MATH 235C – Probability Theory
Lecture Notes, Winter 2022

Dan Romik
Department of Mathematics
UC Davis

DRAFT (version of May 12, 2022)

Contents

Chapter 1: Dynamical systems

Ergodic theory is a mathematical theory that evolved out of the study of global properties of dynamical systems. Here, we speak loosely of a **dynamical system** as consisting of a **phase space** Ω (a set, whose points are the possible states of the system) together with some **dynamics**, which are a notion of how the state of the system evolves over time. Time may flow continuously or in discrete steps. In the simplest case of discrete-time dynamics, the dynamics are encapsulated by a mapping $T : \Omega \to \Omega$. We imagine that if at a given time the state of the system is some point $\omega \in \Omega$, then in the next time step it will be $T(\omega)$. (We assume that the dynamics, i.e., the rules of evolution of the system over time, are themselves unchanging over time.)

The description of the dynamics in a continuous-time dynamical system is more subtle; it consists of a family $(T_s)_{s\geq 0}$ of maps, where for each $s \geq 0$, $T_s : \Omega \to \Omega$ takes the current state of the system $\omega \in \Omega$ and returns a new point $\omega' = T_s(\omega)$ which represents the state of the system s time units into the future. The maps therefore have to satisfy the conditions

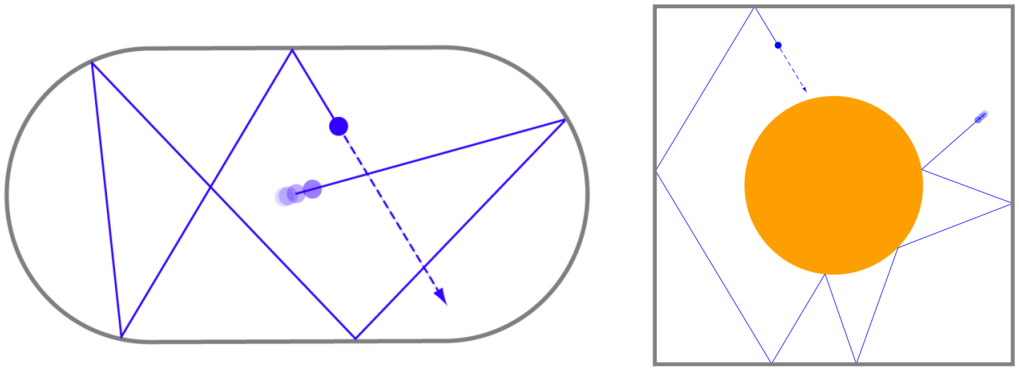
$$\begin{aligned} T_0 &= \mathrm{id}, \\ T_{s+t} &= T_s \circ T_t, \qquad (s, t \geq 0), \end{aligned}$$

i.e., the family $(T_s)_{s\geq 0}$ is a transformation semigroup. In a context where the phase space has a differentiable structure and the dynamics are a result of solving a differential equation, the semigroup $(T_s)_{s\geq 0}$ is often called a **flow**.

Dynamical systems arise naturally in physics, probability, biology, computer science (algorithmic computations can often be interpreted as discrete-time dynamical systems) and many other areas. To illustrate the types of questions that ergodic theory deals with, consider the example of **(mathematical) billiards**: this is a mathematical idealization of the game of billiards in which a small ball is bouncing around without loss of energy in some bounded and odd-shaped region of the plane, being reflected off the walls; see Figure 1 for two examples. The main question of ergodic theory can be roughly formulated as follows:

> *If an observer watches the system for a long time, starting from some arbitrary (random) initial state, can the ideal statistics of the system be recovered?*

The question is formulated in a deliberately vague way, but the idea behind "ideal statistics" is that they are represented by some probability measure $\mathbf{P}$ on the phase space Ω



(a) (b)

Figure 1: Billiard dynamical systems: (a) The "Bunimovich stadium"; (b) The "Sinai billiard" (source: Wikipedia)

(equipped with a suitable measurable structure $\mathcal{F}$) that is compatible with both the way the "arbitrary" initial state of the system is chosen, and with the action of the dynamics of the system (we shall make these ideas more precise soon). The way the observer will try to recover the measure $\mathbf{P}$ is as follows: starting from the initial state x_0 one gets a sequence of subsequent states

$$x_0, \quad x_1 = T(x_0), \quad x_2 = T(T(x_0)), \quad x_3 = T^3(x_0), \ldots$$

in the case of a discrete-time system, or a one-parameter family of states

$$x_s = T_s(x_0), \qquad (s \geq 0)$$

for a continuous-time system (in both the discrete and continuous cases this would be referred to as the **orbit of** x_0 under the dynamics). For a given event $A \in \mathcal{F}$, the observer computes the **empirical frequencies** of occurrence of A in the orbit, namely

$$\mu_A^{(n)}(x_0) = \frac{1}{n} \#\{1 \leq k \leq n \, : \, x_k \in A\} = \frac{1}{n} \sum_{k=0}^{n-1} \mathbf{1}_A(x_k), \qquad (n \geq 1),$$

or, in the case of a continuous-time system

$$\mu_A^{(s)}(x_0) = \frac{1}{s} \int_0^s \mathbf{1}_A(x_s) \, ds, \qquad (s \geq 0).$$

One might expect that in a typical situation, the quantity $\mu_A^{(n)}(x_0)$ or its continuous-time analogue $\mu_A^{(s)}(x_0)$ should converge (as n or s tend to infinity) to a limit that is a constant and independent of x_0, except possibly for some small set of "badly-behaved" initial states x_0. If that is the case, we might denote this limit by $\mathbf{P}(A)$ and say that it represents the "ideal" statistics of the system.

A more general way of recovering the statistics of the system is to look at **observables**, which are measurable functions $f : \Omega \to \mathbb{R}$ on the phase space (an observable is the dynamical systems or physics equivalent term for a random variable, really). For an observable f we can form the **ergodic average**

$$\mu_f^{(n)}(x_0) = \frac{1}{n}\sum_{k=0}^{n-1} f(x_k) = \frac{1}{n}\sum_{k=0}^{n-1} f(T^k(x_0)),$$

(or the analogous continuous-time quantity, whose form we leave to the reader to write down), and hope that again the ergodic averages converge to a limit, which is independent of x_0 and represents the "ideal" average value of the observable f, denoted $\mathbf{E}(f)$ (in physics, usually this would be denoted $\langle f \rangle$). By computing this ideal average for many different observables we can recover all the information on the probability measure $\mathbf{P}$.

One can now ask whether the nice situation described above actually happens in practice. Coming back to the example of billiards, it is easy to see that for some shapes of the billiard "table" one cannot hope to recover any meaningful statistics for the system, for what may be a trivial reason. For example, a rectangular table has the property that the ratio of the absolute values of the horizontal and vertical components of the initial speed of the ball is always preserved (equivalently, the quantity $|\tan(\alpha)|$ where α is the initial angle is preserved). Thus, by observing the trajectory of a single ball we have no hope of recovering any meaningful information on the statistics of the system when started with a ball for which the "invariant" quantity $|\tan(\alpha)|$ is different. In this case we say that the billiard dynamical system on a rectangular domain is **non-ergodic**. Less trivially, an ellipse-shaped billiard can also be shown to be non-ergodic, because of a less obvious geometric invariance property: it can be shown that an orbit will not fill the entire ellipse but will have a non-trivial envelope which is either a smaller ellipse, a hyperbola, a closed polygon or a line (see `http://cage.ugent.be/~hs/billiards/billiards.html`, and Figure 2).

