U for which the Dirichlet problem does not have a solution.

Exercise 2.15. Suppose $f : [0, \infty) \to \mathbb{R}$ is a function such that the right-derivative defined by

$$f'_+(t) = \lim_{\delta \to 0+} \frac{f(t + \delta) - f(t)}{\delta},$$

exists at every point.

- Give an example with f continuous and f'_+ discontinuous.
- Give an example with f'_+ continuous and f discontinuous.
- Prove that if f, f'_+ are both continuous, then f is continuously differentiable and $f'(t) = f'_+(t)$ using the following hints.

– By considering

$$g(t) = f(t) - f(0) - \int_0^t f'_+(s)\, ds,$$

show that it suffices to prove this result with $f'_+ \equiv 0$, $f(0) = 0$.
– Assume $f'_+ \equiv 0$, let $\epsilon > 0$, and let

$$t_\epsilon = \inf\{t > 0 : |f(t)| > t\,\epsilon\}.$$

Show that for every $\epsilon > 0$, $t_\epsilon > 0$.
– Show that if $t_\epsilon < \infty$, then $|f(t_\epsilon)| = t_\epsilon\,\epsilon$.
– Show that if $|f(t)| = t\epsilon$, then there exists a $\delta > 0$ such that $|f(s)| \le s\epsilon$ for $t \le s \le t + \delta$.

Exercise 2.16. Justify (2.22).

Exercise 2.17. Suppose W_t is a two-dimensional Brownian motion. True or false: with probability one, W_t visits every open subset of $\mathbb{R}^2$.

Exercise 2.18. Suppose $d \ge 2$ and $U = \{z \in \mathbb{R}^d : |z| < 1\}$ is the open unit ball with boundary $\partial U = \{z \in \mathbb{R}^d : |z| = 1\}$. Let s denote surface measure so that the $(d-1)$-dimensional area of ∂U is

$$C_d = \int_U 1\, ds(y).$$

(For example, $C_2 = 2\pi$ and $C_3 = 4\pi$.) For $x \in U, z \in \partial U$ let

$$H(x,z) = \frac{1 - |x|^2}{C_d\,|x - z|^d}.$$

- Show that for fixed $z \in \partial U$, $H(x,z)$ is a harmonic function of x.
- Show that for fixed $z \in \partial U$,

$$C_d\,|\nabla H(0,z)| = d.$$

- Show that for fixed $x \in U$,

$$\int_{\partial U} H(x,z)\, ds(z) = 1.$$

- Show that if $F : \partial U \to \mathbb{R}$ is continuous, and we extend F to U by

$$F(x) = \int_{\partial U} F(z)\, H(x,z)\, ds(z),$$

then F is a harmonic function in U that is continuous on ∂U.

Exercise 2.19. Suppose d, k are positive integers. Show that there exists a $c = c(d,k) < \infty$ such that the following holds. Let $f : U \to \mathbb{R}$ be a harmonic function with $|f(x)| \le 1$ for $x \in U$. Then if D denotes any kth order partial derivative,

$$|Df(x)| \le c\, \rho(x)^{-k},$$

where $\rho(x) = \operatorname{dist}(x, \partial U) = \inf\{|x - y| : y \in \partial U\}$.

Exercise 2.20.

- Verify that J_0 as defined in (2.45) satisfies (2.44). Be sure to justify the interchange of derivative and integral in the calculation.
- Show that

$$J_0(x) = \sum_{n=0}^{\infty} \frac{(-1)^n}{(n!)^2\, 2^{2n}}\, x^{2n}.$$

(Hint: the right hand side is absolutely convergent for all x and hence derivatives can be taken by term-by-term differentiation.)

Exercise 2.21. Show that there are an infinite number of positive zeros of J_0 all of which are isolated. (This is actually a very difficult exercise. Feel free to "cheat" by finding a book that discusses asymptotics of Bessel functions.)

Exercise 2.22. Suppose $X_1, X_2, \ldots$ are independent random variables with mean zero. Suppose there exists $K < \infty$ such that $\mathbb{E}[X_j^4] \le K$ for each j. Let $S_n = X_1 + \cdots + X_n$.

- Show that $\mathbb{E}[X_j^2 X_k^2] \le K$ for all $1 \le j, k \le n$.
- Show that $\mathbb{E}[S_n^4] \le 3\, K\, n^2$.
- Show that $\mathbb{P}\{|S_n| \ge (3K)^{1/4}\, n^{7/8}\} \le n^{-3/2}$.

- Prove that with probability one,

$$\lim_{n\to\infty} \frac{S_n}{n} = 0.$$

This is an example of the *strong law of large numbers*.

Exercise 2.23. Let W_t be a standard Brownian motion. For each positive integer n, let

$$Q_n = \sum_{j=1}^{n} \left(W_{j/n} - W_{(j-1)/n}\right)^2.$$

Use the following outline to prove that with probability one,

$$\lim_{n\to\infty} Q_n = 1.$$

Let

$$Y_{n,j} = n\left(W_{j/n} - W_{(j-1)/n}\right)^2 - 1.$$

- Show that $\mathbb{E}[Y_{n,j}] = 0$.
- Show there exists a number K such that for all n, j,

$$\mathbb{E}\left[Y_{n,j}^4\right] = K.$$

- Complete the proof using the ideas of Exercise 2.22.

Exercise 2.24. Let W_t be a standard Brownian motion. For each positive integer n and each real $t \geq 0$, let

$$Q_{n,t} = \sum_{1\leq j\leq nt} \left(W_{j/n} - W_{(j-1)/n}\right)^2.$$

Show that with probability one, for each t,

$$\lim_{n\to\infty} Q_{n,t} = t. \tag{2.50}$$

(Note the order of the quantifiers. This is a stronger statement than saying for each t, (2.50) holds with probability one.)

Exercise 2.25. Suppose W_t, B_t are independent standard Brownian motions. Let

$$Q_n = \sum_{j=1}^{n} \left(W_{j/n} - W_{(j-1)/n}\right)\left(B_{j/n} - B_{(j-1)/n}\right).$$

Prove that with probability one,

$$\lim_{n\to\infty} Q_n = 0.$$

Exercise 2.26. In this exercise we will show that with probability one, there is no $t \in (0,1)$ at which W_t is differentiable. The first two steps are about functions.

- Suppose there exists a $t \in [0,1]$ at which W_t is differentiable. Then, there exists $t \in [0,1], \epsilon > 0, C < \infty$ such that
$$|W_s - W_{s'}| \le C\,\epsilon, \qquad \text{if } s, s' \in [t-\epsilon, t+\epsilon].$$
- Let
$$M(k,n) = \max\left\{\left|W_{\frac{k}{n}} - W_{\frac{k-1}{n}}\right|, \left|W_{\frac{k+1}{n}} - W_{\frac{k}{n}}\right|, \left|W_{\frac{k+2}{n}} - W_{\frac{k+1}{n}}\right|\right\},$$
$$M_n = \min\{M(1,n), \ldots, M(n,n)\}.$$
Suppose there exists a $t \in [0,1]$ at which W_t is differentiable. Then there is a $C < \infty$ and an $n_0 < \infty$ such that for all $n \ge n_0$, $M_n \le C/n$.

We now use the fact that W_t is a Brownian motion.

- Find a constant c such that for all C and all k, n,
$$\mathbb{P}\{M(k,n) \le C/n\} \le \left[\mathbb{P}\{|W_{1/n}| \le C/n\}\right]^3 \le \left[\frac{c\,C}{\sqrt{n}}\right]^3.$$
- Show that this implies that for all C,
$$\lim_{n\to\infty} \mathbb{P}\{M_n \ge C/n\} = 1.$$
- Conclude that with probability one, W_t is nowhere differentiable on $t \in [0,1]$.

Exercise 2.27. Let K_n be as in (2.49). Use this estimate to show that for every $\beta < 1/2$ and every $k < \infty$,
$$\lim_{n\to\infty} 2^{\beta n}\mathbb{E}\left[K_n^k\right] < \infty.$$

Chapter 3

Martingales

3.1. Examples

A martingale is a mathematical model of a "fair game". Before giving the general definition, we will consider a number of examples.

3.1.1. Simple random walk. Let $X_1, X_2, \ldots$ be independent random variables which equal ± 1 each with probability $1/2$. We can consider them as the winnings (or losses) from a simple game where one wins a dollar if a coin comes up heads and loses a dollar if it comes out tails. The total winnings after n plays is

$$S_n = X_1 + \cdots + X_n, \quad S_0 = 0.$$

Of course, this is exactly the same as the simple random walk in one dimension. Note that $\mathbb{E}[S_{n+1} - S_n] = \mathbb{E}[X_{n+1}] = 0$. In fact, a stronger fact is true. Since X_{n+1} is independent of $X_1, \ldots, X_n$, the *conditional* expected value stays the same even if one is given the values of $X_1, \ldots, X_n$,

$$E\left[S_{n+1} - S_n \mid X_1, \ldots, X_n\right] = 0.$$

3.1.2. Simple random walk with betting. Let $X_1, X_2, \ldots$ be as above, but suppose that at each time n we are allowed to place a bet, B_n on the outcome of the nth game. We allow B_n to be negative which is equivalent to betting that the coin will come up tails.

Then the winnings derived from the nth game is $B_n\,X_n$ and the total winnings by time n is

$$W_n = \sum_{j=1}^{n} B_j\,X_j, \quad W_0 = 0.$$

In order to be fair, we are required to make our choice of bet B_n before seeing the result of the nth flip. However, we will allow the bet to depend on the previous results $X_1, \ldots, X_{n-1}$. Then B_n will be some function $\phi(X_1, \ldots, X_{n-1})$ (B_1 will be a constant). Let us consider the conditional expectation

$$E\left[B_n\,X_n \mid X_1, \ldots, X_{n-1}\right].$$

This notation means: we first observe $X_1, \ldots, X_{n-1}$ and then take the "best guess" for $B_n\,X_n$ given this information. For each possible set of observations there is a best guess, and hence this conditional expectation should be a function $F(X_1, \ldots, X_{n-1})$. Given the values $X_1, \ldots, X_{n-1}$, we see that $B_n\,X_n$ equals $\pm F(X_1, \ldots, X_{n-1})$ each with probability $1/2$. In particular, the conditional expectation equals zero,

$$E\left[B_n\,X_n \mid X_1, \ldots, X_{n-1}\right] = 0.$$

This is the martingale property.

Let us write this property slightly differently. Let $\mathcal{F}_n$ denote the "information" available at time n. Then

$$\begin{aligned} E[W_n \mid \mathcal{F}_{n-1}] &= E[W_{n-1} + B_n\,X_n \mid \mathcal{F}_{n-1}] \\ &= E[W_{n-1} \mid \mathcal{F}_{n-1}] + E[B_n\,X_n \mid \mathcal{F}_{n-1}] = W_{n-1}. \end{aligned}$$

Here we have used some properties of conditional expectation that we will discuss in more detail below. The second inequality follows from linearity and the third equality follows from the previous paragraph and the fact that W_{n-1} is completely determined by $\mathcal{F}_{n-1}$.

One interesting choice for ϕ is called the *martingale betting strategy* and is a well known way of "beating a fair game". In order to guarantee winning, one keeps doubling one's bet until one is lucky enough to win. This corresponds to $B_1 = 1$ and for $n > 1$,

$$B_n = \begin{cases} 2^{n-1} & \text{if } X_1 = X_2 = \cdots = X_{n-1} = -1 \\ 0 & \text{otherwise .} \end{cases}$$

Note that for this strategy, the probability distribution of the winnings W_n is given by

$$\mathbb{P}\{W_n = 1 - 2^n\} = 2^{-n}, \quad \mathbb{P}\{W_n = 1\} = 1 - 2^{-n}. \tag{3.1}$$

In particular,

$$\mathbb{E}[W_n] = (1 - 2^{-n})\,\mathbb{P}\{W_n = 1 - 2^n\} + 1\,\mathbb{P}\{W_n = 1\} = 0.$$

Even though this strategy guarantees that one will eventually win, the expected winnings up to any finite time is zero. With probability one,

$$W_\infty := \lim_{n\to\infty} W_n = 1.$$

In particular,

$$\lim_{n\to\infty} \mathbb{E}[W_n] \neq \mathbb{E}\left[\lim_{n\to\infty} W_n\right].$$

3.1.3. A problem in statistics. Suppose $X_1, X_2, \ldots$ are independent random variables taking values in $\{0, 1\}$ with

$$\mathbb{P}\{X_1 = 1\} = 1 - \mathbb{P}\{X_1 = 0\} = \theta.$$

Suppose that we do not know the value of θ, but we observe X_1, $\ldots$, X_n. Can we determine θ? This is an example of a problem in statistics: to determine the distribution given some data. Note the following.

- If we only observe a finite number of data points $X_1, \ldots, X_n$, we cannot determine θ with 100% assurance. Indeed, for any $0 < \theta < 1$, the probability of seeing a particular sequence of points $X_1, \ldots, X_n$ is

$$\binom{n}{k} \theta^k\,(1-\theta)^{n-k} > 0, \tag{3.2}$$

  where $k = X_1 + \cdots + X_n$ denotes the number of 1's in the sequence $X_1, \ldots, X_n$.

- If somehow we could observe the infinite sequence X_1, X_2, $\ldots$, we would be able to determine θ. Indeed, the law of large numbers states that with probability one

$$\theta = \lim_{n\to\infty} \frac{X_1 + \cdots + X_n}{n}.$$

Since we cannot practically observe the infinite sequence, we want to estimate θ as well as possible from the given data.

There are a number of approaches that statisticians use all of which make some assumptions on the observed data. The *Bayesian* approach is to model the (unknown) success probability as a random variable θ with a given density and then to update our density using the data.

♢ In fact, we have already made the *assumptions* that the different data points are independent and that the success probability θ does not change with time. For real data, one would need to worry about testing these assumptions. However, we will simplify our discussion by assuming that this is given.

Since we have no prior knowledge of θ, we might start by assuming that θ is chosen uniformly over the range $[0,1]$, i.e., that at time 0 the density of θ is that of a uniform random variable,

$$f_0(x) = 1, \quad 0 < x < 1.$$

Given $X_1, \ldots, X_n$, the probability that we observe this data given a particular value of θ is given by (3.2). By a form of the *Bayes rule*, the conditional density at time n given that $X_1 + \cdots + X_n = k$ is given by

$$\begin{aligned} f_n(x \mid k) &= \frac{\binom{n}{k} x^k (1-x)^{n-k}}{\int_0^1 \binom{n}{k} y^k (1-y)^{n-k}\, dy} \\ &= (n+1)\binom{n}{k} x^k (1-x)^{n-k}, \quad 0 < x < 1. \end{aligned}$$

The conditional expectation of θ given $X_1 + \cdots + X_n = k$ is

$$\begin{aligned} \int_0^\infty x\, f_n(x \mid k)\, dx &= \int_0^x (n+1)\binom{n}{k} x^{k+1} (1-x)^{n-k}\, dx \\ &= \frac{(n+1)\binom{n}{k}}{(n+2)\binom{n+1}{k+1}} \\ &= \frac{k+1}{n+2}. \end{aligned} \tag{3.3}$$

Note that as $n \to \infty$ this looks like k/n; however, it is not exactly equal. This is because there is a lingering effect from the fact that we

assumed that the a priori distribution was the uniform distribution (note that if $k = n = 0$, then the expected value is $1/2$.)

$\diamondsuit$ In updating the density at time n, we are allowed to use all of the information in $X_1, \ldots, X_n$. For this model, the only thing that is relevant for the updating is the sum $X_1 + \cdots + X_n$. For this reason, the quantity $X_1 + \cdots + X_n$ is called a *sufficient* statistic for this model.

Another question we can ask is: given $X_1, \ldots, X_n = k$ what is the probability that $X_{n+1} = 1$? Again, given the value of θ, we know that the probability is θ and hence

$$\mathbb{P}\{X_{n+1} = 1 \mid X_1 + \cdots + X_n = k\} = \int_0^1 x\, f_n(x \mid k)\, dx = \frac{k+1}{n+2}.$$

This is the same transitions as one would get from a process called *Polya's urn*. Suppose an urn contains red and green balls. At time zero there is one ball of each type. At each integer time n a ball is chosen at random from the urn; the color is checked; and then the ball is returned to the urn along with another ball of the same color. If we let $Y_n + 1$ denote the number of red balls at time n, then the number of green balls is $(n - Y_n) + 1$ and

$$\mathbb{P}\left\{Y_{n+1} = k+1 \mid Y_n = k\right\} = \frac{k+1}{n+2},$$

$$\mathbb{E}[Y_{n+1} \mid Y_n = k] = k\left[1 - \frac{k+1}{n+2}\right] + (k+1)\frac{k+1}{n+2} = k + \frac{k+1}{n+2}.$$

Let $M_n = (Y_n + 1)/(n+2)$ denote the fraction of red balls at time n. Then

$$\mathbb{E}\left[M_{n+1} \mid Y_n = k\right] = E\left[\frac{Y_{n+1}+1}{n+3} \mid Y_n\right] = \tag{3.4}$$

$$\frac{k + \frac{k+1}{n+2} + 1}{n+3} = \frac{k+1}{n+2} = M_n.$$

Returning to the statistics model, let

$$N_n = E[\theta \mid X_1, \ldots, X_n] = E[\theta \mid \mathcal{F}_n],$$

where again we use $\mathcal{F}_n$ for the information in $X_1, \ldots, X_n$. Note that N_n is determined by the values $X_1, \ldots, X_n$; in fact, (3.3) states that

$$N_n = \frac{X_1 + \cdots + X_n + 1}{n+2}.$$

The computation in (3.4) implies that $\{N_n\}$ satisfies the martingale property

$$E[N_{n+1} \mid \mathcal{F}_n] = N_n.$$

3.1.4. A random Cantor set. The *Cantor set* $A \subset [0,1]$ is defined as

$$A = \bigcap_{n=0}^{\infty} A_n,$$

where $A_0 \supset A_1 \supset A_2 \supset \cdots$ are defined as follows: $A_0 = [0,1]$, $A_1 = [0,1/3] \cup [2/3,1]$,

$$A_2 = \left[0, \frac{1}{9}\right] \cup \left[\frac{2}{9}, \frac{1}{3}\right] \cup \left[\frac{2}{3}, \frac{7}{9}\right] \cup \left[\frac{8}{9}, 1\right],$$

and recursively A_{n+1} is obtained from A_n by removing the open "middle third" interval from each of the intervals in A_n. Note that A_n is the disjoint union of 2^n closed intervals each of length 3^{-n}.

We will construct a similar, but more complicated object, that we call a *random Cantor set*. There will be two parameters: k a positive integer greater than 1 and $p \in (0,1)$. Again, we choose $A_0 = [0,1]$ and we will let

$$A = \bigcap_{n=0}^{\infty} A_n,$$

for suitably chosen A_n. To define A_1, we start by dividing $[0,1]$ into k equal intervals

$$\left[0, \frac{1}{k}\right], \left[\frac{1}{k}, \frac{2}{k}\right], \ldots, \left[\frac{k-1}{k}, 1\right].$$

Independently for each of these intervals we decide to retain the interval with probability p and to discard the interval with probability $1-p$. This gives us a random set A_1 which is a union of intervals of length k^{-1}; the interiors of the intervals are disjoint. Let Y_1 denote the number of such intervals so that A_1 is the union of Y_1 intervals of length k^{-1}. Once A_1 is determined, we similarly split each of these Y_1 intervals into k pieces each of length k^{-2}. For each of

these smaller intervals, we retain the interval with probability p and discard the interval with probability $1-p$. All of these decisions are made independently. Then A_2 is the union of Y_2 intervals of length k^{-2}. Recursively, we define A_{n+1} from A_n by retaining each interval of length $k^{-(n+1)}$ in A_n independently with probability p.

Because we are choosing randomly, for each $n > 0$, there is a positive probability that $A_n = \emptyset$. However, by compactness we know that one of two things happens:

- There is a finite n for which $A_n = \emptyset$;
- $A \neq \emptyset$.

The process Y_n is sometimes called a *branching process* or *Galton-Watson process*. It is a simple stochastic model for population growth if we view Y_n as the number of individuals in the nth generation of a population. The (asexual) reproduction rule is that each individual in the nth generation has a random number of offspring j where the probability of j offspring is

$$p_j = \binom{k}{j} p^k \, (1-p)^{k-j}. \tag{3.5}$$

(This is the binomial distribution with parameters p and k. One can also define branching processes with other distributions for the offspring process, but this is the distribution that corresponds to the random Cantor set.) Let μ, σ^2 denote the mean and variance of the distribution (3.5); it is well known that

$$\mu = kp, \qquad \sigma^2 = kp(1-p).$$

The conditional distribution of Y_{n+1} given $Y_n = m$ is not easy to write down explicitly. However, the construction shows that is should be the distribution of the sum of m independent random variables each with distribution (3.5). In particular,

$$\mathbb{E}(Y_{n+1} \mid Y_n = m) = m\,\mu, \quad \operatorname{Var}(Y_{n+1} \mid Y_n = m) = m\,\sigma^2. \tag{3.6}$$

$\diamondsuit$ We are using the fact that in order to determine the distribution of Y_{n+1} given all the information up to time n, the only relevant information is Y_n, the number of individuals in the nth generation. Note, however, if we

are interested in the set A_{n+1}, we need to know the set A_n which cannot be determined only from the information in Y_n, or even from $Y_0, Y_1, \ldots, Y_n$.

Let $M_n = \mu^{-n} Y_n$. Then (3.6) implies the martingale property

$$E[M_{n+1} \mid \mathcal{F}_n] = E\left[\mu^{-n-1} Y_{n+1} \mid Y_n\right] = \mu^{-n-1} \mu Y_n = M_n.$$

3.2. Conditional expectation

Although we have already computed the conditional expectation informally in our examples, it is useful to make this concept more precise. Suppose that $X_1, X_2, \ldots$ is a sequence of random variables. We write $\mathcal{F}_n$ as a shorthand for the information available in $X_1, \ldots, X_n$. We are assuming that information is never lost. By convention, $\mathcal{F}_0$ will be "no information". If Y is a random variable with $\mathbb{E}[|Y|] < \infty$, then the conditional expectation $E(Y \mid \mathcal{F}_n)$ is the best guess for Y given the information in $\mathcal{F}_n$. Since $\mathcal{F}_0$ contains no information, $E(Y \mid \mathcal{F}_0) = \mathbb{E}[Y]$.

$\diamondsuit$ We use the "blackboard bold" notation $\mathbb{E}$ for expectations (which are numbers), but we use E for conditional expectations which are random variables (since their values depend on $X_1, \ldots, X_n$). We hope that this makes the concept easier to learn. However, most texts use the same typeface for both expectation and conditional expectation.

Let us list some of the properties that the conditional expectation has.

- The random variable $E(Y \mid \mathcal{F}_n)$ is a function of $X_1, \ldots, X_n$.

For each possible value for the random vector $(X_1, \ldots, X_n)$, there is the conditional expectation given that value. We say that $E(Y \mid \mathcal{F}_n)$ is $\mathcal{F}_n$-*measurable*. When computing $E(Y \mid \mathcal{F}_n)$ we treat $X_1, \ldots, X_n$ as constants and then in the end have an expression in terms of $X_1, \ldots, X_n$.

- To say that Y is $\mathcal{F}_n$-measurable means that if we treat $X_1, \ldots, X_n$ as constants, then Y is a constant. Hence, if the random variable Y is $\mathcal{F}_n$-measurable,

$$E(Y \mid \mathcal{F}_n) = Y.$$

Expectation is a linear operation, and this is still true for conditional expectation,

- If Y, Z are random variables and a, b are constants, then

$$E(aY + bZ \mid \mathcal{F}_n) = a\, E(Y \mid \mathcal{F}_n) + b\, E(Z \mid \mathcal{F}_n).$$

We can generalize this. If Z is $\mathcal{F}_n$-measurable, then we can treat Z like a constant. This implies the following.

- If Z is $\mathcal{F}_n$-measurable, then

$$E(ZY \mid \mathcal{F}_n) = Z\, E(Y \mid \mathcal{F}_n). \tag{3.7}$$

The conditional expectation $E(Y \mid \mathcal{F}_n)$ is a random variable that depends on the value of $X_1, \ldots, X_n$. Hence, we can consider $\mathbb{E}[E(Y \mid \mathcal{F}_n)]$. This can be considered as the operation of first averaging over all the randomness other than that given by $X_1, \ldots, X_n$ and then averaging over the randomness in $X_1, \ldots, X_n$. This two stage process should give the same result as that given by averaging all at once, therefore

- $\mathbb{E}\left[E(Y \mid \mathcal{F}_n)\right] = \mathbb{E}[Y]$.

Let us combine the last two properties. Suppose V is an event that is $\mathcal{F}_n$-*measurable*, i.e., an event that depends only on the values of $X_1, \ldots, X_n$. Then the indicator random variable 1_V is an $\mathcal{F}_n$-measurable random variable and (3.7) implies that $E(1_V\, Y \mid \mathcal{F}_n) = 1_V\, \mathbb{E}(Y \mid \mathcal{F}_n)$. Taking expectations of both sides, we get the following.

- If V is an $\mathcal{F}_n$-measurable event,

$$\mathbb{E}\left[1_V\, Y\right] = \mathbb{E}\left[1_V\, E(Y \mid \mathcal{F}_n)\right]. \tag{3.8}$$

At this point, we have given many of the properties that we expect conditional expectation to satisfy, but we have not given a formal definition. It turns out that (3.8) is the property that characterizes the conditional expectation.

Definition 3.1. The conditional expectation $E(Y \mid \mathcal{F}_n)$ is the unique $\mathcal{F}_n$-measurable random variable such that (3.8) holds for each $\mathcal{F}_n$-measurable event V.

In order to show this is well defined, we must show that there exists a unique such $\mathcal{F}_n$-measurable random variable. To show uniqueness, suppose that W, Z were two $\mathcal{F}_n$-measurable random variables with

$$\mathbb{E}\left[1_V\, W\right] = \mathbb{E}\left[1_V\, Y\right] = \mathbb{E}\left[1_V\, Z\right].$$

Then for every $\mathcal{F}_n$-measurable event V, we have

$$\mathbb{E}\left[(W - Z)\, 1_V\right] = 0.$$

If we apply this to the events $\{W - Z > 0\}$ and $\{W - Z < 0\}$ we can see that these events must have probability zero and hence $\mathbb{P}\{W = Z\} = 1$. (This uniqueness up to an event of probability zero is how we define uniqueness in this definition.) Existence takes more work and generally is established making use of a theorem from measure theory, the Radon-Nikodym theorem. We will just assume the existence in this book. In many cases, we will be able to give the conditional expectation explicitly so we will not need to use the existence theorem.

For a rigorous approach to conditional expectation, one needs to verify all the bulleted properties for the conditional expectation from our definition. This is not difficult. There are two other properties that will be important to us. First, if Y is independent of $\mathcal{F}_n$, then any information about $X_1, \ldots, X_n$ should be irrelevant.

- If Y is independent of $X_1, \ldots, X_n$, then

$$E(Y \mid \mathcal{F}_n) = \mathbb{E}[Y].$$

To justify this from the definition note that $\mathbb{E}[Y]$ is $\mathcal{F}_n$-measurable ($\mathbb{E}[Y]$ is a constant random variable) and if V is $\mathcal{F}_n$-measurable, then Y and 1_V are independent. Hence,

$$\mathbb{E}\left[1_V\, Y\right] = \mathbb{E}\left[1_V\right]\, \mathbb{E}\left[Y\right] = \mathbb{E}\left[1_V\, \mathbb{E}\left[Y\right]\right].$$

♢ When we say Y is independent of $X_1, \ldots, X_n$ (or, equivalently, independent of $\mathcal{F}_n$) we mean that none of the information in $X_1, \ldots, X_n$ is useful

for determining Y. It is possible for Y to be independent of each of the X_j separately but not independent of $X_1, \ldots, X_n$ (Exercise 3.1).

One final property is a "projection" property for conditional expectation.

- If $m < n$,

$$E\left(E(Y \mid \mathcal{F}_n) \mid \mathcal{F}_m\right) = E\left(Y \mid \mathcal{F}_m\right).$$

We use the definition to verify this. Clearly, the right hand side is $\mathcal{F}_m$-measurable, so we need to show that for all $\mathcal{F}_m$-measurable events

$$\mathbb{E}\left[E(Y \mid \mathcal{F}_n)\, 1_V\right] = \mathbb{E}\left[E(Y \mid \mathcal{F}_m)\, 1_V\right]. \tag{3.9}$$

We leave this as Exercise 3.2.