On the other hand, in many cases, such as the domains shown in Figure 1, it can be

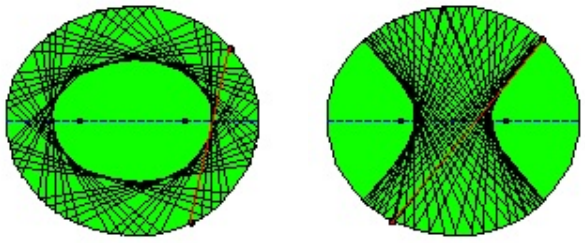


Figure 2: Billiard in an ellipse-shaped domain

proved that the nice situation exists, i.e., the billiard *is* **ergodic** (we will define later what that actually means). This is related (in a way that is difficult to articulate precisely), to the emergence of a kind of "chaos" – i.e., the billiard ball trajectories are erratic and irregular rather than forming a nice pattern as in the trivial examples discussed above. When ergodicity holds, the statistics of the system can be recovered from the typical trajectory of a single ball; in the case of billiards, it turns out that these statistics are quite interesting: the underlying measure $\mathbf{P}$ on the phase space (which may be parametrized in terms of three parameters ϕ, θ, ℓ — see the article [3] for the meaning of these quantities) takes the form

$$\mathbf{P}(A) = \iiint_A \frac{\sin\theta}{\sin\theta_1}\, d\theta\, d\phi\, d\ell.$$

Note that even when the system is ergodic, there may be exceptional orbits from which one cannot recover any statistics. For example, in the Bunimovich stadium shown in Figure 1, a trajectory that starts in a vertical direction starting in the rectangular area bounded between the two semi-circles will be a periodic vertical line. However, the key point is that such trajectories are atypical examples that only occur on a measure 0 set of the phase space.

It should also be noted that in any given example, *proving* that the ergodicity property holds may be extremely difficult. In fact, the family of dynamical systems (and even more restrictively billiard systems) for which ergodicity has been proved rigorously is quite limited, and in practical dynamical systems that one encounters in physics or other applied areas usually this is assumed without proof, as long as there is a sufficiently strong intuition that allows one to rule out a "trivial" reason why ergodicity should fail to hold. (This assumption is sometimes referred to as the ergodic hypothesis.)

In the next few sections, we shall start developing the basic ideas of ergodic theory in a

more formal and precise way. The key concept is of a **measure-preserving system**, which is a probability space together with a **measure-preserving map** representing the dynamics of the system. The main result we will prove is the fundamental result of ergodic theory, known as **Birkhoff's pointwise ergodic theorem**. It explains precisely the connection between the notion of ergodicity and the ability to "recover the statistics of the system" as illustrated above. We shall also give some important examples and explain why the study of ergodic theory is natural from the point of view of probability theory, since one can consider the Birkhoff ergodic theorem as a powerful generalization of the strong law of large numbers.

Chapter 2: Measure preserving systems

2.1 Measure preserving systems

In the previous section we cheated a little bit by considering dynamical systems without an underlying measurable structure or notion of measure (in fact, such a structure was implicit in the discussion of the orbit of a "typical" or "random" initial state). In ergodic theory we concentrate on dynamical systems which come equipped with a measure, and furthermore, we require the measure to be preserved under the action of the dynamics. This idea leads to the following definitions.

Definition 2.1. *Let $(\Omega, \mathcal{F}, \mathbf{P})$ be a probability space. A measurable map $T : \Omega \to \Omega$ is called* ***measure preserving*** *if for any event $E \in \mathcal{F}$ we have*

$$\mathbf{P}(T^{-1}(E)) = \mathbf{P}(E). \tag{1}$$

If T is measure preserving, we say that the probability measure $\mathbf{P}$ is ***invariant under*** *T.*

The condition (1) is sometimes written in the form $\mathbf{P} = \mathbf{P} \circ T^{-1}$. This can be interpreted as the statement that the push-forward of $\mathbf{P}$ under T is again $\mathbf{P}$; that is, if X is an Ω-valued random variable with distribution $\mathbf{P}$, then $T(X)$ has the same distribution.

Definition 2.2. *A* ***measure preserving system*** *is a probability space equipped with a measure preserving map, i.e., a quadruple $(\Omega, \mathcal{F}, \mathbf{P}, T)$, where $(\Omega, \mathcal{F}, \mathbf{P})$ is a probability space and $T : \Omega \to \Omega$ is a measure preserving map.*

Measure preserving systems are the fundamental objects studied in ergodic theory (just like vector spaces are the fundamental objects of linear algebra, topological spaces are the fundamental objects of topology, etc.). It makes sense to ask to see some examples of such systems before proceeding with their theoretical study. Aside from some very interesting measure preserving systems that originate in dynamical systems (such as the billiard systems mentioned in the previous chapter), a huge class of examples arise in a very natural way in probability theory, and are intimately related to the notion of a **stationary sequence**, which is the subject of the next section.

2.2 Stationary sequences

Let $(X_n)_{n=1}^{\infty}$ be a sequence of random variables. The sequence is called **stationary** if for any $n, m \geq 1$, we have the equality in distribution

$$(X_n, \ldots, X_{n+m-1}) \stackrel{\mathcal{D}}{=} (X_1, \ldots, X_m). \tag{2}$$

Note that in particular this implies that the variables $X_1, X_2, \ldots$ are identically distributed. Stationarity is a stronger property that also ensures that any pair of successive variables (X_n, X_{n+1}) is equal in distribution to the first pair (X_1, X_2), any triple (X_n, X_{n+1}, X_{n+2}) is equal in distribution to the first triple (X_1, X_2, X_3), etc.; that is, any probabilistic question about a block of adjacent variables does not depend on the "origin" of the block. An i.i.d. sequence is a trivial example of a stationary sequence.

A stationary sequence gives rise in a natural way to a measure preserving system known as the **shift dynamics**. To define it, first note that although the variables may be defined on a generic probability space $(\Omega, \mathcal{F}, \mathbf{P})$, there is no real loss of generality in assuming that the probability space is the **canonical product space**

$$\Omega = \mathbb{R}^{\mathbb{N}}$$

(sometimes denoted by $\mathbb{R}^{\infty}$) together with the product σ-algebra $\mathcal{B} = \mathcal{B}(R^{\mathbb{N}})$, and the probability measure μ defined by

$$\mu(E) = \mathbf{P}((X_1, X_2, \ldots) \in E),$$

(i.e., the distribution measure of the infinite-dimensional vector $(X_1, X_2, \ldots)$). In this representation, the random variables are simply the **coordinate functions**

$$X_n(\omega) = \pi_n(\omega) = \omega_n,$$

where $\omega = (\omega_1, \omega_2, \ldots) \in \mathbb{R}^{\mathbb{N}}$.

On the space $(\mathbb{R}^{\mathbb{N}}, \mathcal{B}, \mu)$ we define the **shift map** $S : \mathbb{R}^{\mathbb{N}} \to \mathbb{R}^{\mathbb{N}}$ by

$$S(\omega_1, \omega_2, \omega_3, \ldots) = (\omega_2, \omega_3, \omega_4, \ldots).$$

Lemma 2.3. *The shift map S is a measure preserving map of the probability space $(\mathbb{R}^{\mathbb{N}}, \mathcal{B}, \mu)$ if and only if the sequence $(X_n)_{n=1}^{\infty}$ is stationary.*

Exercise 2.4. *Prove Lemma 2.3.*

Definition 2.5. *If $(X_n)_{n=1}^{\infty}$ the measure preserving system $(\mathbb{R}^{\mathbb{N}}, \mathcal{B}, \mu, S)$ described above is called the* ***one-sided shift map*** *(or sometimes just* ***shift map****) associated to $(X_n)_{n=1}^{\infty}$.*

What about a two-sided shift map? One can consider a two-sided infinite sequence $(X_n)_{n=-\infty}^{\infty}$, and say that it is stationary if the equation (2) holds for any $m \geq 1$ and $n \in \mathbb{Z}$. One may associate with such a stationary sequence the **two-sided shift dynamics**, which is the measure preserving system $(\mathbb{R}^{\mathbb{Z}}, \mathcal{B}(\mathbb{R}^{\mathbb{Z}}), \mu, S)$, where as before μ is the distribution measure of the sequence $(X_n)_{n\in\mathbb{Z}}$, and S is the two-sided shift, given by

$$S((\omega_n)_{n\in\mathbb{Z}}) = (\omega_{n+1})_{n\in\mathbb{Z}}.$$

One may check easily that Lemma 2.3 remains true when replacing the one-sided concepts of stationary sequence and shift dynamics with their two-sided analogues.

From the definitions it may appear that the notion of a two-sided stationary sequence is more general than that of a one-sided shift, since half of the elements of a two-sided stationary sequence $(X_n)_{n\in\mathbb{Z}}$ can be removed to give a one-sided stationary sequence $(X_n)_{n\geq 1}$. However, in fact this is not the case, as the next result shows.