We define the conditional variance in the natural way

$$\mathrm{Var}[Y \mid \mathcal{F}_n] = E\left((Y - E(Y \mid F_n))^2 \mid \mathcal{F}_n\right).$$

By expanding the square and using linearity and (3.7), we get the usual alternative formula for the conditional variance,

$$\mathrm{Var}[Y \mid \mathcal{F}_n] = E(Y^2 \mid \mathcal{F}_n) - \left[E(Y \mid \mathcal{F}_n)\right]^2.$$

If $\mathbb{E}[Y^2] < \infty$, the conditional variance is well defined.

3.3. Definition of martingale

Definition 3.2. If $X_0, X_1, X_2, \ldots$ and $M_0, M_1, M_2, \ldots$ are sequences of random variables, then $\{M_n\}$ is called a *martingale* with respect to $\{X_n\}$ if $\mathbb{E}[|M_n|] < \infty$ for each n, each M_n is $\mathcal{F}_n$-measurable, and for all n,

$$E\left(M_{n+1} \mid \mathcal{F}_n\right) = M_n. \tag{3.10}$$

Here $\mathcal{F}_n$ denotes the information in $X_0, X_1, \ldots, X_n$.

The projection rule for conditional expectation shows that if M_n is a martingale, then

$$E\left[M_{n+2} \mid \mathcal{F}_n\right] = E\left(E(M_{n+2} \mid \mathcal{F}_{n+1}) \mid \mathcal{F}_n\right) = E(M_{n+1} \mid \mathcal{F}_n) = M_n,$$

and similarly for all $n < m$,

$$E(M_m \mid \mathcal{F}_n) = M_n.$$

In particular, $\mathbb{E}(M_n \mid \mathcal{F}_0) = M_0$ and

$$\mathbb{E}\,[M_n] = \mathbb{E}\,[E(M_n \mid \mathcal{F}_0)] = \mathbb{E}[M_0].$$

A number of examples were given in Section 3.1. For another, consider a time homogeneous, Markov chain on a finite or countably infinite state space S. In other words, we have random variables $Y_0, Y_1, Y_2, \ldots$, taking values in S, such that

$$\mathbb{P}\{Y_{n+1} = y \mid Y_0, \ldots, Y_n\} = \mathbb{P}\{Y_{n+1} = y \mid Y_n\} = p(Y_n, y).$$

Here $p : S \times S \to [0,1]$ are the transition probabilities. Suppose $f : S \to \mathbb{R}$ is a function. Then if $\mathcal{F}_n$ denotes the information in $Y_0, \ldots, Y_n$,

$$E(f(Y_{n+1}) \mid \mathcal{F}_n) = E(f(Y_{n+1}) \mid Y_n) = \sum_y p(Y_n, y)\, f(y).$$

The condition on f so that $M_n = f(X_n)$ is a martingale is that the function is *harmonic* with respect to the Markov chain which means that at every x,

$$f(x) = \sum_y p(x,y)\, f(y). \tag{3.11}$$

If the condition (3.11) holds only for a subset $S_1 \subset S$, then we can guarantee that $f(M_n)$ is a martingale if we change the Markov chain so it does not move once it leaves S_1, i.e, $p(y,y) = 1$ for $y \in S \setminus S_1$. This shows that there is a very close relationship between martingales and harmonic functions.

A function f is called *subharmonic* or *superharmonic* if (3.11) is replaced by

$$f(x) \le \sum_y p(x,y)\, f(y),$$

or

$$f(x) \ge \sum_y p(x,y)\, f(y),$$

respectively. Using this as motivation, we define a process to be a *submartingale* or *supermartingale* if (3.10) is replaced by

$$E(M_{n+1} \mid \mathcal{F}_n) \ \ge M_n,$$

or

$$E(M_{n+1} \mid \mathcal{F}_n) \le M_n,$$

respectively. In other words, a submartingale is a game in one's favor and a supermartingale is an unfair game.

♢ The terminology can be confusing. The prefix "sub" is used for processes that tend to get bigger and "super" is used for processes that tend to decrease. The terminology was chosen to match the definitions of subharmonic and superharmonic.

3.4. Optional sampling theorem

The most important result about martingales is the optional sampling theorem which states that under certain conditions "you can't beat a fair game". We have already given a "counterexample" to this principle in the martingale betting strategy so we will have to be careful in determining under what conditions the result is true.

Our first result states that one cannot make a game in one's favor (or even against one!) in a finite amount of time. Suppose $M_0, M_1, \ldots$ is a martingale with respect to $X_0, X_1, \ldots$. We think of $M_n - M_{n-1}$ as the winnings on the nth game of a fair game. Before the nth game is played, we are allowed to decide to stop playing. The information in $\mathcal{F}_{n-1}$ may be used in making the decision, but we are not allowed to see the result of the nth game. More mathematically, we say that a *stopping time* is a random variable T taking values in $\{0, 1, 2, \ldots\}$ such that the event $\{T = n\}$ is $\mathcal{F}_n$-measurable. Recalling that $T \wedge n = \min\{T, n\}$, we see that $M_{n\wedge T}$ equals the value of the stopped process at time n,

$$M_{n\wedge T} = \begin{cases} M_n & \text{if } T > n, \\ M_T & \text{if } T \le n. \end{cases}$$

Proposition 3.3. *If $M_0, M_1, \ldots$ is a martingale and T is a stopping time each with respect to $X_0, X_1, \ldots$, then the process $Y_n = M_{T\wedge n}$ is a martingale. In particular, for each n, $\mathbb{E}[M_{T\wedge n}] = \mathbb{E}[M_0]$.*

Proof. Using the indicator function notation, we can write

$$M_{T\wedge n} = \sum_{j=0}^{n-1} M_j \, 1\{T = j\} + M_n \, 1\{T \ge n\}.$$

In other words, if we stop before time n we get the value when we stop; otherwise, we get the value at time n. Then,

$$\begin{aligned} M_{T\wedge(n+1)} - M_{T\wedge n} &= M_{n+1}\,1\{T \geq n+1\} - M_n\,1\{T \geq n+1\} \\ &= [M_{n+1} - M_n]\,1\{T > n\}. \end{aligned}$$

The event $\{T > n\}$ which corresponds to not stopping by time n is $\mathcal{F}_n$-measurable since the decision not to stop by time n uses only the information in $\mathcal{F}_n$. Hence, by (3.8),

$$E\left([M_{n+1} - M_n]\,1\{T > n\} \mid \mathcal{F}_n\right) = \\ 1\{T > n\}\,E\left(M_{n+1} - M_n \mid \mathcal{F}_n\right) = 0.$$

□

We will now consider the question: for which martingales and stopping times can we conclude that the game is fair in the sense that $\mathbb{E}[M_T] = \mathbb{E}[M_0]$? The last proposition shows that this is the case if T is bounded with probability one, for in this case $M_T = M_{T\wedge n}$ for all large n. Let us try to prove the fact and on the way try to figure out what assumptions are needed (the martingale betting strategy tells us that there certainly need to be more assumptions).

♢ In mathematics texts and papers, theorems are stated with certain assumptions and then proof are given. However, this is not how the research process goes. Often there is a result that one wants, one writes a proof, and in the process one discovers what assumptions are needed in order for the argument to be valid. We will approach the optional sampling theorem from this research perspective.

Suppose $M_0, M_1, \ldots$ is a martingale with respect to $X_0, X_1, \ldots$ and T is a stopping time. Assume that we will eventually stop,

$$\mathbb{P}\{T < \infty\} = 1. \tag{3.12}$$

This is a weaker assumption than saying there is a K such that $\mathbb{P}\{T \leq K\} = 1$. The martingale betting strategy satisfies (3.12) since the probability that $T \leq n$ is the same as the probability that one has not lost the game n times in a row, which equals $1 - 2^{-n}$. Hence

$$\mathbb{P}\{T < \infty\} = \lim_{n\to\infty} \mathbb{P}\{T \leq n\} = 1.$$

For each n, note that

$$M_T = M_{T\wedge n} + M_T\, 1\{T > n\} - M_n\, 1\{T > n\}.$$

By Proposition 3.3, we know that $\mathbb{E}[M_{T\wedge n}] = \mathbb{E}[M_0]$. So to conclude that $\mathbb{E}[M_0] = \mathbb{E}[M_T]$ it suffices to show that

$$\lim_{n\to\infty} \mathbb{E}\left[M_T\, 1\{T > n\}\right] = 0, \tag{3.13}$$

and

$$\lim_{n\to\infty} \mathbb{E}\left[M_n\, 1\{T > n\}\right] = 0. \tag{3.14}$$

Let us start with (3.13). Note that with probability one, $1\{T > n\} \to 0$; this is another way of stating (3.12). This is not quite enough to sufficient to conclude (3.13), but we can conclude it if $\mathbb{E}\left[|M_T|\right] < \infty$ for then we can use the dominated convergence theorem. In the case of the martingale betting strategy, $W_T \equiv 1$, so $\mathbb{E}[W_T] < \infty$.

The equation (3.14) is trickier. This is the one that the martingale betting strategy does not satisfy. If one has lost n times in a row, $W_n = 1 - 2^n$. This happens with probability 2^{-n} and hence

$$\mathbb{E}\left[W_n\, 1\{T > n\}\right] = \left[1 - 2^n\right] 2^{-n} \to -1.$$

We will make (3.14) an assumption in the theorem which we have now proved.

Theorem 3.4 (Optional Sampling Theorem)**.** *Suppose M_0 , M_1, ... is a martingale and T is a stopping time both with respect to X_0, X_1, ... Suppose that $\mathbb{P}\{T < \infty\} = 1$, $\mathbb{E}\left[|M_T|\right] < \infty$, and* (3.14) *holds. Then,*

$$\mathbb{E}[M_T] = \mathbb{E}[M_0].$$

3.4.1. Gambler's ruin estimate. Suppose S_n is one-dimensional simple random walk starting at the origin as in Section 3.1.1. Suppose j, k are positive integers and let

$$T = \min\{n : S_n = -j \text{ or } k\}.$$

It is easy to check that $\mathbb{P}\{T < \infty\} = 1$, and

$$\begin{aligned}
\mathbb{E}\left[S_T\right] &= -j\,\mathbb{P}\{S_T = -j\} + k\,\mathbb{P}\{S_T = k\} \\
&= -j\,\left[1 - \mathbb{P}\{S_T = k\}\right] + k\,\mathbb{P}\{S_T = k\} \\
&= -j + (j+k)\,\mathbb{P}\{S_T = k\}.
\end{aligned}$$

Since $S_{T\wedge n}$ is bounded, it is easy to see that it satisfies the conditions of the optional sampling theorem, and hence

$$\mathbb{E}\,[S_T] = S_0 = 0.$$

Solving for $\mathbb{P}\{S_T = k\}$ yields

$$\mathbb{P}\{S_T = k\} = \frac{j}{j+k}.$$

3.4.2. Asymmetric random walk. Suppose $\frac{1}{2} < p < 1$, and X_1, X_2, ... are independent random variables with $\mathbb{P}\{X_j = 1\} = 1 - \mathbb{P}\{X_j = -1\} = p$. Let $S_0 = 0$ and

$$S_n = X_1 + \cdots + X_n.$$

We will find a useful martingale by finding a harmonic function for the random walk. The function f is harmonic if

$$E\,[f(S_n) \mid S_{n-1}] = f(S_{n-1}).$$

Writing this out, we get the relation

$$f(x) = p\,f(x+1) + (1-p)\,f(x-1).$$

A function that satisfies this equation is

$$f(x) = \left(\frac{1-p}{p}\right)^x,$$

and using this, we see that

$$M_n = \left(\frac{1-p}{p}\right)^{S_n}$$

is a martingale. Let j, k, T be as in the previous section. Note that $M_0 = 1$, and since $M_{T\wedge n}$ is a bounded martingale

$$\mathbb{E}\,[M_T] = M_0 = 1.$$

If $r = \mathbb{P}\{S_T = k\}$, then

$$\mathbb{E}\,[M_T] = (1-r)\,\left(\frac{1-p}{p}\right)^{-j} + r\,\left(\frac{1-p}{p}\right)^{k}.$$

Solving for r gives

$$\mathbb{P}\{S_T = k\} = \frac{1-\theta^j}{1-\theta^{j+k}} \quad \text{where} \quad \theta = \frac{1-p}{p} < 1.$$

Note what happens as $k \to \infty$. Then

$$\mathbb{P}\{ \text{ random walker ever reach } -j\} = \lim_{k\to\infty} \mathbb{P}\{S_T = j\} = \theta^j.$$

3.4.3. Polya's urn. We will use the optional sampling theorem to deduce some facts about Polya's urn as in Section 3.1.3. To be more general, assume that we start with an urn with J red balls and K green balls so that the fraction of red balls at the start is $b = J/(J + K)$. We let M_n denote the fraction of red balls at time n so that $M_0 = b$. Let $a < b$, and let T denote the smallest n such that $M_n \le a$. Then T is a stopping time although it is possible that $T = \infty$. We will give an upper bound on the probability that $T < \infty$. The optional sampling theorem in the form of Proposition 3.3 implies

$$b = \mathbb{E}[M_0] = \mathbb{E}[M_{T\wedge n}] =$$

$$\mathbb{P}\{T \le n\}\, \mathbb{E}[M_T \mid T \le n] + [1 - \mathbb{P}\{T \le n\}]\, \mathbb{E}\left[M_T \mid T > n\right].$$

Solving for $\mathbb{P}\{T \le n\}$ gives

$$\mathbb{P}\{T \le n\} = \frac{\mathbb{E}\left[M_T \mid T > n\right] - b}{\mathbb{E}\left[M_T \mid T > n\right] - \mathbb{E}\left[M_T \mid T \le n\right]}. \tag{3.15}$$

A little thought (verify this!) shows that

$$0 \le \mathbb{E}\left[M_T \mid T \le n\right] \le a, \quad b \le \mathbb{E}\left[M_T \mid T > n\right] \le 1,$$

and (with the aid of some simple calculus) we can see that the right hand side of (3.15) is at most equal to $(1 - b)/(1 - a)$. Therefore,

$$\mathbb{P}\{T < \infty\} = \lim_{n\to\infty} \mathbb{P}\{T \le n\} \le \frac{1-b}{1-a}.$$

By essentially the same argument (or by the same fact for the fraction of green balls), we can see that if $M_0 \le a$, then the probability that the fraction of balls ever gets as large as b is at most a/b.

Let us continue this. Suppose $M_0 = b$. Consider the event that at some time the fraction of red ball becomes less than or equal to a and then after that time it becomes greater than or equal to b again. By our argument, we see that the probability of this event is at most

$$\left(\frac{1-b}{1-a}\right) \frac{a}{b}.$$

Let us call such an event an (a, b) fluctuation. (In the martingale literature, the terms *upcrossing* and *downcrossing* are used. For us an

(a, b) fluctuation is a downcrossing of a followed by an upcrossing of b.) The probability of at least k (a, b) fluctuations (to be precise, an (a, b) fluctuation, followed by another (a, b) fluctuation, followed by another, for a total of k fluctuations) is no larger than

$$\left[\left(\frac{1-b}{1-a}\right)\frac{a}{b}\right]^k.$$

Note that this goes to zero as $k \to \infty$. What we have shown is

$$\mathbb{P}\{\text{there are infinitely many } (a, b) \text{ fluctuations}\} = 0.$$

Up to now we have fixed $a < b$. But since there are only countably many rationals, we can see that

$$\mathbb{P}\{\exists \text{ rational } a < b \text{ with infinitely many } (a, b) \text{ fluctuations}\} = 0.$$

This leads to an interesting conclusion. We leave this simple fact as an exercise.

Lemma 3.5. *Suppose $x_0, x_1, x_2, \ldots$ is a sequence of numbers such that for every rational $a < b$, the sequence does not have an infinite number of (a, b) fluctuations. Then there exists $C \in [-\infty, \infty]$ such that*

$$\lim_{n\to\infty} x_n = C.$$

If the sequence is bounded, then $C \in (-\infty, \infty)$.

Proof. Exercise 3.8. □

Given this, then we can see we have established the following: with probability one, there is an $M_\infty \in [0, 1]$ such that

$$\lim_{n\to\infty} M_n = M_\infty.$$

We should point out that M_∞ is a *random variable* , i.e., different realizations of the experiment of drawing balls will give different values for M_∞ (see Exercises 3.9 and 3.10).

3.5. Martingale convergence theorem

We will generalize the convergence fact that was just established for the Polya urn model. The next theorem shows that under a relatively weak condition, we can guarantee that a martingale converges.

Theorem 3.6 (Martingale Convergence Theorem)**.** *Suppose M_0, M_1, ... is a martingale with respect to $X_0, X_1, \ldots$. Suppose that there exists $C < \infty$ such that for all n,*

$$\mathbb{E}\left[|M_n|\right] \leq C.$$

Then there is a random variable M_∞ such that with probability one

$$\lim_{n\to\infty} M_n = M_\infty.$$

Before proving the theorem let us consider some examples.

- If $S_n = X_1 + \cdots + X_n$ denotes simple random walk, then S_n is a martingale. However,

  $$\lim_{n\to\infty} \mathbb{E}\left[|S_n|\right] = \infty,$$

  so this does not satisfy the condition of the theorem. Also, S_n does not converge as $n \to \infty$.
- Suppose $M_0, M_1, \ldots$ only take nonnegative values. Then

  $$\mathbb{E}[M_n] = \mathbb{E}[M_0] < \infty,$$

  and hence the conditions are satisfied.
- Let W_n be the winnings in the martingale betting strategy as in (3.1). Then

  $$\mathbb{E}\left[|W_n|\right] = (1 - 2^{-n})\, 1 + 2^{-n}\left[2^n - 1\right] \leq 2.$$

  Therefore, this does satisfy the conditions of the theorem. In fact, with probability one,

  $$\lim_{n\to\infty} W_n = W_\infty,$$

  where $W_\infty \equiv 1$. Note that $\mathbb{E}[W_\infty] \neq \mathbb{E}[W_0]$. In particular, it is *not* a conclusion of the martingale convergence theorem that $\mathbb{E}[M_\infty] = \mathbb{E}[M_0]$.

Proof. For ease we will assume that $M_0 = 0$; otherwise we can consider $M_0 - M_0, M_1 - M_0, \ldots$. As in the Polya urn case, we will show that for every $a < b$, the probability that there are infinitely many (a, b) fluctuations is zero. The basic idea of the proof can be considered a financial strategy: "buy low, sell high". To make this precise, let us write

$$M_n = \Delta_1 + \cdots + \Delta_n$$

where $\Delta_j = M_j - M_{j-1}$. We will consider

$$W_n = \sum_{j=1}^{n} B_j \, \Delta_j,$$

where B_j are "bets" (or "investments") which equal zero or one and, as in Section 3.1.2, the bet B_j must be measurable with respect to $\mathcal{F}_{j-1}$. The total winnings W_n is a martingale which can be seen by

$$\begin{aligned} E[W_n \mid \mathcal{F}_{n-1}] &= E[W_{n-1} + B_n(M_n - M_{n-1}) \mid \mathcal{F}_{n-1}] \\ &= E[W_{n-1} \mid \mathcal{F}_{n-1}] + E[B_n(M_n - M_{n-1}) \mid \mathcal{F}_{n-1}] \\ &= W_{n-1} + B_n \, E[M_n - M_{n-1} \mid \mathcal{F}_{n-1}] = W_{n-1}. \end{aligned}$$

The third equality uses the fact that W_{n-1} and B_n are measurable with respect to $\mathcal{F}_{n-1}$. In particular,

$$\mathbb{E}[W_n] = \mathbb{E}[W_0] = 0.$$

This is true for every acceptable betting rule B_n. We now choose a particular rule. For ease, we will assume $a \leq 0$ (the $a > 0$ case is done similarly). We let $B_j = 0$ for all $j < T$ where T is the smallest n such that $M_{n-1} \leq a$. We let $B_T = 1$ and we keep $B_n = 1$ until the first time $m > n$ that $M_m \geq b$. At this point, we change the bet to zero and keep it at zero until the martingale M drops below a again. Every time we have an (a, b) fluctuation, we gain at least $b - a$ in this strategy. Let J_n denote the number of (a, b) fluctuations by time n. Then

$$W_n \geq J_n \, (b - a) + (M_n - a) \, 1\{M_n \leq a\} \geq J_n \, (b - a) - |M_n|.$$

The term

$$(M_n - a) \, 1\{M_n \leq a\},$$

which is nonpositive, comes from considering the amount we have lost in the last part of the process where we started "buying" at

the last drop below a before time n. Since $\mathbb{E}[W_n] = 0$, we can take expectations of both sides to conclude

$$\mathbb{E}[J_n] \le \frac{\mathbb{E}[|M_n|]}{b-a} \le \frac{C}{b-a}.$$

The right hand side does not depend on n, so if $J = J_\infty$ denotes the total number of (a, b) fluctuations,

$$\mathbb{E}[J] \le \frac{C}{b-a} < \infty.$$

If a nonnegative random variable has finite expectation, then with probability one it is finite. Hence the number of (a, b) fluctuations is finite with probability one. As in the Polya urn case, we can use Lemma 3.5 to see that with probability one the limit

$$\lim_{n\to\infty} M_n = M_\infty$$

exists. We also claim that with probability one $M_\infty \in (-\infty, \infty)$. This can be seen from the estimate

$$\mathbb{P}\{|M_n| \ge K\} \le K^{-1}\,\mathbb{E}[|M_n|] \le \frac{C}{K},$$

from which one can conclude

$$\mathbb{P}\{|M_\infty| \ge K\} \le \frac{C}{K}.$$

□

3.5.1. Random Cantor set. Consider the random Cantor set from Section 3.1.4. We use the notation from that section. Recall that Y_n denotes the number of individuals at the nth generation and $M_n = \mu^{-n} Y_n$ is a martingale, where μ is the mean number of offspring (or remaining subintervals) per individual (interval). In the example $Y_0 = M_0 = 1$. Since this is a nonnegative martingale, the martingale convergence theorem implies that there is an M_∞ such that with probability one,

$$\lim_{n\to\infty} M_n = M_\infty.$$

Proposition 3.7. *If $\mu \le 1$, then $M_\infty = 0$. In fact, with probability one, $Y_n = 0$ for all large n.*

Proof. Note that $\mathbb{E}[Y_n] = \mu^n\,\mathbb{E}[M_n] = \mu^n$. If $\mu < 1$, then

$$\mathbb{P}\{Y_n \geq 1\} = \sum_{k=1}^{\infty} \mathbb{P}\{Y_n = k\} \leq$$

$$\sum_{k=1}^{\infty} k\,\mathbb{P}\{Y_n = k\} = \mathbb{E}[Y_n] = \mu^n \longrightarrow 0.$$

If $\mu = 1$, then $Y_n = M_n$ which implies

$$\lim_{n\to\infty} Y_n = M_\infty.$$

Since Y_n takes on only integer values, the only way that this limit can exist is for Y_n to take the same value for all sufficiently large n. But the nature of the process shows immediately that if $k > 0$, then $\mathbb{P}\{Y_{n+1} = 0 \mid Y_n = k\} > 0$ from which one case see that it cannot be the case that $Y_n = k$ for all large n. □

For $\mu \leq 1$, we see that $\mathbb{E}[M_\infty] \neq \mathbb{E}[M_0]$. In the next section, we will show that if $\mu > 1$, then $\mathbb{E}[M_\infty] = \mathbb{E}[M_0] = 1$. In particular, $\mathbb{P}\{M_\infty \neq 0\} > 0$ which implies that with positive probability $Y_n \to \infty$. (We cannot hope for this to be true with probability one, because there is a positive probability that we will be unlucky early on and die out.) For large n,

$$Y_n \sim M_\infty\,\mu^n.$$

This means that there is geometric growth of the number of offspring with rate μ. The constant factor M_∞ is random and depends on the randomness in the early generations. With the aid of Exercise 3.19, we will see that there are only two possibilities: either $Y_n = 0$ for some n and hence the random Cantor set is empty, or $M_\infty > 0$ so that the number of intervals in the nth generation grows like $M_\infty\,\mu^n$. Roughly speaking, once the population gets large, the growth rate is almost deterministic with rate μ, $Y_{n+1} \sim \mu Y_n$. The constant factor M_∞ is determined by the randomness in the first few generations.

3.6. Uniform integrability

Suppose $M_0, M_1, \ldots$ is a martingale for which there exists M_∞ with

$$M_\infty = \lim_{n\to\infty} M_n.$$

We know that for each n,

$$\mathbb{E}[M_n] = M_0.$$

If we could interchange the limit and expectation, we would have

$$\mathbb{E}\left[M_\infty\right] = \mathbb{E}\left[\lim_{n\to\infty} M_n\right] = \lim_{n\to\infty} \mathbb{E}\left[M_n\right] = M_0.$$

We have already seen examples for which the conclusion is false, so we can see that the interchange is not always valid.

This leads to asking: suppose $Y_1, Y_2, \ldots$ is a sequence of random variables such that there is a random variable Y such that with probability one,

$$\lim_{n\to\infty} Y_n = Y.$$

Under what conditions can we conclude that

$$\lim_{n\to\infty} \mathbb{E}\left[Y_n\right] = Y \quad ? \tag{3.16}$$

There are two main theorems that are learned in a first course in measure theory.

- **Monotone Convergence Theorem**. If $0 \le Y_1 \le Y_2 \le \cdots$, then (3.16) is valid.
- **Dominated Convergence Theorem**. If there exists a random variable $Z \ge 0$ with $\mathbb{E}(Z) < \infty$ such that $|Y_n| \le Z$ for every n, then (3.16) is valid.

In this section, we will derive a generalization of the dominated convergence theorem that is very useful in studying martingales.

To motivate our main definition, let us first consider a random variable Z such that $\mathbb{E}[|Z|] < \infty$. Such a random variable is called *integrable*. Note that

$$\mathbb{E}\left[|Z|\right] = \sum_{n=0}^{\infty} \mathbb{E}\left[|Z|\, 1\{n \le |Z| < n+1\}\right].$$

Since the sum on the right is finite, we can see that for every $\epsilon > 0$, there is a $K < \infty$ such that

$$\mathbb{E}\left[|Z|\, 1\{|Z| \ge K\}\right] = \sum_{n=K}^{\infty} \mathbb{E}\left[|Z|\, 1\{n \le |Z| < n+1\}\right] < \epsilon.$$

Our definition builds on this observation.

Definition 3.8. A collection of random variables $\{Y_1, Y_2, \ldots\}$ is *uniformly integrable* if for every $\epsilon > 0$, there is a $K < \infty$ such that for each j,

$$\mathbb{E}\left[|Y_j| \, 1\{|Y_j| \geq K\}\right] < \epsilon.$$

It follows from our discussion about that any collection $\{Y_1\}$ consisting of only one integrable random variable is uniformly integrable. It is not difficult (Exercise 3.15) to show that a finite collection of integrable random variables is uniformly integrable. However, it is possible for an infinite collection of integrable random variables to be not uniformly integrable.

♢ The use of the word uniformly is similar to that in the terms uniform continuity and uniform convergence. For example, a collection of continuous random variables $f_1, f_2, \ldots$ on $[0, 1]$ is uniformly continuous if for every $\epsilon > 0$, there is a $\delta > 0$ such that $|x - y| < \delta$ implies $|f_j(x) - f_j(y)| < \epsilon$. In particular, for a given ϵ, there must be a single δ that works for every function f_j. In the definition of uniform integrability, for each ϵ, there is a K that works for each Y_j.