Lemma 2.6. *Given a one-sided stationary sequence $(X_n)_{n\geq 1}$, there exists a two-sided stationary sequence $(Y_n)_{n\in\mathbb{Z}}$ defined on some probability space such that $(Y_n)_{n\geq 1} \stackrel{\mathcal{D}}{=} (X_n)_{n\geq 1}$.*

Proof. This is a simple example of an application of the Kolmogorov extension theorem, a useful result from measure theory that enables one to construct measures on infinite product spaces with prescribed finite-dimensional marginals (see [4, Sec. A.3]). Here, the stationarity condition (2) determines the joint m-dimensional distribution of any block $(Y_n, \ldots, Y_{n+m-1})$ of m successive random variables in the sequence, where $m \geq 1$ and $n \in \mathbb{Z}$. These distributions satisfy the consistency condition in the Kolmogorov extension theorem, and therefore are indeed the m-dimensional marginals of some infinite sequence $(Y_n)_{n\in\mathbb{Z}}$ defined on a single probability space. □

We saw that we can associate with any stationary sequence a measure preserving system. Going in the opposite direction, if we start with a measure preserving system $(\Omega, \mathcal{F}, \mathbf{P}, T)$,

any random variable $X : \Omega \to \mathbb{R}$ (what we called an *observable* in the previous chapter) can be transformed by T to a new variable

$$X \circ T = X(T).$$

The measure preserving property implies that $X \circ T$ is equal in distribution to X. By starting with X and repeatedly iterating the transformation T we get a sequence $(X_n)_{n=1}^{\infty}$ given by

$$X_n = X \circ T^{n-1}.$$

Lemma 2.7. *$(X_n)_n$ is a stationary sequence.*

Exercise 2.8. *Prove Lemma 2.7*

The conclusion from the above discussion is that the study of stationary sequences is roughly equivalent to the study of measure preserving systems with a distinguished observable, and indeed much of ergodic theory could be developed using just the language of stationary sequences, although this would come at great cost to the elegance and beauty of the theory.

2.3 Examples of measure preserving systems

1. **i.i.d. sequences.** As mentioned in the previous section, any i.i.d. sequence is stationary and hence has an associated shift measure preserving system, referred to as an **i.i.d. shift**. In the case when the i.i.d. random variables take on only a finite number of values with positive probability this measure preserving system is known as a **Bernoulli shift**.

2. **A shift-equivariant function of a stationary sequence.** Given a stationary sequence $(X_n)_{n=1}^{\infty}$ and a measurable function $F : \mathbb{R}^{\mathbb{N}} \to \mathbb{R}^{\mathbb{N}}$ one can manufacture a new stationary sequence $(Y_n)_{n=1}^{\infty}$ via the equation

$$Y_n = F(X_n, X_{n+1}, X_{n+2}, \ldots), \qquad (n \geq 1). \tag{3}$$

   The verification that $(Y_n)_n$ is stationary is easy and is left to the reader. In this way one can generate starting from a known stationary sequence (e.g., an i.i.d. sequence) a large class of new and interesting sequences.

3. **Stationary finite-state Markov chains.** Let $A = \{\alpha_1, \ldots, \alpha_d\}$ be a finite set. A **finite-state Markov chain with state space** A is a sequence $(X_n)_{n=0}^{\infty}$ of A-valued random variables such that for each $n \geq 0$ and $1 \leq j_1, j_2, \ldots, j_{n+1} \leq d$ we have that

$$\mathbf{P}(X_{n+1} = \alpha_{j_{n+1}} \,|\, X_1 = \alpha_1, \ldots, X_n = \alpha_n) = \mathbf{P}(X_{n+1} = \alpha_{j_{n+1}} \,|\, X_n = \alpha_n). \tag{4}$$

That is, the conditional distribution of X_{n+1} given the n preceding values $X_1, \ldots, X_n$ is only dependent on the value of the last observed variable X_n; this property is known as the **Markov property**. In most cases the chain is also assumed to be **time-homogoneous**, meaning that the expression in (4) is independent of n. In this case, if we denote

$$p_{i,j} = \mathbf{P}(X_2 = \alpha_j \,|\, X_1 = \alpha_i), \qquad (1 \leq i, j \leq d),$$

then the matrix $P = (p_{i,j})_{i,j=1}^{d}$ together with the probability distribution of the initial state X_0 determine the distribution of the entire sequence. The probability $p_{i,j}$ is referred to as the **transition probability from state** i **to** j, and the matrix P is called the **transition matrix** of the chain. The distribution of X_0 is usually given as a probability vector $\pi = (\pi_1, \ldots, \pi_d)$ where $\pi_j = \mathbf{P}(X_0 = j)$. It is easy to show that the vector $\pi^{(n)} = (\pi_1^{(n)}, \ldots, \pi_d^{(n)})$ representing the probability distribution of X_n is obtained from π and P via

$$\pi^{(n)} = \pi P^n,$$

the linear-algebraic result of multiplying the row vector π by the matrix P multiplied by itself n times.

Assume now that π is chosen to be a probability vector satisfying the equation $\pi = \pi P$; i.e., π is a left-eigenvector of the transition matrix P with eigenvalue 1. By the above remarks, this means that the sequence $(X_n)_n$ is a sequence of identically distributed random variables, and furthermore it is easy to see that $(X_n)_n$ is in fact a stationary sequence. A Markov chain started with such an initial state distribution is called a **stationary Markov chain**. The associated shift measure preserving system is known as a **Markov shift**.

4. **Tossing a randomly chosen coin.** Let $0 \leq U \leq 1$ be a random variable. We can define a stationary sequence $X_1, X_2, \ldots$ by the following "two-step experiment": first, pick a random coin with bias U; then, toss the chosen coin infinitely many times (the

coin tosses being independent of each other), denoting the results (encoded as 0's or 1's) by $X_1, X_2, \ldots$. Formally, we can define the distribution of the sequence by

$$\mathbf{P}(X_1 = a_1, \ldots, X_n = a_n) = \mathbf{E}\left[U^{\sum_j a_j}(1-U)^{n-\sum_j a_j}\right], \qquad a_1, \ldots, a_n \in \{0,1\}.$$

Note that the X_n's are identically distributed (in fact, the sequence is stationary), but *not* independent (except in the extreme case when U is a.s. constant); rather, they are said to be conditionally independent given U.

5. **Pólya's urn experiment.** Pólya's urn is a famous probabilistic model that illustrates many concepts in probability theory, including the notion of martingale, stationary sequences, exchangeable sequences, and more. Picture an urn that originally contains a white balls and b black balls. The experimenter samples a uniformly random ball from the urn, examines its color, then puts the ball back and adds another ball of the same color; this is repeated to infinity. Let X_n denote the number of white balls in the urn after the nth step. Clearly $X_0 = a$ and the distribution of X_{n+1} can be expressed most naturally by conditioning on X_n, namely

$$X_{n+1}\big|_{X_n=m} = \begin{cases} m+1 & \text{with probability } \frac{m}{n+a+b}, \\ m & \text{with probability } \frac{n+a+b-m}{n+a+b}. \end{cases} \tag{5}$$

Let I_n be the indicator random variable of the event that in the nth sampling step a white ball was drawn. A surprising property of the sequence $(I_n)_{n=1}^{\infty}$ is that it is a stationary sequence, and therefore has an associated measure preserving shift. In fact, we'll prove that a stronger claim is true:

Lemma 2.9. *The sequence $(I_n)_{n=1}^{\infty}$ is invariant in distribution under finite permutations.*[1] *That is, for any permutation σ of the numbers $\{1, 2, \ldots, n\}$, the random vector $(I_{\sigma(1)}, I_{\sigma(2)}, \ldots, I_{\sigma(n)})$ is equal in distribution to the random vector $(I_1, \ldots, I_n)$.*

Proof. It is enough to prove this when the permutation is an adjacent transposition, that is, a permutation that swaps the positions of the numbers k and $k+1$ for some value

[1] A sequence of r.v.'s with this property is called **exchangeable**. There is an important result about such sequences (that we will not talk about here) called **De-Finetti's theorem**, which you might want to read about.