An example of a collection of integrable random variables that is not uniformly integrable is given by the winnings using the martingale betting strategy as in (3.1). In this case, if $n \geq 2$,

$$\mathbb{E}\left[|W_n| \, 1\{|W_n| \geq 2^n - 1\}\right] = 2^{-n}\left[2^n - 1\right] = 1 - 2^{-n}.$$

Hence if we choose $\epsilon = 1/2$, there is no choice of K such that for each n.

$$\mathbb{E}\left[|W_n| \, 1\{|W_n| \geq K\}\right] \leq \frac{1}{2}.$$

Theorem 3.9. *Suppose $Y_1, Y_2, \ldots$ is a uniformly integrable sequence of random variables such that with probability one*

$$Y = \lim_{n \to \infty} Y_n,$$

then

$$\mathbb{E}[Y] = \lim_{n \to \infty} \mathbb{E}\left[Y_n\right].$$

Proof. Without loss of generality, we may assume $Y \equiv 0$; otherwise, we can consider $Y_1 - Y, Y_2 - Y, \ldots$. Note that

$$|\mathbb{E}[Y_n]| \leq \mathbb{E}\left[|Y_n|\right].$$

Hence it suffices to show that

$$\lim_{n\to\infty} \mathbb{E}\left[|Y_n|\right] = 0,$$

and to show this we need to show that for every $\epsilon > 0$, there is an N such that if $n \geq N$,

$$\mathbb{E}\left[|Y_n|\right] < \epsilon.$$

Let $\epsilon > 0$. Since the collection $\{Y_n\}$ is uniformly integrable, there is a K such that for each n,

$$\mathbb{E}\left[|Y_n|\, 1\{|Y_n| \geq K\}\right] < \frac{\epsilon}{3}. \tag{3.17}$$

With probability one, we know that

$$\lim_{n\to\infty} Y_n = 0.$$

This implies that for each $\epsilon > 0$, there is a (random!) J such that if $n \geq J$,

$$|Y_n| < \frac{\epsilon}{3}.$$

Since J is a random variable, we can find an N such that

$$\mathbb{P}\{J \geq N\} < \frac{\epsilon}{3K}.$$

We now choose $n \geq N$ and write

$$\mathbb{E}\left[|Y_n|\right] = \mathbb{E}\left[|Y_n|\, 1\{|Y_n| \leq \frac{\epsilon}{3}\}\right] +$$

$$\mathbb{E}\left[|Y_n|\, 1\{\frac{\epsilon}{3} < |Y_n| < K\}\right] + \mathbb{E}\left[|Y_n|\, 1\{|Y_n| \geq K\}\right].$$

We estimate the three terms on the right hand side. The third is already done in (3.17) and obviously

$$\mathbb{E}\left[|Y_n|\, 1\{|Y_n| \leq \frac{\epsilon}{3}\}\right] \leq \frac{\epsilon}{3}.$$

Also,

$$\begin{aligned}\mathbb{E}\left[|Y_n|\,1\{\frac{\epsilon}{3}<|Y_n|<K\}\right] &\leq K\,\mathbb{P}\{|Y_n|\geq\frac{\epsilon}{3}\}\\ &\leq K\,\mathbb{P}\{J\geq n\}\\ &\leq K\,\frac{\epsilon}{3K}=\frac{\epsilon}{3}.\end{aligned}$$

By summing we see that if $n \geq N$, then $\mathbb{E}\,[|Y_n|] < \epsilon$. □

The condition of uniform integrability can be hard to verify. It is useful to have some simpler conditions that imply this. Although the next proposition is stated for all $\alpha > 0$, it is most often applied with $\alpha = 1$.

Proposition 3.10. *Suppose $Y_1, Y_2, \ldots$ is a sequence of integrable random variables such at least one of these conditions holds:*

- *There is an integrable Z, such that $|Y_j| \leq |Z|$ for each j.*
- *There exist $\alpha > 0$ and $C < \infty$, such that for each j,*

$$\mathbb{E}[|Y_j|^{1+\alpha}] \leq C.$$

Then the sequence is uniformly integrable.

Proof. The argument to show that the first condition implies uniform integrability is very similar to the argument to show that a single random variable is uniformly integrable; we leave this to the reader. Assume $\mathbb{E}[|Y_j|^{1+\alpha}] \leq C$ for each j. Then,

$$\mathbb{E}\,[|Y_j|\,1\{|Y_j|\geq K\}] \leq K^{-\alpha}\,\mathbb{E}\left[|Y_j|^{1+\alpha}\,1\{|Y_j|\geq K\}\right] \leq K^{-\alpha}\,C.$$

In particular, if $\epsilon > 0$ is given and $K > (C/\epsilon)^{1/\alpha}$, then for all j,

$$\mathbb{E}\,[|Y_j|\,1\{|Y_j|\geq K\}] < \epsilon.$$

□

3.6.1. Random Cantor set. Let us return to the random Cantor set with $\mu > 1$. We use the notation of Section 3.1.4.

Proposition 3.11. *Suppose $\mu > 1$. Then there exists $C < \infty$ such that for all n,*

$$\mathbb{E}\left[M_n^2\right] \leq C.$$

In particular, $M_0, M_1, M_2, \ldots$ is a uniformly integrable collection of random variables.

Proof. We start with (3.6) that tell us that

$$\begin{aligned}\mathbb{E}[Y_{n+1}^2 \mid Y_n = k] &= \mathrm{Var}[Y_{n+1} \mid Y_n = k] + \mathbb{E}[Y_{n+1}^2 \mid Y_n = k] \\ &= k\sigma^2 + k^2\,\mu^2.\end{aligned}$$

Therefore,

$$\begin{aligned}\mathbb{E}[Y_{n+1}^2] &= \sum_{k=0}^{\infty} \mathbb{P}\{Y_n = k\}\,\mathbb{E}[Y_{n+1}^2 \mid Y_n = k] \\ &= \sigma^2 \left[\sum_{k=0}^{\infty} \mathbb{P}\{Y_n = k\}\,k\right] + \mu^2 \left[\sum_{k=0}^{\infty} \mathbb{P}\{Y_n = k\}\,k^2\right] \\ &= \sigma^2\,\mathbb{E}[Y_n] + \mu^2\,\mathbb{E}[Y_n^2] = \sigma^2\,\mu^n + \mu^2\,\mathbb{E}[Y_n^2].\end{aligned}$$

Therefore,

$$\mathbb{E}\left[M_{n+1}^2\right] = \mu^{-2(n+1)}\,\mathbb{E}\left[Y_{n+1}^2\right] = \sigma^2\,\mu^{-n-2} + \mathbb{E}\left[M_n^2\right].$$

By induction, we see that

$$\mathbb{E}\left[M_{n+1}^2\right] = \sigma^2 \sum_{j=0}^{n} \mu^{-j-2} < \sigma^2 \sum_{j=0}^{\infty} \mu^{-j-2} < \infty.$$

Note that the last inequality uses $\mu > 1$. Uniform integrability follows from Proposition 3.10. □

Exercises

Exercise 3.1. Give an example of random variables Y, X_1, X_2 such that Y is independent of X_1, Y is independent of X_2, but Y is not independent of X_1, X_2.

Exercise 3.2. Prove (3.9).

Exercise 3.3. Suppose Y is a random variable with $\mathbb{E}[Y] = 0$ and $\mathbb{E}[Y^2] < \infty$. Let $\mathcal{F} = \mathcal{F}_n$ be the information contained in $X_1, \ldots, X_n$ and let $Z = \mathbb{E}(Y \mid \mathcal{F})$. Show that

$$\mathbb{E}\left[Y^2\right] = \mathbb{E}\left[Z^2\right] + \mathbb{E}\left[(Y - Z)^2\right].$$

Exercise 3.4. Suppose that M_n is a martingale with respect to $\{\mathcal{F}_n\}$ with $M_0 = 0$ and $\mathbb{E}[M_n^2] < \infty$. Show that

$$\mathbb{E}\left[M_n^2\right] = \sum_{j=1}^{n} \mathbb{E}[(M_j - M_{j-1})^2].$$

Exercise 3.5. Let $M_0, M_1, \ldots$ be a martingale with respect to $\mathcal{F}_n$ such that for each n, $\mathbb{E}[M_n^2] < \infty$.

(1) Show that if Y is a random variable with $\mathbb{E}[Y^2] < \infty$, then $E[Y^2 \mid \mathcal{F}] \geq (E[Y \mid F])^2$. Hint: consider $[Y - E(Y \mid \mathcal{F}_n)]^2$.

(2) Show that if $Y_n = M_n^2$, then $Y_0, Y_1, \ldots$ is a submartingale with respect to $\mathcal{F}_n$.

Exercise 3.6. Suppose $Y_0, Y_1, \ldots$ is a submartingale with $Y_j \geq 0$ for each j. Let

$$\overline{Y}_n = \max\{Y_0, Y_1, \ldots, Y_n\}.$$

Show that for every $a > 0$,

$$\mathbb{P}\left\{\overline{Y}_n \geq a\right\} \leq a^{-1}\,\mathbb{E}[Y_n].$$

Hint: Let $T = \min\{j : Y_j \geq a\}$ and let E_j be the event $\{T = j\}$. Show that for $j \leq n$,

$$a\,\mathbb{P}\{T = j\} \leq \mathbb{E}\left[Y_j\, 1\{T = j\}\right] \leq \mathbb{E}\left[Y_n\, 1\{T = j\}\right].$$

Exercise 3.7. Use the previous two exercises to conclude the following generalization of Chebyshev's inequality. Suppose $M_0, M_1, \ldots$ is a martingale with respect to X_0, X_1, X_2, ... Then for every positive integer n and every $a > 0$,

$$\mathbb{P}\left\{\max\{|M_0|, \ldots, |M_n|\} \geq a\right\} \leq a^{-2}\,\mathbb{E}\left[M_n^2\right].$$

Exercise 3.8. Prove Lemma 3.5.

Exercise 3.9. Suppose M_n is the fraction of red balls in the Polya urn where at time 0 there exist one red and one green ball.

(1) Show that for every n, the distribution of M_n is uniform on the set

$$\left\{\frac{1}{n+2}, \frac{2}{n+2}, \ldots, \frac{n+1}{n+2}\right\}.$$

Hint: use induction.

(2) Let $M_\infty = \lim_{n\to\infty} M_n$. What is the distribution of M_∞?

Exercise 3.10. Do a computer simulation of Polya's urn. Start with one red ball and one green ball and do the ball selection until there are 502 balls and then continue until there are 1002 balls.

(1) Use the simulations to try to guess (without the benefit of the previous exercise) the distribution of M_{500} and M_{1000}.

(2) For each run of the simulation, compare M_{500} and M_{1000} and see how much they vary.

Exercise 3.11. Suppose in Polya's urn there are k different colors and we initially start with one ball of each color. At each integer time, a ball is chosen from the urn and it and another ball of the same color are returned to the urn.

(1) Let $M_{n,j}$ denote the fraction of balls of color j after n steps. Show that with probability one the limits

$$M_{\infty,j} = \lim_{n\to\infty} M_{n,j}$$

exist.

(2) In the case $k = 3$ find the distribution of the random vector $(M_{\infty,1}, M_{\infty,2}, M_{\infty,3})$.

Exercise 3.12. Consider random walk on the set $\{0, 1, \ldots, N\}$ where one stops when one reaches 0 or N and otherwise at each step move to the right one unit with probability p and left one unit with probability $1 - p$. Suppose that the initial position is $S_0 = j \in \{1, \ldots, N\}$. Let

$$T = \min\{n : S_n = 0 \text{ or } N\}.$$

Show that there exists a $\rho > 0$ and $C < \infty$ such

$$\mathbb{P}\{T > n\} \le C\, e^{-\rho n}.$$

Conclude that $\mathbb{E}[T] < \infty$. Hint: Show that there is a $u > 0$ such that for all n,

$$\mathbb{P}\{T > N + n \mid T > n\} \le 1 - u.$$

Exercise 3.13. Let S_n and T be as in Section 3.4.1.

(1) Let
$$M_n = S_n^2 - n.$$
Show that M_n is a martingale with respect to S_0, S_1, ...

(2) Show that
$$\lim_{n\to\infty} \mathbb{E}\left[|M_n|\, 1\{T > n\}\right] = 0.$$
Hint: Exercise 3.12 may be helpful.

(3) Use the optional sampling theorem to show that
$$\mathbb{E}\left[T\right] = jk.$$

Exercise 3.14. Let S_n, T be as in Section 3.4.2.

(1) Show that
$$M_n = S_n + n(1-2p)$$
is a martingale with respect to $S_0, S_1, S_2, \ldots$.

(2) Use this martingale and the optional sampling theorem to compute $\mathbb{E}[T]$.

Exercise 3.15. Show that any finite collection $Y_1, \ldots, Y_m$ of integrable random variables is uniformly integrable. More generally, show that if $Z_1, Z_2, \ldots$ is a uniformly integrable collection, then so is $Y_1, \ldots, Y_m, Z_1, Z_2, \ldots$.

Exercise 3.16. Let $X_1, X_2, \ldots,$ be independent random variables each with
$$\mathbb{P}\{X_1 = 2\} = \frac{1}{3}, \qquad \mathbb{P}\left\{X_j = \frac{1}{2}\right\} = \frac{2}{3}.$$
Let $M_0 = 1$ and for $j \geq 1$,
$$M_n = X_1\, X_2 \cdots X_n.$$

(1) Show that M_n is a martingale with respect to M_0, X_1, X_2, ...

(2) Use the martingale convergence theorem to show that there is an M_∞ such that with probability one
$$\lim_{n\to\infty} M_n = M_\infty.$$

(3) Show that $M_\infty \equiv 0$. (Hint: consider $\log M_n$ and use the law of large numbers.)

(4) Show that the sequence $M_0, M_1, M_2, \ldots$ is not uniformly integrable.

(5) Show that for every $\alpha > 0$,

$$\sup_n \mathbb{E}\left[M_n^{1+\alpha}\right] = \infty.$$

Exercise 3.17. Suppose $X_1, X_2, \ldots$ are independent random variables with

$$\mathbb{P}\{X_j = 1\} = \mathbb{P}\{X_j = -1\} = \frac{1}{2}.$$

Let $M_0 = 0$ and for $n \geq 1$,

$$M_n = \sum_{j=1}^{n} \frac{X_j}{j}.$$

M_n can be considered as a random harmonic series.

(1) Show that M_n is a martingale.

(2) Show that for each n,

$$\mathbb{E}\left[M_n^2\right] \leq \frac{\pi^2}{6}.$$

(3) Use the martingale convergence theorem to show that there exists M_∞ such that with probability one

$$\lim_{n\to\infty} M_n = M_\infty.$$

In other words, the random harmonic series converges.

Exercise 3.18. Suppose $Y_1, Y_2, \ldots$ is a uniformly integrable collection of random variables. Show that the following holds. If $\epsilon > 0$, there exists $\delta > 0$ such that if V is an event with $\mathbb{P}(V) < \delta$, then for every j,

$$\mathbb{E}\left[|Y_j|\, 1_V\right] < \epsilon.$$

Exercise 3.19. Consider the random Cantor set with $\mu > 1$ and $Y_0 = 1$. We have shown that $\mathbb{E}[M_\infty] = \mathbb{E}[M_0] = 1$ which implies that $q := \mathbb{P}\{M_\infty > 0\} > 0$. Let p be the probability that the random Cantor set is empty. We have seen that

$$p = \mathbb{P}\{\text{there exists } \ n \ \text{ with } \ Y_n = 0\}.$$

If $Y_n = 0$ for some n, then $Y_m = 0$ for $m \geq n$ and $M_\infty = 0$. The goal of this exercise is to prove the converse, that is, $q = 1 - p$.