$k \geq 1$; if the claim holds for such permutations then it holds for permutations obtained by composing two or more adjacent transpositions, and it is well-known that any permutation of n numbers can be obtained as such a composition.

To prove the claim for the adjacent transposition that swaps k and $k+1$, note that, conditioned on the event that at stage $k-1$ of the experiment the urn contained A white and B black balls, the probability to draw "white then black" in the next two steps would be

$$\frac{A}{k+a+b} \cdot \frac{B}{k+a+b+1}.$$

On the other hand, the probability of drawing "black then white" is

$$\frac{B}{k+a+b} \cdot \frac{A}{k+a+b+1}.$$

Since these two probabilities are equal, it follows that the two-dimensional random vectors (I_k, I_{k+1}) and (I_{k+1}, I_k) are equal in distribution conditionally on $(I_1, \ldots, I_{k-1})$. From here it is a small step (left as an exercise) to conclude the equality in distribution

$$(X_1, \ldots, X_{k-1}, X_{k+1}, X_k, X_{k+2}, \ldots, X_n) \stackrel{\mathcal{D}}{=} (X_1, \ldots, X_{k-1}, X_k, X_{k+1}, X_{k+2}, \ldots, X_n),$$

which finishes the proof of the lemma. □

Exercise 2.10. *Complete the argument in the proof of Lemma 2.9 above.*

Exercise 2.11. *Explain why Lemma 2.9 implies that the sequence $(I_n)_{n=0}^{\infty}$ is stationary.*

Exercise 2.12. (a) *Use the result of Lemma 2.9 to prove that the joint distribution of $I_1, \ldots, I_n$ can be written in terms of the following explicit formula: for any $x_1, \ldots, x_n \in \{0, 1\}$, we have*

$$\mathbf{P}(I_1 = x_1, \ldots, I_n = x_j) = \frac{a(a+1)\ldots(a+k-1) \cdot b(b+1)\ldots(b+n-k-1)}{(a+b)(a+b+1)\ldots(a+b+n)}, \qquad (6)$$

where $k = \sum_{j=1}^{n} x_j$.

(b) *From* (6)*, deduce that the probability of having p white balls after n steps is*

$$\mathbf{P}(X_n = p) = \binom{n}{p-a} \frac{a(a+1)\ldots(a+p-a-1) \cdot b(b+1)\ldots(b+n-p+a-1)}{(a+b)(a+b+1)\ldots(a+b+n)} \qquad (7)$$

for $a \le p \le n + a$.

(c) *Use the above formula* (7) *for the distribution of* X_n *to prove that the* proportion *of white balls in the urn converges to a limiting beta distribution. Specifically, prove that*

$$\frac{X_n}{n+a+b} \implies \text{Beta}(a, b). \tag{8}$$

(In fact, $X_n/(n + a + b)$ *converges* almost surely *to a limiting random variable* Y *that has the beta distribution* $\text{Beta}(a, b)$ *as its probability distribution. This follows from the martingale convergence theorem; see [11, Sec. 4.1].)*

6. **Rotation of the circle** (a.k.a. $x+\alpha$ **modulo 1**). Let $(\Omega, \mathcal{F}, \mathbf{P})$ be the unit interval $[0, 1]$ with Lebesgue measure. One can consider $[0, 1]$ to be topologically a circle by identifying both endpoints 0 and 1 as a single point. Fix $0 \le \alpha < 1$. The **circle rotation map** $R_\alpha : [0, 1) \to [0, 1)$, which rotates the circle by a fraction α, is defined by

$$R_\alpha(x) = x + \alpha \bmod 1 = \begin{cases} x + \alpha & \text{if } x + \alpha < 1, \\ x + \alpha - 1 & \text{otherwise.} \end{cases}$$

Lemma 2.13. *R_α preserves Lebesgue measure.*

Proof. If $A \subset [0, 1]$ is a Borel set, then

$$\begin{aligned} \text{Leb}(R_\alpha^{-1}(A)) &= \text{Leb}\Big(\big(A \cap [0, \alpha) + 1\big) \sqcup \big(A \cap [\alpha, 1) - 1\big)\Big) \\ &= \text{Leb}\big(A \cap [0, \alpha) + 1\big) + \text{Leb}\big(A \cap [\alpha, 1) - 1\big) \\ &= \text{Leb}\big(A \cap [0, \alpha)\big) + \text{Leb}\big(A \cap [\alpha, 1)\big) \\ &= \text{Leb}\Big(\big(A \cap [0, \alpha)\big) \sqcup \big(A \cap [\alpha, 1)\big)\Big) = \text{Leb}(A). \end{aligned}$$

□

7. **The** $2x$ **mod** 1 **map.** Similarly to the previous example, the $2x$ **mod** 1 **map** or **doubling map** is also defined on the probability space $[0, 1]$ with Lebesgue measure, and is given by

$$D(x) = 2x \bmod 1 = \begin{cases} 2x & x < \frac{1}{2}, \\ 2x - 1 & x \ge \frac{1}{2}. \end{cases}$$

Lemma 2.14. *D preserves Lebesgue measure.*

Proof. If $A \subset [0,1]$ is a Borel set, then

$$\mathrm{Leb}(D^{-1}(A)) = \mathrm{Leb}\left(\tfrac{1}{2}A \sqcup \left(\tfrac{1}{2}A + \tfrac{1}{2}\right)\right) = \tfrac{1}{2}\,\mathrm{Leb}(A) + \tfrac{1}{2}\,\mathrm{Leb}(\tfrac{1}{2} + A) = \mathrm{Leb}(A).$$

□

We have seen before that the measure space $[0,1]$ with Lebesgue measure is isomorphic to the product space of an infinite sequence of i.i.d. unbiased coin tosses. It is easy to see that under this isomorphism, the doubling map translates to the shift map S of the Bernoulli sequence. So, the doubling map is really a disguised version of the Bernoulli shift associated with i.i.d. unbiased coin tosses.

8. **The continued fraction map.** A well-known fact from number theory says that any rational number $x \in (0,1)$ has a unique **continued fraction expansion** of the form

$$x = \cfrac{1}{n_1 + \cfrac{1}{n_2 + \cfrac{1}{n_3 + \cfrac{1}{\ddots + \frac{1}{n_k}}}}},$$

where $k \geq 1$, $n_1, \ldots, n_k \in \mathbb{N}$ and $n_k > 1$. Such an expansion is said to be finite, or terminating. Similarly, any irrational $x \in (0,1)$ has a unique *infinite* continued fraction expansion, which takes the form

$$x = \cfrac{1}{n_1 + \cfrac{1}{n_2 + \cfrac{1}{n_3 + \cfrac{1}{n_4 + \frac{1}{\cdots}}}}},$$

where $n_1, n_2, n_3 \ldots \in \mathbb{N}$. The numbers $n_1, n_2, \ldots$ are called the **quotients** of the expansion, and are analogous to the digits in the decimal (or base-b) expansion of a real number. They are computed using a process that is a natural generalization of the Euclidean al-

gorithm to real numbers, namely:

$n_1 =$ the number of times a stick of length x "fits" inside a stick of length 1,
$n_2 =$ the number of times a stick of length $x_2 = (1 - n_1 x)$ fits inside a stick of length x,
$n_3 =$ the number of times a stick of length $x_3 = (x - n_2 x_1)$ fits inside a stick of length x_2,
$\vdots$