(1) Let $T_k = \min\{n : Y_n \geq k\}$. Explain why
$$\mathbb{P}\{M_\infty > 0 \mid T_k < \infty\} \geq 1 - (1-q)^k.$$
(2) Show that for each k with probability one one of two things happens: either $T_k < \infty$ or $Y_n = 0$ for some n.

Exercise 3.20. This is a continuation of Exercise 3.19. We will find the *extinction probability* p. Let r_j denote the probability that in the first iteration of the random Cantor set there are exactly j intervals. Let ϕ be the function
$$\phi(s) = \sum_{j=0}^{\infty} r_j \, s^j.$$
(This is really a finite sum and hence ϕ is a polynomial.)

(1) Explain why $p = \phi(p)$.
(2) Show that $\phi(s) < \infty$ for $0 \leq s \leq 1$ and ϕ is an increasing function on this range.
(3) Let $r_{j,n}$ denote the probability that there are exactly j intervals at the nth generation (in particular, $r_{j,1} = r_j$. Let
$$\phi^{(n)}(s) = \sum_{j=0}^{\infty} r_{j,n} \, s^j.$$
Show that for all n, $\phi^{n+1}(s) = \phi(\phi^n(s))$.
(4) Explain why
$$p = \lim_{n\to\infty} \phi^{(n)}(0).$$
(5) Show that p is the smallest positive solution to the equation $p = \phi(p)$.
(6) Suppose the random Cantor set is constructed by dividing intervals into three pieces and choosing each piece with probability 2/3. Find the extinction probability p.

Chapter 4

Fractal Dimension

The idea of dimension arises in a number of areas of mathematics. All reasonable definitions consider the real line as a one-dimensional set, the plane as a two-dimensional set, and in general $\mathbb{R}^d$ to be a d-dimensional set. In linear algebra, one characterizes the dimension as the minimal number of vectors in a spanning set or as the maximum number of vectors in a linearly independent collection. This algebraic definition is not very useful for describing subsets of $\mathbb{R}^d$ that are not vector subspaces. Fractal dimensions are a way to assign dimensions to irregular subsets of $\mathbb{R}^d$. We use the plural *dimensions* to indicate that there are a number of nonequivalent ways of defining this. One particular aspect is that the fractal dimension does not have to be an integer. We will discuss two definitions, *box dimension* and *Hausdorff dimension*.

4.1. Box dimension

Suppose A is a bounded subset of $\mathbb{R}^d$. For every $\epsilon > 0$, let $N_\epsilon = N_\epsilon(A)$ denote the minimal number of closed balls of diameter ϵ needed to cover the set A. The box dimension of A, $D = D(A)$, is defined roughly by the relation

$$N_\epsilon(A) \approx \epsilon^{-D}, \qquad \epsilon \to 0+ .$$

For nonmathematicians, this might suffice as a definition. However, we will want to be more precise. We start by defining the notation $\approx$.

Definition 4.1. If f, g are positive functions on $(0, \infty)$ such that $f(0+) = 0, g(0+) = 0$ or $f(0+) = \infty, g(0+) = \infty$, we write

$$f(x) \approx g(x), \quad x \to 0+,$$

if

$$\lim_{x \to 0+} \frac{\log f(x)}{\log g(x)} = 1.$$

Definition 4.2. A function $\phi : (0, \infty) \to (0, \infty)$ is called a *subpower function (as $x \to 0$)* if for every $a > 0$, there is an $\epsilon > 0$ such that

$$x^a \le \phi(x) \le x^{-a}, \quad 0 < x \le \epsilon.$$

The following properties are straightforward to show and are left as an exercise (Exercise 4.1).

Lemma 4.3.

- *Constants are subpower functions.*
- *If $\beta \in \mathbb{R}$, $[\log(1/x)]^\beta$ is a subpower function.*
- *If ϕ_1, ϕ_2 are subpower functions, so are $\phi_1 + \phi_2$ and $\phi_1\, \phi_2$.*
- *If ϕ_1, ϕ_2 are subpower functions and $a \neq 0$, then*

$$x^a\, \phi_1(x) \approx x^a\, \phi_2(x), \quad x \to 0 + .$$

Definition 4.4. A bounded subset $A \subset \mathbb{R}^d$ has *box dimension* $D = D(A)$ if there exist subpower functions ϕ_1, ϕ_2 such that for all ϵ

$$\phi_1(\epsilon)\, \epsilon^{-D} \le N_\epsilon(A) \le \phi_2(\epsilon)\, \epsilon^{-D}.$$

If $D > 0$, this is equivalent to the relation

$$N_\epsilon(A) \approx \epsilon^{-D}, \quad \epsilon \to 0 + .$$

♢ The box dimension is not defined for all sets A. One can define the *upper box dimension* by

$$\overline{D}(A) = \limsup_{\epsilon \to 0+} \frac{\log N_\epsilon(A)}{\log(1/\epsilon)}.$$

The *lower box dimension* is defined similarly using $\liminf$. The upper and lower box dimensions are defined for every set, and the box dimension exists if the upper and lower box dimensions are equal.

It can be challenging to determine box dimensions because the quantity $N_\epsilon(A)$ can be hard to compute exactly. However, as long as we can estimate it well enough on both sides, we can find the box dimension. The following simple lemma helps.

Lemma 4.5. *Suppose $A \subset \mathbb{R}^d$ is a bounded set and $S = \{x_1, \ldots, x_k\} \subset A$.*

- *If for every $y \in A$, there exists $x_j \in S$ with $|y - x_j| \le \epsilon/2$, then $N_\epsilon(A) \le k$.*
- *If $|x_j - x_i| > 2\epsilon$ for all $j \ne i$, then $N_\epsilon(A) \ge k$.*

Proof. The first assertion follows by considering the covering of A by the balls of diameter ϵ centered at $x_j \in S$. If $|x_j - x_i| > 2\epsilon$, then no ball of diameter ϵ can include both x_i and x_j. Hence any cover must include at least k balls. □

We consider some examples.

- If $A = [0,1]$, then it takes about ϵ^{-1} balls (intervals) of diameter ϵ to cover $[0,1]$. Therefore, $D(A) = 1$.
- If $A = [0,1] \times [0,1]$, we can see that it takes $c\,\epsilon^{-2}$ balls of diameter ϵ to cover A. Therefore, $D(A) = 2$.
- Let A be the Cantor set as defined in Section 3.1.4. Then A_n is the disjoint union of 2^n intervals each of length (diameter) 3^{-n}. Since $A \subset A_n$, we have $N_{3^{-n}}(A) \le 2^n$. Since the left endpoints of these intervals are all retained in A, we can use Lemma 4.5 to see that $N_{3^{n+1}}(A) \ge 2^n$. Hence
$$N_\epsilon(A) \approx \epsilon^{-D} \text{ where } D = \frac{\log 2}{\log 3}.$$
- Let A be the random Cantor set as in Section 3.1.4. Then A_n is the union of Y_n intervals each of length k^{-n}. If $\mu \le 1$, then with probability one $A = \emptyset$. Let us consider the case

$\mu > 1$. With probability $p < 1$, $A = \emptyset$ and with probability $q = 1 - p$, as $n \to \infty$,

$$Y_n \sim M_\infty \mu^n,$$

with $M_\infty > 0$. Since $A_n \supset A$, we see that $N_{k^{-n}}(A) \leq Y_n$. It requires a little more argument, but one can show that on the event $A \neq \emptyset$

$$D(A) = \frac{\log \mu}{\log k}.$$

We list some properties of box dimension.

- If $A_1, \ldots, A_r$ have box dimensions $D(A_1)$, $\ldots$, $D(A_r)$, respectively, then the box dimension of $A_1 \cup \cdots \cup A_r$ is

$$\max\{D(A_1), \ldots, D(A_r)\}.$$

Indeed, this follows immediately from

$$\max_j N_\epsilon(A_j) \leq N_\epsilon(A_1 \cup \cdots \cup A_r) \leq N_\epsilon(A_1) + \cdots + N_\epsilon(A_r).$$

- If $A \subset \mathbb{R}^d$ with box dimension $D(A)$, then the box dimension of the closure of A is also $D(A)$. Indeed, any cover of A with a finite number of closed sets is also a cover of the closure of A (why?). For example, the box dimension of the set of rational numbers between 0 and 1 is the same as the dimension of $[0, 1]$ which is 1. Note that this shows that the countable union of sets of box dimension zero can have nonzero box dimension.

4.2. Cantor measure

In this section we compare two measures on $[0, 1]$: length (or more fancily called Lebesgue measure) which is a one-dimensional measure and *Cantor measure* which is related to the Cantor set and is an α-dimensional measure where $\alpha = \log 2 / \log 3$. In both cases, we can think of the measures probabilistically in terms of an infinite number of independent trials.

Suppose we pick a number from $[0, 1]$ at random from the uniform distribution. Then for any set $A \subset [0, 1]$, we would like to say that the probability that our number is in A is given by the length of

A, $l(A)$. This gets a little tricky when A is a very unusual subset, but certainly if A is an interval or a finite union of intervals, this interpretation seems reasonable.

Any $x \in [0,1]$ can be written in its dyadic expansion

$$x = .a_1a_2a_3\cdots = \sum_{j=1}^{\infty} \frac{a_j}{2^j}, \qquad a_j \in \{0,1\}.$$

This expansion is unique unless x is a dyadic rational (e.g., $.0111\cdots = .1000\cdots = 1/2$), but the chance that we choose such a dyadic rational is zero so we will just assume that the expansion is unique. We can think of $a_1, a_2, \ldots$ as independent random variables taking values 0 and 1 each with probability $1/2$. Another way to think of the measure is in terms of mass distribution. We have a total mass of 1 to distribute. We start by splitting $[0,1]$ into two intervals $[0, \frac{1}{2}]$ and $[\frac{1}{2}, 1]$ and we divide the mass equally between these two sets, giving each mass $1/2$. Then each of these two intervals is divided in half and the mass is split evenly between the intervals. This gives us four intervals, each with mass $1/4$. The total mass for the interval $[0,x]$ becomes x.

The Cantor measure is defined similarly except that we write real numbers in their *ternary* expansion and consider only those real numbers for which 1 does not appear:

$$x = .b_1b_2\cdots = \sum_{j=1}^{\infty} \frac{b_j}{3^j}, \quad b_j \in \{0,2\}.$$

The random variables $b_1, b_2, \ldots$ are independent taking the values 0 and 2 each with probability $1/2$. In terms of mass distribution, we have a total mass of 1 for the interval $[0,1]$. When the interval is divided into three pieces $[0, \frac{1}{3}], [\frac{1}{3}, \frac{2}{3}], [\frac{2}{3}, 1]$, the first and third intervals each get mass $1/2$ and the middle interval gets mass 0. Each of these intervals is in turn split into three equal intervals with the first and third intervals receiving half of the mass each and the middle interval receiving none. At the nth stage there are 3^n intervals of length 3^{-n}. The mass is distributed equally among 2^n of the intervals (each having mass 2^{-n}) and the other intervals have zero mass. As can be seen, the 2^n intervals with positive mass are the intervals appearing in the

nth approximation of the Cantor set. In the limit, we get a measure μ on $[0,1]$ called the *Cantor measure*.

The Cantor measure is closely related to the *Cantor function*

$$F(x) = \mu[0,x]. \tag{4.1}$$

The function F is a continuous, nondecreasing function which equals $1/2$ for $1/3 \le x \le 2/3$; $1/4$ for $1/9 \le x \le 2/9$; $3/4$ for $7/9 \le x \le 8/9$, etc.

The Cantor measure gives zero measure to every point. However, it gives measure one to the Cantor set; this can be seen by noting the the measure of A_n is one for each n. Note that the length of the Cantor set is 0 since the length of A_n is $2^n\, 3^{-n} \to 0$. This measure is in some sense an α-dimensional measure where $\alpha = \log 2/\log 3$. There are several ways to measure the "dimension" of a measure. One way is to say that a measure is D-dimensional if for small ϵ, balls of diameter ϵ tend to have measure ϵ^D (we are being somewhat vague here). This characterization is consistent with length being a one-dimensional measure, area a two-dimensional measure, volume a three-dimensional measure, etc. In the Cantor measure, if I is one of the intervals of length 3^{-n} in A_n, then the measure of $A \cap I$ is $2^{-n} = \operatorname{diam}(I)^\alpha$.

Proposition 4.6. *Suppose I is a closed subinterval of $[0,1]$. Then*

$$m(I) \le 3^\alpha\, l(I)^\alpha, \quad \alpha = \frac{\log 2}{\log 3}. \tag{4.2}$$

Proof. Suppose $l(I) = 3^{-k}$ for some positive integer k. Since the kth approximation of the Cantor set consists of 2^k intervals of length 3^{-k} and none of the intervals are adjacent, we can see that the interior of I can intersect at most one of these intervals. Therefore $m(I) \le 2^{-k} = l(I)^\alpha$. Similarly, if $3^{-k-1} < l(I) \le 3^{-k}$, we can conclude

$$m(I) \le 2^{-k} = (3^{-k})^{-\alpha} \le 3^\alpha\, l(I)^\alpha.$$

□

In Exercise 4.4 you are asked to improve this to show that for all intervals

$$m(I) \le l(I)^\alpha.$$

4.2.1. Some notes on measures and integrals. We used the term measure somewhat loosely in this section. Here we introduce some precise definitions. A measure is a function m from a collection of subsets of $\mathbb{R}^d$ into $[0, \infty)$. For technical reasons, it is often impossible to define a measure on *all* subsets, but one can define it on all reasonable sets. The next definition describes the sets we will consider.

Definition 4.7. The *Borel subsets* of $\mathbb{R}^d$, denoted by $\mathcal{B}$, is the smallest collection of subsets of $\mathbb{R}^d$ with the following properties:

- All open sets are in $\mathcal{B}$;
- If $V \in \mathcal{B}$, then $\mathbb{R}^d \setminus V \in \mathcal{B}$;
- If $V_1, V_2, \ldots$ is a finite or countable collection of sets in $\mathcal{B}$, then
$$\bigcup_{j=1}^{\infty} V_j \in \mathcal{B}.$$

It may not be obvious to the reader that this definition makes sense; see Exercise 4.5 for more details. All closed sets are Borel sets since closed sets are of the form $\mathbb{R}^d \setminus V$ for an open set V.