Gauss studied in 1812 the statistical distribution of the quotients for a number x chosen uniformly at random in $(0, 1)$. In this case, since x is irrational with probability 1 we need not worry about terminating expansions, and can consider the quotients $n_1, n_2, \ldots$ to be random variables defined on the measure space $(0, 1)$ equipped with Lebesgue measure. Gauss reformulated the problem in terms of a measure preserving system (before this concept even existed!) now called the **continued fraction map** or **Gauss map**. To see how this reformulation works, note first that, in the computation above, the first quotient n_1 can be represented in the form

$$n_1 = \left\lfloor \frac{1}{x} \right\rfloor$$

(where $\lfloor z \rfloor$ denotes as usual the integer part of a real number z). Next, observe that, to continue with the computation of the next quotients $n_2, n_3, \ldots$, instead of replacing the two yardsticks of lengths 1 and x (which are used in the computation of the first quotient n_1) by a pair of yardsticks of lengths x and $x_2 = 1 - n_1 x$, one can instead rescale the yardstick of length x to be of length 1, so that the yardstick of length x_2 becomes of length

$$x' = \frac{1 - n_1 x}{x} = \frac{1}{x} - n_1 = \left\{ \frac{1}{x} \right\}$$

(where $\{z\} = z - \lfloor z \rfloor$ is the **fractional part** of z). The quotient n_2 can be computed from this rescaled value x' in the same way that n_1 is computed from x. By continuing in this way one can obtain all the quotients by successive rescaling operations. Formally, define the Gauss map $G : (0, 1) \to [0, 1)$ and a function $N : (0, 1) \to \mathbb{N}$ by

$$G(x) = \left\{ \frac{1}{x} \right\}, \qquad N(x) = \left\lfloor \frac{1}{x} \right\rfloor.$$

Then the above comments show that the quotients $n_1, n_2, \ldots$ are obtained by

$$\begin{aligned} n_1 &= N(x), \\ n_2 &= N(G(x)), \\ n_3 &= N(G^2(x)), \ \ldots \\ n_k &= N(G^{k-1}(x)), \ \ldots \end{aligned}$$

(Note that the range of G is $[0, 1)$ instead of the open interval $(0, 1)$ since $G(x) = 0$ exactly when x is a rational number of the form $x = 1/m$; this is related to the fact that if we start with any rational number x, after a finite number of iterations of G we will reach 0 and will not be able to extract any more quotients.)

If you guessed that the Gauss map G preserves Lebesgue measure, you guessed wrong. The real situation is more interesting:

Lemma 2.15. *The map G preserves the* ***Gauss measure*** γ *on* $(0, 1)$*, given by*

$$\gamma(A) = \frac{1}{\log 2} \int_A \frac{dx}{1+x}.$$

Exercise 2.16. *Prove Lemma 2.15.*

An important observation is that Gauss measure and Lebesgue measure are mutually absolutely continuous with respect to each other. This means that any event which has probability 1 with respect to one is also a probability 1 event with respect to the other. Thus any almost-sure statistical results about the measure preserving system $((0, 1), \mathcal{B}, \gamma, G)$ (which will be obtained from the Birkhoff ergodic theorem once we develop the theory a bit more) will translate immediately to statements about the behavior of the continued fraction expansion of a *uniformly random* real number.

The continued fraction map described above is intimately related to the Euclidean algorithm for computing the greatest common divisor (GCD) of two integers, since iterating the map starting from a rational fraction p/q reproduces precisely the sequence of quotients (and remainders, if one takes care to record them) in the execution of the Euclidean algorithm, and the last non-zero iteration $T^k(p/q)$ is of the form $1/d$, where d is precisely the GCD of p and q. Given the usefulness of the Euclidean algorithm and its historical

status as one of the earliest algorithms ever described, is not surprising that already in the early days of the theory of algorithms (a.k.a. the 1960's) researchers were interested in giving a quantitative analysis of the running time of this venerable procedure. Such analyses lead directly to ergodic theoretic questions about the continued fraction map; the renewed interest in this classical problem has stimulated new and extremely interesting studies into the mathematics of the Gauss map. A highly readable account of these fascinating developments (some of which are topics of active current research) is told in [5, Sections 4.5.2–4.5.3].

9. **The binary GCD algorithm.** Continuing the discussion above, a fact that is little-known outside computer science circles is that in modern times a new algorithm for computing GCD's was proposed that gives the Euclidean algorithm a serious run for its money, and is actually faster in some implementations. This algorithm was proposed by Josef Stein in 1967 and is known as **the binary GCD algorithm** or **Stein's algorithm**. It replaces the integer division operations of the Euclidean algorithm, which are costly in some computer architectures, with a clever use of subtractions (which are generally cheap) and divisions by 2, which can be implemented in machine language as (also cheap) bit shift operations.

   [Here is a summary of the algorithm: start with two integers $u < v$. First, extract the common power-of-2 factor to get to a situation where at least one of u, v is odd. Then, successively replace (u, v) with the new pair $(v - u)/2^k, u$ (sorted so that the smaller one gets called "u" and the bigger one "v"), where 2^k is the maximal power of 2 dividing $v - u$. Eventually one of the numbers becomes 0 and the remaining one represents the odd component of the GCD of the original numbers.]

   The computer scientist Richard Brent noticed in 1976 that this algorithm can also be reformulated in terms of a dynamical system. Similarly to the case of the Euclidean algorithm, the theoretical analysis of the running time of the binary GCD algorithm leads to highly nontrivial questions (most of them still open) about the behavior of this dynamical system. In particular, this system has an invariant measure that is mutually absolutely continuous with respect to Lebesgue measure, and is analogous to the Gauss measure, but no good formula for it is known. See [5, Sec. 4.5.3] for more details.

10. **The $3x+1$ map.** The **3x+1 problem** or **Collatz problem** is a famous open problem (studied since the 1950's, and originating in work of L. Collatz around 1932) about a discrete dynamical system on the positive integers. It pertains to iterations of the map $T : \mathbb{N} \to \mathbb{N}$ defined by

$$T(n) = \begin{cases} 3n+1 & \text{if } n \text{ is odd,} \\ \frac{n}{2} & \text{if } n \text{ is even.} \end{cases}$$

The conjecture is that for any initial number n_0, iterating the map will eventually lead to the cycle $1, 4, 2, 1, 4, 2, 1, \ldots$ The mathematician Paul Erdös was quoted as saying "Mathematics is not yet ready for such problems" and offered a $500 prize for its solution.

One of the many (ultimately unsuccessful) attempts to study the problem was based on the beautiful observation that this dynamical system can be turned into a measure preserving system, by extending its domain of definition to the ring $\mathbf{Z}_2$ of **2-adic integers**. This is an extension of the usual ring $\mathbb{Z}$ of integers in which every element has a binary expansion that extends *infinitely far to the left* (instead of to the right as a real number would). That is, a dyadic integer is a formal expression of the form

$$a_0 + 2 \cdot a_1 + 4 \cdot a_2 + 8 \cdot a_3 + \ldots + 2^n a_n + \ldots = \sum_{n=0}^{\infty} a_n 2^n$$

where $a_0, a_1, a_2, \ldots \in \{0, 1\}$. It can be shown that one can do algebra, and even an exotic form of calculus, on these numbers (and more generally over similar sets of numbers in which the binary expansion is replaced by a base-p expansion where p is an arbitrary prime number — these are the so-called **p-adic integers**). Since the notion of the parity of a number extends to 2-adic integers, the $3x+1$ map T extends in an obvious way to a map $\tilde{T} : \mathbf{Z}_2 \to \mathbf{Z}_2$. It can be shown that $\tilde{T}$ preserves the natural volume measure of $\mathbf{Z}_2$. For more information, see Wikipedia and [6].

11. **Billiards.** In Chapter 5 we discussed billiard dynamical systems, and mentioned a formula on the limiting statistics of such a system, in the case when it is ergodic. This is related to the fact that the billiard dynamics also has an invariant measure, given (in a suitable parametrization of the phase space) by

$$\mu(A) = \iiint_A \frac{\sin\theta}{\sin\theta_1} \, d\theta \, d\phi \, d\ell.$$

12. **Hamiltonian flow. Hamiltonian mechanics** is a formalism for modeling a mechanical system of particles and rigid bodies interacting via physical forces, with no external influences. The phase space is some set Ω representing the possible states of the system (formally, it is a **symplectic manifold**, and has a smooth structure — i.e., one can solve differential equations on it and do other calculus-type operations). The **Hamiltonian flow** is a semigroup of maps $(H_s)_{s\geq 0}$ representing the time-evolution of the system, i.e., $H_s(\omega)$ takes an initial state $\omega \in \Omega$ of the system and returns a new state representing the state of the system s time units in the future. A result known as Liouville's theorem says that the natural volume measure of the manifold is preserved under the Hamiltonian system. Thus, the Hamiltonian flow is a **measure preserving flow** (the continuous-time analogue of a measure preserving system, which we will not discuss in detail). Such flows provided some of the original motivation for questions of ergodic theory, since, e.g., statistical physicists in the 19th century wanted to understand the statistical behavior of ideal gases (note that billiard can be thought of a toy model for a gas in an enclosed region).

13. **Geodesic flow.** On a compact Riemann surface (or more generally a Riemannian manifold), the geodesic flow $(\varphi_s)_{s\geq 0}$ is a family of maps, where each φ_s takes a point on the manifold together with a "direction" at s (formally, an element of the tangent space at s), and returns a new pair "point+direction" that is obtained by proceeding s units of distance along the unique geodesic curve originating from s in the given direction. (For a more formal description, see Wikipedia or a textbook on differential geometry). The geodesic flow preserves the volume measure and is thus a measure preserving flow.

14. **The logistic map.** The logistic map was originally studied as a simple model for the dynamics of population growth of animal and plant species. It is given by the formula

$$L_r(x) = rx(1-x) \qquad (0 < x < 1),$$

where $r > 0$ is a parameter of the system. Here, x represents the size of the population, and $L_r(x)$ represents the size of the population one generation later, so successive iterations $L_r^n(x)$ correspond to the evolution of the population sizes over time starting from some initial size x. The assumptions underlying the model are that when x is small one should observe roughly exponential growth when iterating the map, but as the size

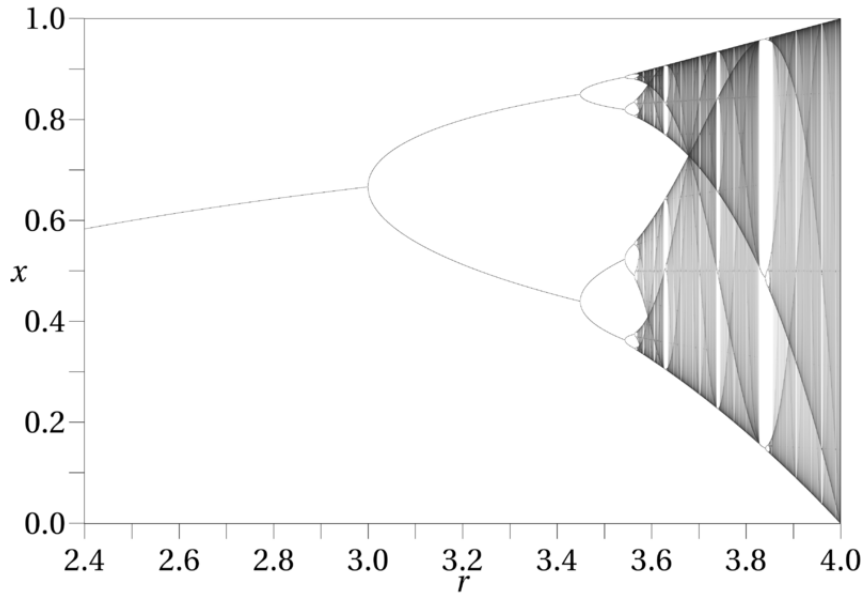


Figure 3: Chaos in the logistic map (source: Wikipedia)

of the population increases, the environmental resources required to support growth are depleted, leading to starvation and a sharp decrease in the population size.

The logistic map is a famous example of the emergence of **chaos**: for values of r between 0 and 3, the system stabilizes around a unique value (0 if $r \leq 1$, or $(r-1)/r$ if $1 \leq r \leq 3$). When r becomes slightly bigger than 3 a **bifurcation** occurs, leading to an oscillation between 2 values; as r increases further, additional bifurcations occur (oscillation betweeen 4 values, 8 values etc.) until chaotic behavior emerges at $r \approx 3.57$ and continues (with occasional intervals of stability) until $r = 4$, after which point the range of the map leaves $[0, 1]$ so the model stops making sense as a dynamical system. See Figure 3 for an illustration of this remarkable phenomenon.

Lemma 2.17. *When $r = 4$, the map L_4 has an invariant measure λ on $(0, 1)$ given by*

$$\lambda(dx) = \frac{1}{\pi\sqrt{x(1-x)}}\,dx$$

(also known as the Beta$(\frac{1}{2}, \frac{1}{2})$ *distribution).*

Exercise 2.18. *Prove Lemma 2.17*

2.4 Ergodicity

Let $(\Omega, \mathcal{F}, \mathbf{P}, T)$ be a measure preserving system. An event $A \in \mathcal{F}$ is called **T-invariant** (or **invariant under** T, or just **invariant** if the context is clear) if

$$T^{-1}(A) = A \text{ a.s.},$$

with the convention that two events A, B are considered equal almost surely if their symmetric difference has probability 0. That is, A is invariant if

$$\mathbf{P}(A \triangle T^{-1}(A)) = 0,$$

(where $A \triangle B = (A \setminus B) \cup (B \setminus A)$ denotes the symmetric difference of two sets). We denote by $\mathcal{I}$ the collection of a.s. invariant events.

Lemma 2.19. *$\mathcal{I}$ is a σ-algebra.*

Exercise 2.20. *Prove Lemma 2.19.*

Definition 2.21. *The measure preserving system $(\Omega, \mathcal{F}, \mathbf{P}, T)$ is called **ergodic** if for any invariant event A, $\mathbf{P}(A) = 0$ or $\mathbf{P}(A) = 1$.*

A sub-σ-algebra of $\mathcal{F}$ all of whose events have probability 0 or 1 is called **trivial**. (We already saw an example: the σ-algebra of tail events of an i.i.d. sequence of random variables is trivial, according to the Kolmogorov 0-1 law.) So, another way of saying that a measure preserving system is ergodic is that its σ-algebra $\mathcal{I}$ of invariant events is trivial.

There is an equivalent way to characterize ergodicity in terms of invariant random variables rather than events, given in the following exercise.

Exercise 2.22. *If $(\Omega, \mathcal{F}, \mathbf{P}, T)$ is a measure preserving system, a random variable $X : \Omega \to \mathbb{R}$ is called **invariant** if $X \circ T \equiv X$ almost surely. Prove that a random variable is invariant if and only if it is measurable with respect to $\mathcal{I}$, and that a system is ergodic if and only the only invariant random variables are almost surely constant.*

Exercise 2.23. *Show that a measure preserving system $(\Omega, \mathcal{F}, \mathbf{P}, T)$ is ergodic if and only if the probability measure $\mathbf{P}$* cannot *be represented in the form*

$$\mathbf{P} = \alpha Q_1 + (1 - \alpha) Q_2,$$

where $0 < \alpha < 1$ and Q_1, Q_2 are two distinct T-invariant probability measures on the measurable space $(\Omega, \mathcal{F})$. (In words, this means that an ergodic system cannot be decomposed into a nontrivial convex combination of two simpler systems.)

To get a feel for this new concept, let us examine which of the measure preserving systems discussed in the previous section are ergodic.

1. **i.i.d. sequence.** Let A be an invariant event in the i.i.d. shift. A is in the product σ-algebra, in other words, it is measurable with respect to $\sigma(X_1, X_2, \ldots)$, where $X_1, X_2, \ldots$ denote the coordinate functions of the product space. Then

$$S^{-1}(A) = \{\omega \in \mathbb{R}^{\mathbb{N}} : (\omega_2, \omega_3, \ldots) \in A\}$$

is measurable with respect to $\sigma(X_2, X_3, \ldots)$, and similarly, for any $n \geq 1$,

$$S^{-n}(A) = \{\omega \in \mathbb{R}^{\mathbb{N}} : (\omega_{n+1}, \omega_{n+2}, \ldots) \in A\}$$

is in $\sigma(X_{n+1}, X_{n+2}, \ldots)$. It follows that

$$A' = \bigcap_{N=1}^{\infty} \bigcup_{n=N}^{\infty} S^{-n}(A) = \{S^{-n}(A) \text{ i.o.}\}$$

is a tail event, and hence has probability 0 or 1 by the Kolmogorov 0-1 law. But we assumed that A was invariant, which implies that $A = S^{-n}(A)$ almost surely for all $n \geq 1$, and therefore also $A = A'$ almost surely. It follows that A is also a 0-1 event[2]. We have proved:

Lemma 2.24. *Any i.i.d. shift map is ergodic.*

2. **A shift-equivariant function of an ergodic stationary sequence.**[3] Let $(X_n)_n$ be a stationary sequence whose associated shift system is ergodic (such a sequence is called simply a **stationary ergodic sequence**), and let Y_n be defined as in (3).