Definition 4.8. A *(Borel) measure* is a function $m : \mathcal{B} \to [0, \infty]$ such that $m(\emptyset) = 0$ and if $V_1, V_2, \ldots \in \mathcal{B}$ are disjoint,
$$m\left(\bigcup_{j=1}^{\infty} V_j\right) = \sum_{j=1}^{\infty} m(V_j).$$

We will also need the *Lebesgue integral* with respect to a measure. We will only need it for positive functions.

Definition 4.9. If m is a Borel measure and f is a positive function on $\mathbb{R}$, then
$$\int f(x)\, m(dx)$$
is defined as follows.

- If f is a *simple function*, that is, a function of the form
$$f(x) = \sum_{j=1}^{n} a_j \, 1\{x \in V_j\}, \tag{4.3}$$

for Borel sets V_j, then

$$\int f(x)\, m(dx) = \sum_{j=1}^{\infty} a_j\, m(V_j).$$

- If $f_1 \le f_2 \le \dots$ are simple functions and

$$f = \lim_{n\to\infty} f_n,$$

then

$$\int f(x)\, m(dx) = \lim_{n\to\infty} \int f_n(x)\, m(x). \tag{4.4}$$

The careful reader will note that we have only defined the integral for functions that can be written in the form (4.3). This does not include all positive functions, and, in fact, we cannot define the integral for all functions. This will suffice for our purposes. Also, we should define the integral to be the supremum of the right hand side of (4.4) where the supremum is over all such sequence $\{f_n\}$. In fact, one can show the value is the same for all such sequences (this is the monotone convergence theorem).

4.3. Hausdorff measure and dimension

Hausdorff dimension is another definition of fractal dimension that agrees with the box dimensions on many sets such as the Cantor set but is not mathematically equivalent. It has some nice mathematical properties, e.g., it is well defined for every set, that makes it a more useful tool for many applications. However, the definition is more complicated which makes it more difficult to compute the dimension of sets.

4.3.1. Hausdorff measure. Suppose $A \subset \mathbb{R}^d$ is a bounded set. For each $\epsilon > 0$, we define

$$\mathcal{H}^{\alpha}_{\epsilon}(A) = \inf \sum_{j=1}^{\infty} \operatorname{diam}(A_j)^{\alpha},$$

where the infimum is over all finite or countable collections of sets $A_1, A_2, \ldots$ with

$$A \subset \bigcup_{j=1}^{\infty} A_j$$

and $\operatorname{diam}(A_j) \le \epsilon$. Here diam denotes the diameter of a set,

$$\operatorname{diam}(V) = \sup\{|x-y| : x, y \in V\}.$$

If $\delta < \epsilon$, then

$$\mathcal{H}^\alpha_\delta(A) \ge \mathcal{H}^\alpha_\epsilon(A)$$

since any covering of A by sets of diameter at most δ is also a covering of A by sets of diameter at most ϵ. The *Hausdorff α-measure* of A is defined by

$$\mathcal{H}^\alpha(A) = \lim_{\epsilon \to 0+} \mathcal{H}^\alpha_\epsilon(A).$$

The existence of the limit follows from the fact that increasing limits exist. However, it is not clear that this definition is interesting; for example, it could give the value 0 or ∞. In fact, our next proposition will show that for any set A, there is at most one value of α such that $\mathcal{H}^\alpha(A) \notin \{0, \infty\}$.

Proposition 4.10. *Suppose $A \subset \mathbb{R}^d$ is bounded and $0 \le \alpha < \beta < \infty$.*

- *If $\mathcal{H}^\alpha(A) < \infty$, then $\mathcal{H}^\beta(A) = 0$.*
- *If $\mathcal{H}^\beta(A) > 0$, then $\mathcal{H}^\alpha(A) = \infty$.*

Proof. For any $\epsilon > 0$ and any cover $A_1, A_2, \ldots$ of A of sets with $\operatorname{diam}(A_j) \le \epsilon$,

$$\sum \operatorname{diam}(A_j)^\beta \le \epsilon^{\beta-\alpha} \sum \operatorname{diam}(A_j)^\alpha.$$

Therefore, by taking infimums, we get

$$\mathcal{H}^\beta_\epsilon(A) \le \epsilon^{\beta-\alpha}\, \mathcal{H}^\alpha_\epsilon(A).$$

Since $\epsilon^{\beta-\alpha} \to 0$ as $\epsilon \to 0+$, we get the result. □

We will not prove the next proposition which justifies the term Hausdorff *measure*. We define $\mathcal{H}^\alpha(A)$ for unbounded A by

$$\mathcal{H}^\alpha(A) = \lim_{R \to \infty} \mathcal{H}^\alpha(A \cap \{x \in \mathbb{R}^d; |x| \le R\}).$$

Proposition 4.11. *For each $\alpha > 0$, $\mathcal{H}^\alpha$ is a measure on the Borel sets. In other words, $\mathcal{H}^\alpha : \mathcal{B} \to [0, \infty]$ with $\mathcal{H}^\alpha(\emptyset) = 0$ and if $V_1, V_2, \ldots$ are disjoint*

$$\mathcal{H}^\alpha \left[\bigcup_{j=1}^{\infty} V_j \right] = \sum_{j=1}^{\infty} \mathcal{H}^\alpha(V_j).$$

Proposition 4.12. *Suppose A is a bounded subset of $\mathbb{R}^d$ and $\alpha \geq 0$.*

- *If $x \in \mathbb{R}^d$, and $A + x = \{y + x : y \in A\}$, then $\mathcal{H}^\alpha(A + x) = \mathcal{H}^\alpha(A)$.*
- *If $r > 0$, and $rA = \{ry : y \in A\}$, then $\mathcal{H}^\alpha(rA) = r^\alpha \, \mathcal{H}^\alpha(A)$.*

Proof. If $A_1, A_2, \ldots$ is a cover of A, then $A_1 + x, A_2 + x, \ldots$ is a cover of $A + x$ and $rA_1, rA_2, \ldots$ is a cover of rA. Using this one can see that

$$\mathcal{H}^\alpha_\epsilon(A + x) = \mathcal{H}^\alpha_\epsilon(A), \qquad \mathcal{H}^\alpha_{r\epsilon}(rA) = r^\alpha \, \mathcal{H}^\alpha_\epsilon(A).$$

□

Example 4.13. Let A be the Cantor set. We will show that $3^{-\alpha} \leq \mathcal{H}^\alpha(A) \leq 1$ where $\alpha = \log 2 / \log 3$. (In Exercise 4.4, it is shown that $\mathcal{H}^\alpha(A) = 1$.) Let $\epsilon > 0$ and choose n sufficiently large so that $3^{-n} < \epsilon$. We can cover A with 2^n intervals of length (diameter) 3^{-n}. Therefore

$$\mathcal{H}^\alpha_\epsilon(A) \leq 2^n \, (3^{-n})^\alpha = 1.$$

Letting $\epsilon \to 0+$, we get $\mathcal{H}^\alpha(A) \leq 1$. To prove the lower bound we use the Cantor measure m and the estimate (4.2). If $A_1, A_2, \ldots$ is a cover of A, then

$$1 = m(A) \leq \sum_{j=1}^{\infty} m(A_j) \leq \sum_{j=1}^{\infty} 3^\alpha \operatorname{diam}(A_j)^\alpha.$$

By taking the infimum, we see that $\mathcal{H}^\alpha(A) \geq 3^{-\alpha}$.

4.3.2. Hausdorff dimension. If $A \subset \mathbb{R}^d$, then the *Hausdorff dimension* of A which we denote by $\dim_h(A)$ is defined by

$$\dim_h(A) = \inf\{\alpha : \mathcal{H}^\alpha(A) = 0\}.$$

Using Proposition 4.10, we can also describe $\dim_h(A)$ by the relation

$$\mathcal{H}^\alpha(A) = \begin{cases} \infty, & \alpha < \dim_h(A) \\ 0, & \alpha > \dim_h(A) \end{cases}.$$

If $\alpha = \dim_h(A)$, then $\mathcal{H}^\alpha(A)$ can be $0, \infty$, or a finite positive number. If $0 < \mathcal{H}^\beta(A) < \infty$ for some β then $\beta = \dim_h(A)$.

Proposition 4.14. *If A is a bounded set with box dimension D, then*

$$\dim_h(A) \leq D.$$

Proof. To show that $\dim_h(A) \leq D$ it suffices to show for every $\alpha > D$, $\mathcal{H}^\alpha(D) = 0$. Let $D < \beta < \alpha$. Since the box dimension of A is D, for all ϵ sufficiently small, we can cover D by $\epsilon^{-\beta}$ sets of diameter ϵ. Using this cover, we see that

$$\mathcal{H}^\alpha_\epsilon(D) \leq \epsilon^{-\beta}\, \epsilon^\alpha = \epsilon^{\alpha-\beta}.$$

Therefore,

$$\mathcal{H}^\alpha(D) = \lim_{\epsilon \to 0+} \mathcal{H}^\alpha_\epsilon(D) \leq \lim_{\epsilon \to 0+} \epsilon^{\alpha-\beta} = 0.$$

□

The next proposition gives a property of Hausdorff dimension that is not satisfied by box dimension.

Proposition 4.15. *Suppose $A_1, A_2, \ldots$ are subsets of $\mathbb{R}^d$. Then*

$$\dim_h \left[\bigcup_{j=1}^{\infty} A_j \right] = \sup \left\{ \dim_h(A_j) : j = 1, 2, \ldots \right\}.$$

Proof. By monotonicity, for each j,

$$\dim_h \left[\bigcup_{j=1}^{\infty} A_j \right] \geq \dim_h(A_j),$$

and hence

$$\dim_h \left[\bigcup_{j=1}^{\infty} A_j \right] \geq \sup \left\{ \dim_h(A_j) : j = 1, 2, \ldots \right\}.$$

Conversely, suppose $\alpha > \dim_h(A_j)$ for every j. Then $\mathcal{H}^\alpha(A_j) = 0$ and by properties of measures

$$\mathcal{H}^\alpha \left[\bigcup_{j=1}^{\infty} A_j \right] = 0.$$

This implies that $\alpha \geq \dim_h[\cup A_j]$.

□

It follows from the proposition that every countable set has Hausdorff dimension zero. In particular, the set of rationals between 0 and 1 has dimension zero. Recall that the box dimension of a set is the same as its closure so this set has box dimension one. This example shows that the Hausdorff dimension of a set is not necessarily the same as its closure.

4.3.3. Computing Hausdorff dimensions. Computing Hausdorff measure and Hausdorff dimension can be difficult. Let us start with an example.

Proposition 4.16. *If $A = [0, 1]$, then $\mathcal{H}^1(A) = 1$. In particular, $\dim_h([0, 1]) = 1$.*

Proof. We will prove the stronger fact that $\mathcal{H}^1_\epsilon(A) = 1$ for each $\epsilon > 0$. Fix $\epsilon > 0$. To give an upper bound, choose $n > 1/\epsilon$ and consider the cover of A by the n intervals $A_{j,n} = [\frac{j-1}{n}, \frac{j}{n}]$ for $j = 1, \ldots, n$. Then

$$\mathcal{H}^1_\epsilon(A) \leq \sum_{j=1}^{n} \text{diam}(A_{j,n}) = 1.$$

To prove the lower bound, suppose $A_1, A_2, \ldots$ is a collection of sets such that $A \subset \bigcup_{j=1}^{\infty} A_j$. Without loss of generality assume that A_j are actually intervals. Then $\text{diam} A_j = l(A_j)$ where l denotes length.

Then

$$\begin{aligned} 1 = l(A) &\leq l\left(\bigcup_{j=1}^{\infty} A_j\right) \\ &\leq \sum_{j=1}^{\infty} l(A_j) \\ &= \sum_{j=1}^{\infty} \operatorname{diam}(A_j). \end{aligned} \tag{4.5}$$

□

This proof is typical of calculations of Hausdorff measure and Hausdorff dimension. In order to get an upper bound, one needs only find a good cover. However, to get lower bounds one needs bounds that hold for all covers. In the case $A = [0, 1]$, we used a measure on $[0, 1]$ (length) to estimate the sum in (4.5) for any cover. This is the most common way to give lower bounds. The idea is to show a set is big by showing there is a way to distribute mass on the set so that it is sufficiently spread out. Somewhat more precisely, the idea is that if one can put an "α-dimensional" measure on a set, then the set must be at least α-dimensional.

We say that a measure m is *carried* on A if $m(\mathbb{R}^d \setminus A) = 0$. For example, the Cantor measure is carried on the Cantor set A. We will call a measure m that is carried on A and satisfies $0 < m(A) < \infty$ a *mass distribution* on A.

Proposition 4.17. *Suppose A is a bounded subset of $\mathbb{R}^d$. Suppose m is a mass distribution on A and that there exist $\epsilon_0 > 0, c < \infty$ such that for all B with* $\operatorname{diam}(B) < \epsilon_0$,

$$m[B] \leq c \operatorname{diam}(B)^{\alpha}.$$

Then $\mathcal{H}^{\alpha}(A) \geq c^{-1} m(A) > 0$. In particular,

$$\dim_h(A) \geq \alpha.$$

Proof. Let $A_1, A_2, \ldots$ be a collection of sets with $A \subset \bigcup A_j$ and $\operatorname{diam}(A_j) < \epsilon_0$. Then

$$m(A) \le \sum_{j=1}^{\infty} m(A_j) \le c \sum_{j=1}^{\infty} \operatorname{diam}(A_j)^{\alpha}.$$

By taking the infimum, we can see that for each $\epsilon < \epsilon_0$,

$$\mathcal{H}^{\alpha}_{\epsilon}(A) \ge c^{-1}\, m(A).$$

□

The last proposition is not difficult to prove, but unfortunately the conditions are too strong when dealing with random fractals such as the random Cantor set. We will give a different way of saying that a measure is spread out. To motivate, let us consider an integral in $\mathbb{R}^d$,

$$\int_{|x|\le 1} \frac{d^d x}{|x|^{\alpha}}.$$

Using spherical coordinates, we can see that the integral is finite if $\alpha < d$ and is infinite if $\alpha \ge d$ (check this!). Using this as motivation, we can see that if m is a mass distribution with

$$\int_{|x|\le 1} \frac{m(dx)}{|x|^{\alpha}} < \infty,$$

then m looks "at least α-dimensional" near zero. We need to focus locally at all points, so we will consider instead a somewhat more complicated integral:

$$\mathcal{E}_{\alpha}(m) = \int \int \frac{m(dx)\, m(dy)}{|x-y|^{\alpha}}.$$

We will not prove the following generalization of Proposition (4.17).