[2]The above argument shows that $\mathcal{I} \subseteq \mathcal{T}$ (the σ-algebra of invariant subsets is contained in the tail σ-algebra), as long as we identify sets which are a.s. equal.

[3]In more abstract treatments of ergodic theory, this example would be called a **factor map** or **homomorphism**. An important family of problems in ergodic theory is concerned with identifying when one measure preserving system can be obtained as a homomorphism of another (usually simpler) measure preserving system, and especially when one can find an *invertible* homomorphism, also known as an **isomorphism**, between the two systems.

Lemma 2.25. *The stationary sequence $(Y_n)_n$ is also ergodic.*

Proof. Let A be an invariant event for the (Y_n) sequence. We can think of A as "living" in the original product space $\mathbb{R}^{\mathbb{N}}$ associated with the shift map for the sequence $(X_n)_n$. (Formally, the sequence $(Y_n)_n$ is an infinite-dimensional random vector, i.e., it maps the $\mathbb{R}^{\mathbb{N}}$ "of" the $(X_n)_n$ sequence into a "different copy" of $\mathbb{R}^{\mathbb{N}}$; by pulling back the event A with respect to this mapping we get a "copy" of A in the original product space.) The fact that A is invariant under shifting the Y_n's means it is also invariant under the original shift of the X_n's, hence is a 0-1 event by the assumption that the $(X_n)_n$ sequence is ergodic. □

3. **Stationary finite-state Markov chains.** A Markov chain is called **irreducible** if any state can be reached in a sequence of steps from any other state. It is not hard to prove (see Example 6.1.6 on page 333 of [4]) that a stationary finite-state Markov chain is ergodic if and only if it is irreducible.

4. **Tossing a randomly chosen coin.** In this experiment we have

$$U = \lim_{n\to\infty} \frac{1}{n} \sum_{k=1}^{n} X_k$$

by the strong law of large numbers (conditioned on the value of U). So, the random coin bias U is an invariant random variable, and thus the sequence $(X_n)_n$ is ergodic if and only if U is a.s. constant (equivalently, if and only if the sequence is i.i.d.).

We should note that this process is in some sense an archetypical example of a non-ergodic process, in the sense that non-ergodicity is precisely the behavior in which the experiment chooses "at the dawn of time" some random data or information (represented by the σ-algebra of invariant events), and then performs a stationary ergodic sequence of experiments that depends on this initial data. In other words, a general stationary sequence can always be represented as a mixture, or a kind of weighted average, of stationary ergodic sequences, where the weights in the mixture correspond to the probability distribution of the initial data. (The precise formulation of this statement leads to the concept of the **ergodic decomposition** of a measure preserving system, which we will not discuss in detail since it requires some slightly advanced notions from functional analysis.)

Exercise 2.26. *Show that the σ-algebra $\mathcal{I}$ of invariant subsets for this process coincides with the σ-algebra $\sigma(U)$ generated by the random coin bias U. That means that, not only is U an invariant random variable, but any other invariant random variable can be computed once the value of U is known.*

5. **Pólya's urn.** Recall that in the Pólya's urn model discussed in Example 5 on page 13, it was mentioned in Exercise 2.12(c) (page 15) that the fraction of white balls in the urn converges almost surely to a limiting random variable Y. It is easy to see that this is random variable is invariant with respect to the shift map associated with the stationary sequence $(I_n)_{n=1}^{\infty}$ described in the discussion of the Pólya's urn model. By (8), Y is a beta-distributed random variable, so in particular it is non-constant. This shows that the measure preserving shift associated with the sequence $(I_n)_n$ is not ergodic.

   It is an amusing and rather counter-intuitive fact that the Pólya urn experiment is actually a special case of the "tossing a randomly chosen coin" family of examples discussed above. In fact, the "random coin bias" U is equal to the limiting fraction Y of white balls in the urn. To see this, note that by a short computation (6) can be massaged into the form

$$\mathbf{P}(I_1 = x_1, \ldots, I_n = x_n) = \frac{B(a+k, b+n-k)}{B(a,b)},$$

   where $B(u,v) = \int_0^1 x^{u-1}(1-x)^{v-1}\,dx$ denotes the Euler beta function, and $k = \sum_{j=1}^n x_j$ (check!). We can further recognize the quantity on the right hand side as an expectation, namely

$$\frac{B(a+k, b+n-k)}{B(a,b)} = \frac{1}{B(a,b)} \int_0^1 x^{a-1}(1-x)^{b-1} x^k (1-x)^{n-k}\,dx = \mathbf{E}(U^k(1-U)^{n-k}),$$

   where $U \sim \text{Beta}(a,b)$. Thus, we have the amazing fact that, in effect, Pólya's urn behaves as if at the beginning of time, it *chooses* the random limiting fraction Y of white balls (without telling the experimenter!), and subsequently tosses an i.i.d. sequence of coin tosses with bias Y to choose the successive colors of the balls that get added to the urn. Furthermore, by the exercise above, the σ-algebra of invariant subsets is the one generated by Y, so intuitively one can say that this random variable measures the precise extent of non-ergodicity in the process, i.e., the decomposition of the process into its ergodic components.

6. **Rotation of the circle.** The following result has a natural number theoretic interpretation, which we'll discuss later after proving the Birkhoff pointwise ergodic theorem.

   Theorem 2.27. *The circle rotation map R_α is ergodic if and only if α is irrational.*

   Proof. If $\alpha = p/q$ is rational, the set

   $$E = \left[0, \tfrac{1}{2q}\right] \cup \left[\tfrac{1}{q}, \tfrac{3}{2q}\right] \cup \left[\tfrac{2}{q}, \tfrac{5}{2q}\right] \cup \left[\tfrac{3}{q}, \tfrac{7}{2q}\right] \cup \ldots \cup \left[\tfrac{q-1}{q}, \tfrac{2q-1}{2q}\right]$$

   is an example of a nontrivial invariant set. Conversely, assume that α is irrational. Let $A \subset [0,1]$ be an invariant event. The indicator variable $\mathbf{1}_A$ is a bounded measurable function, hence an element of $L_2[0,1]$, and can therefore be expanded in the Fourier basis

   $$\mathbf{1}_A(x) = \sum_{n=-\infty}^{\infty} c_n e^{2\pi i n x}.$$

   (The equation represents an equality in L_2, i.e., it is true for almost every $x \in [0,1]$.) The coefficients c_n in the expansion are given by $c_n = \frac{1}{2\pi}\int_0^1 \mathbf{1}_A(x) e^{-2\pi i n x}\, dx$. Then we have

   $$\mathbf{1}_{R_\alpha^{-1}(A)}(x) = (\mathbf{1}_A \circ R_\alpha)(x) = \mathbf{1}_A(R_\alpha(x)) = \sum_{n=-\infty}^{\infty} c_n e^{2\pi i n R_\alpha(x)} = \sum_{n=-\infty}^{\infty} c_n e^{2\pi i n (x+\alpha \bmod 1)}$$
   $$= \sum_{n=-\infty}^{\infty} c_n e^{2\pi i n (x+\alpha)} = \sum_{n=-\infty}^{\infty} d_n e^{2\pi i n x},$$

   where we denote $d_n = c_n e^{2\pi i n \alpha}$. Since A is invariant, i.e., $\mathbf{1}_{R_\alpha^{-1}(A)} = \mathbf{1}_A$ a.s., we get that $c_n = d_n$ for all $n \in \mathbb{Z}$. But α is irrational, so $e^{2\pi i n \alpha} \neq 1$ if $n \neq 0$. It follows that $c_n = 0$ for all $n \neq 0$, which leaves only the constant Fourier coefficient, i.e., $\mathbf{1}_A \equiv c_0$ a.s., which proves that A is a trivial event. □

7. **The $2x$ mod 1 map.** As we discussed earlier, this system is equivalent to the $(\frac{1}{2}, \frac{1}{2})$ Bernoulli shift, so by Lemma 2.24 above, the doubling map is ergodic.