Theorem 4.18. *Suppose A is a compact subset of $\mathbb{R}^d$ and m is a mass distribution on A with $\mathcal{E}^{\alpha}(m) < \infty$. Then* $\dim_h(A) \ge \alpha$.

4.3.4. Random Cantor set. We will consider the random Cantor set A as in Section 3.1.4 with $\mu = kp$. We have already noted that the probability that A is nonempty is strictly positive if and only if $\mu > 1$. We will discuss the proof of the following theorem, leaving some parts as exercises.

Theorem 4.19. *With probability one, the random Cantor set A is either empty or has Hausdorff dimension* $\log \mu / \log k$.

For ease we will only consider $k = 2$, but the argument is essentially the same for all k. We choose $p \in (\frac{1}{2}, 1)$ so that $\mu = kp > 1$ and we let $\alpha = \log \mu / \log 2$ so that $2^\alpha = \mu$.

Let $\mathcal{I}_n$ denote the set of dyadic intervals in $[0,1]$ of length 2^{-n}. Recall that $A = \bigcap A_n$ where A_n is the union of Y_n intervals in $\mathcal{I}_n$. Recall from Sections 3.5.1 and 3.6.1 that with probability one the limit

$$M_\infty = \lim_{n\to\infty} \mu^{-n} Y_n \tag{4.6}$$

exists and is positive provided that $A \neq \emptyset$. In particular, for n sufficiently large,

$$Y_n \leq 2\, M_\infty\, \mu^n.$$

Since the random Cantor set can be covered by Y_n intervals of diameter 2^{-n}, we can see that

$$\mathcal{H}^\alpha(A) \leq Y_n\, 2^{-n\alpha} \leq 2\, M_\infty,\ \mu^n\, 2^{-n\alpha} = 2\, M_\infty < \infty.$$

Therefore, with probability one, $\dim_h(A) \leq \alpha$. The lower bound is more difficult.

It is not difficult to extend (4.6) as follows. If $I \in \mathcal{I}_j$ and $n \geq j$, let $Y_n(I)$ denote the number of intervals $I' \in \mathcal{I}_n$ such that $I' \subset I$. Then with probability one, for each I,

$$M_\infty(I) := \lim_{n\to\infty} \mu^{-n} Y_n(I),$$

exists and is strictly positive if $I \cap A \neq \emptyset$. Recalling that $\mathbb{E}[M_\infty] = 1$, we get

$$\begin{aligned} 1 = \mathbb{E}[M_\infty] &= \mathbb{E}\left[\sum_{I\in\mathcal{I}_n} M_\infty(I)\right] \\ &= \sum_{I\in\mathcal{I}_n} \mathbb{E}\left[M_\infty(I)\right] \\ &= \sum_{I\in\mathcal{I}_n} \mathbb{P}\{I \in A_n\}\mathbb{E}[M_\infty(I) \mid I \in A_n] \\ &= 2^n\, p^n\, \mathbb{E}[M_\infty(I) \mid I \in A_n]. \end{aligned}$$

The last equality uses the fact that $\mathbb{E}[M_\infty(I) \mid I \in A_n]$ must be the same for all $I \in \mathcal{I}_n$. Therefore,

$$\mathbb{E}\left[M_\infty(I) \mid I \in A_n\right] = \mu^{-n}, \quad I \in \mathcal{I}_n.$$

Viewed in this way, we can think of M_∞ as a *random measure* that assigns to dyadic intervals I the measure $M_\infty(I)$. We can find $M_\infty(I)$ for other intervals I by writing I as a countable union of dyadic intervals. Let us write this measure as m. Then $\mathcal{E}_\beta(m)$ is a random variable. The next proposition is the key estimate.

Proposition 4.20. *If* $0 \le \beta < \alpha$,

$$\mathbb{E}\left[\mathcal{E}_\beta(m)\right] < \infty.$$

In particular, with probability one, $\mathcal{E}_\beta(m) < \infty$.

Before proving this, let us say why this implies the lower bound. On the event that $A \neq \emptyset$, we know that $m([0,1]) > 0$. Since for each $\beta < \alpha$, $\mathbb{E}\left[\mathcal{E}_\beta(m)\right] < \infty$, we know that with probability one for all $\beta < \alpha$, $\mathcal{E}_\beta(m) < \infty$.

♢ This is another case where we show that a positive random variable Z is finite with probability one by showing that $\mathbb{E}[Z] < \infty$. The technique does not work for the converse. We cannot conclude that $Z = \infty$ with positive probability by showing that $\mathbb{E}[Z] = \infty$.

Proof. We will consider two distances or metrics ρ, d on the set $\mathcal{I}_n$. Let ρ be defined by $\rho(I, I') = j$ if j is the smallest integer such that there is an interval $I'' \in \mathcal{I}_{n-j}$ such that $I \subset I''$ and $I' \subset I''$. We will use d for the Euclidean distance between intervals

$$d(I, I') = \min\{|x - y| : x \in I, y \in I'\}.$$

(This is not quite a metric because adjacent intervals are distance zero apart, but this will not matter.) Note that if $I, I' \in \mathcal{I}$ with $\rho(I, I') \le j$, then $d(I, I') \le 2^{j-n}$. (There is no converse relation. It is possible for intervals to be close in the d metric but far away in the ρ metric). Also, if $I \in \mathcal{I}_n$, then the number of intervals $I' \in \mathcal{I}_n$ with $d(I, I') \le 2^j$ is less than 2^{j+2}.

If $I, I' \in \mathcal{I}_n$, then

$$\int_I \int_{I'} \frac{m(dx)\, m(dy)}{|x-y|^\beta} \le \int_I \int_{I'} \frac{m(dx)\, m(dy)}{d(I,I')^\beta} = \frac{M_\infty(I)\, M_\infty(I')}{d(I,I')^\beta}.$$

Let

$$\mathcal{E}_{\beta,\epsilon}(m) = \int \int_{|x-y|\ge\epsilon} \frac{m(dx)\, m(dy)}{|x-y|^\beta}.$$

We claim that it suffices to show that there is a $C = C(\beta)$ such that for all ϵ,

$$\mathbb{E}[\mathcal{E}_{\beta,\epsilon}(m)] \le C.$$

Indeed,

$$\mathcal{E}_\beta(m) = \lim_{\epsilon\to 0+} \mathcal{E}_{\beta,\epsilon}(m)$$

and this is a monotone limit, so we may use the monotone convergence theorem to conclude

$$\mathbb{E}[\mathcal{E}_\beta(m)] \le C.$$

For each ϵ, if n is sufficiently large

$$\begin{aligned} (4.7) \qquad & \mathbb{E}\,[\mathcal{E}_{\beta,\epsilon}(m)] \\ &\le \sum_{I,I'\in\mathcal{I}_n, d(I,I')>2^{-n}} \frac{\mathbb{E}[M_\infty(I)\, M_\infty(I')]}{d(I,I')^\beta} \\ &\le \sum_{j=0}^{n} \sum_{2^{-n+j}<d(I,I')\le 2^{-n+j+1}} \frac{\mathbb{E}[M_\infty(I)\, M_\infty(I')]}{d(I,I')^\beta} \\ &\le \sum_{j=0}^{n} 2^{(n-j)\beta} \sum_{2^{-n+j}<d(I,I')\le 2^{-n+j+1}} \mathbb{E}[M_\infty(I)\, M_\infty(I')]. \end{aligned}$$

Suppose $I, I' \in \mathcal{I}_n$ with $\rho(I,I') = j$. Then there is a common "ancestor" $I'' \in \mathcal{I}_{n-j}$. Let $\mathcal{F} = \mathcal{F}_{n-j}$ denote the information up through the $(n-j)$th level. Then,

$$\begin{aligned} &E[M_\infty(I)\, M_\infty(I') \mid \mathcal{F}] \\ &\quad = 1\{I'' \in A_{n-j}\}\, \mathbb{E}\,[M_\infty(I) \mid I'' \in A_{n-j}]\, \mathbb{E}\,[M_\infty(I') \mid I'' \in A_{n-j}] \\ &\quad = 2^{-2j}\, \mu^{2(j-n)}\, 1\{I'' \in A_{n-j}\}. \end{aligned}$$

Here we use

$$\mathbb{E}\,[M_\infty(I) \mid I'' \in A_{n-k}] = 2^{-k}\, \mathbb{E}\,[M_\infty(I'') \mid I'' \in A_{n-k}] = 2^{-k}\, \mu^{k-n}.$$

Hence,

$$\begin{aligned}\mathbb{E}\,[M_\infty(I)\,M_\infty(I')] &= \mathbb{E}\,[E[M_\infty(I)\,M_\infty(I') \mid \mathcal{F}]] \\ &= 2^{-2j}\,\mu^{2(j-n)}\,\mathbb{P}\{I'' \in A_{n-j}\} \\ &= p^{n-j}\,2^{-2j}\,\mu^{2(j-n)} = p^{n+j}\,\mu^{-2n}.\end{aligned}$$

If $d(I,I') > 2^{j-n}$, then $\rho(I,I') > j$, and hence

$$\mathbb{E}\,[M_\infty(I)\,M_\infty(I')] \le p^{n+j}\,\mu^{-2n}, \quad d(I,I') > 2^{j-n}.$$

Let us return to (4.7). For a given I, there are at most $c2^j$ intervals I' with $d(I,I') \le 2^{j-n+1}$. Hence the number of ordered pairs (I,I') satisfying this is bounded by $c2^{n+j}$. Therefore,

$$\sum_{2^{-n+j} < d(I,I') \le 2^{-n+j+1}} \mathbb{E}[M_\infty(I)\,M_\infty(I')] \le c\,2^{n+j}\,p^{n+j}\,\mu^{-2n} = c\,\mu^{j-n},$$

and

$$\mathbb{E}\,[\mathcal{E}_{\beta,\epsilon}(m)] \le c\sum_{j=0}^{n} 2^{n-j}\beta\,\mu^{j-n} \le c\sum_{j=0}^{\infty} 2^{j\beta}\,\mu^{-j} := C < \infty.$$

The last inequality uses $2^\beta < \mu$ which follows from $2^\alpha = \mu$.

□

Exercises

Exercise 4.1. Prove Lemma 4.3. Give an example to show that the final statement is false for $a = 0$.

Exercise 4.2. We will construct a set $A \subset [0,1]$ for which the box dimension does not exist. It will be a generalization of the Cantor set. Suppose k_n, j_n are sequences of positive integers greater than 1. Here is the construction.

- $A_0 = [0,1]$
- A_1 is the finite union of j_1 intervals of length $l_1 = (j_1k_1)^{-1}$. It is obtained by splitting $[0,1]$ into j_1k_1 equal intervals and selecting j_1 intervals by taking every k_1th one. For example,

if $j_1 = 3, k_1 = 2$, then $l_1 = 1/6$ and the interval $[0,1]$ is written as

$$[0,1] = \left[0, \frac{1}{6}\right] \cup \left[\frac{1}{6}, \frac{2}{6}\right] \cup \cdots \cup \left[\frac{5}{6}, 1\right],$$

and

$$A_1 = \left[\frac{1}{6}, \frac{2}{6}\right] \cup \left[\frac{3}{6}, \frac{4}{6}\right] \cup \left[\frac{5}{6}, 1\right].$$

- Inductively, if A_n is given which is a disjoint union of Y_n intervals of length l_n, then A_{n+1} is obtained by dividing each of the intervals into $j_{n+1}k_{n+1}$ equal pieces and taking every k_{n+1}th subinterval. Then A_{n+1} is the disjoint union of Y_{n+1} intervals of length l_{n+1} where $Y_{n+1} = j_{n+1}\, Y_n, l_{n+1} = l_n\, (j_{n+1}k_{n+1})^{-1}$.

(1) Show that if $j_n = j, k_n = k$, then $D(A)$ is well-defined and find $D(A)$.

(2) Find an example of j_n, k_n such that $D(A)$ is not defined.

Exercise 4.3. Let $F : [0,1] \to [0,1]$ be the Cantor function as defined in (4.1).

- Show that F is continuous.
- Show that $F'(x) = 0$ for x in the complement of the Cantor set.
- True or false: if $F : [0,1] \to \mathbb{R}$ is continuous and F' exists except perhaps on a set of measure zero, them

$$F(x) = F(0) + \int_0^x F'(s)\, ds.$$

Exercise 4.4. Let m denote Cantor measure on $[0,1]$. Show that if I is a closed subinterval of $[0,1]$, then

$$m(I) \leq l(I)^\alpha, \quad \alpha = \frac{\log 2}{\log 3}.$$

Use this to show that the Hausdorff α-measure of the Cantor set equals one.

Exercise 4.5. Call a collection of subsets $\mathcal{A}$ of $\mathbb{R}^d$ a sigma-algebra of subsets if

- $\emptyset \in \mathcal{A}$.
- If $V \in \mathcal{A}$, then $\mathbb{R}^d \setminus V \in \mathcal{A}$;
- If $V_1, V_2, \ldots \in \mathcal{A}$, then

$$\bigcup_{n=1}^{\infty} V_n \in \mathcal{A}.$$

(1) Find a sigma-algebra $\mathcal{A}$ that contains all the open sets.

(2) Let

$$\mathcal{B} = \bigcap \mathcal{A}$$

where the intersection is over all sigma-algebras $\mathcal{A}$ that contain all the open sets. Show that $\mathcal{B}$ is a sigma-algebra containing the open sets. (This can be considered as the formal definition of the Borel sets.)

Exercise 4.6.

(1) Show that $\mathcal{H}^0(A)$ equals the number of elements of A.

(2) Show that if A is a countable, then $\mathcal{H}^\alpha_\epsilon(A) = 0$ for all $\alpha, \epsilon > 0$.

Exercise 4.7. Let A be a random Cantor set as in Section 3.1.4. Show that for each $x \in [0,1]$,

$$\mathbb{P}\{x \in A\} = 0.$$

Conclude that if $V \subset [0,1]$ is a countable set,

$$\mathbb{P}\{A \cap V = \emptyset\} = 1.$$

Exercise 4.8. In the scientific literature, fractal dimensions of irregular sets in the plane are sometimes estimated by dividing the plane into squares of side length ϵ for small and counting the number of these squares that intersect the set. Is this technique more like the box dimension or the Hausdorff dimension? Can you make any precise mathematical statements relating this procedure to either box or Hausdorff dimension?