8. **The continued fraction map.** The Gauss map is ergodic, a fact which has important consequences (which we will discuss in the next chapter) for understanding the distribution of continued fraction quotients of a typical real number. There are many proofs of this result, see for example [1, 2].

9. **The $3x+1$ map.** K. R. Matthews and A. M. Watts [8] studied the extension $\tilde{T}$ of the $3x+1$ map to the 2-adic integers, and in particular proved that $\tilde{T}$ is ergodic. See also [6].

10. **Billiards.** In Chapter 5 we described some example of billiard systems which are known to be ergodic, and some which aren't (for relatively trivial reasons). In general it is extremely difficult to prove that a given billiard system is ergodic, but, similarly to the example of Markov chains described above, there is a kind of philosophical principle (that applies to billiard and other types of dynamical systems) that says that unless a system is non-ergodic for a relatively obvious or trivial reason (e.g., because there is some obvious quantity that is conserved such as the energy of a mechanical system), one would expect the system to be ergodic, even though in practice one may have no idea how to prove it in a given situation. As with any philosophical principle, one should take care in deciding how to apply it[4].

11. **Hamiltonian flow.** The situation is similar to that of billiard systems: most systems are assumed to be ergodic unless there are obvious reasons why they are not, but as far as I know this cannot be proved in virtually any example which has any real-world relevance.

12. **Geodesic flow.** Some geodesic flows are not ergodic (e.g., the sphere), and others are (for example, hyperbolic space). The main property required to have ergodicity is negative curvature, but I am not familiar with the specific details. It is also interesting to note that there is a beautiful theory linking the continued fraction map and other dynamical systems with a number-theoretic flavor to geodesic flows on compact hyperbolic surfaces (in the case of the continued fraction map, it can be related to the geodesic flow on the **modular surface** $\mathbb{H}/PSL(2,\mathbb{Z})$, the quotient of the hyperbolic plane by the modular group).

13. **The logistic map.** This map is ergodic, a fact that follows as a consequence of a much more general result proved in the paper [2].

[4]Note: a *philosophical principle* is what mathematicians invent when they can't say anything rigorous.

Chapter 3: Ergodic theorems

3.1 Von Neumann's L_2 ergodic theorem

Our first ergodic theorem is von Neumann's ergodic theorem, which in fact is a result in operator theory that has a nice interpretation for our problem of the convergence of ergodic averages in a measure preserving system.

Theorem 3.1 (Von Neumann's ergodic theorem.)**.** *Let H be a Hilbert space, and let U be a unitary operator on H. Let P be the orthogonal projection operator onto the subspace* $\mathrm{Ker}(U-I)$ *(the subspace of H consisting of U-invariant vectors). For any vector $v \in H$ we have*

$$\frac{1}{n}\sum_{k=0}^{n-1} U^k v \to Pv \quad \text{as } n \to \infty. \tag{9}$$

(Equivalently, the sequence of operators $\frac{1}{n}\sum_{k=0}^{n-1} U^k$ converges to P in the strong operator topology.)

Proof. Define two subspaces

$$V = \mathrm{Ker}(U-I) = \{v \in H \,:\, Uv = v\},$$
$$V' = \mathrm{Range}(U-I) = \{Uw - w \,:\, w \in H\}.$$

Note that (9) holds trivially for $v \in V$. For a different reason, we also show that it holds for $v \in V'$: if $v = Uw - w$ then we have

$$\frac{1}{n}\sum_{k=0}^{n-1} U^k v = \frac{1}{n}\left(U^n w - w\right) \to 0 \quad \text{as } n \to \infty.$$

On the other hand, one can verify that $v \in V^\perp$, and therefore $Pv = 0$, by observing that if $z \in V$ then

$$\langle w, z\rangle = \langle Uw, Uz\rangle = \langle Uw, z\rangle,$$

hence $\langle Uw - w, z\rangle = 0$.

Combining the above observations we see that (9) holds for $v \in V + V'$. Next, we claim that it also holds for $v \in \overline{V+V'}$, the norm closure of $V + V'$. Indeed, if $v \in \overline{V+V'}$ then

for an arbitrary $\epsilon > 0$ we can take $w \in V + V'$ such that $\|v - w\| < \epsilon$ and conclude that

$$\begin{aligned}\left\| \left(\frac{1}{n} \sum_{k=0}^{n-1} U^k - P \right) v \right\| &\leq \left\| \left(\frac{1}{n} \sum_{k=0}^{n-1} U^k - P \right) w \right\| + \left\| \left(\frac{1}{n} \sum_{k=0}^{n-1} U^k - P \right) (v - w) \right\| \\ &\leq \left\| \left(\frac{1}{n} \sum_{k=0}^{n-1} U^k - P \right) w \right\| + \epsilon.\end{aligned}$$

This implies that $\limsup_{n\to\infty} \left\| \left(\frac{1}{n} \sum_{k=0}^{n-1} U^k - P \right) v \right\| < \epsilon$, and since ϵ was an arbitrary positive number we get (9).

Finally, we claim that $H = \overline{V + V'}$. Since $\overline{V + V'}$ is a *closed* subspace of H, we have

$$\overline{V + V'} = \left(\left(\overline{V + V'} \right)^{\perp} \right)^{\perp},$$

(in general, the orthogonal complement of the orthogonal complement of a subspace W of a Hilbert space is equal to $\overline{W}$). So, it suffices to show that $\left(\overline{V + V'} \right)^{\perp} = \{0\}$, i.e., that the only vector orthogonal to all of $\overline{V + V'}$ is the zero vector. Assume w is such a vector. Then $w \perp Uw - w$. But note that we have the identity

$$\begin{aligned}\|Uw - w\|^2 &= \langle Uw - w, Uw - w \rangle = \|Uw\|^2 + \|w\|^2 - 2\,\mathrm{Re}\langle Uw, w \rangle \\ &= 2\|w\|^2 - 2\,\mathrm{Re}\langle Uw, w \rangle = -2\,\mathrm{Re}\langle Uw - w, w \rangle\end{aligned}$$

which means that $Uw - w = 0$, i.e., $w \in V$. Since $w \in \left(\overline{V + V'} \right)^{\perp}$ we get that w is orthogonal to itself and therefore $w = 0$, as claimed. □

Let $(\Omega, \mathcal{F}, \mathbf{P}, T)$ be a measure preserving system. We associate with T an operator U_T on the Hilbert space $L_2(\Omega)$, defined by

$$U_T(f) = f \circ T.$$

The fact that T is measure preserving implies that U_T is unitary:

$$\langle U_T f, U_T g \rangle = \mathbf{E}((U_T f)\overline{(U_T g)}) = \mathbf{E}((f \circ T)\overline{(g \circ T)}) = \mathbf{E}((f\overline{g}) \circ T) = \mathbf{E}(f\overline{g}) = \langle f, g \rangle.$$

Note also that the subspace $\mathrm{Ker}(U - I)$ consists exactly of the invariant (square-integrable) random variables, or equivalently those random variables which are measurable with respect to the σ-algebra $\mathcal{I}$ of invariant events. Recalling some of the standard properties of conditional expectations, we also see that the orthogonal projection operator P is exactly the conditional expectation operator $\mathbf{E}(\cdot \,|\, \mathcal{I})$ with respect to the σ-algebra of invariant events! Thus, Theorem 3.1 applied to this setting gives the following result.

Theorem 3.2 (The L_2 ergodic theorem)**.** *Let $(\Omega, \mathcal{F}, \mathbf{P}, T)$ be a measure preserving system. For any random variable $X \in L_2(\Omega, \mathcal{F}, \mathbf{P})$, we have*

$$\frac{1}{n}\sum_{k=0}^{n-1} X \circ T^k \to \mathbf{E}\,(X \,|\, \mathcal{I}) \quad \textit{in } L_2 \textit{ as } n \to \infty.$$

In particular, if the system is ergodic then

$$\frac{1}{n}\sum_{k=0}^{n-1} X \circ T^k \to \mathbf{E}(X) \quad \textit{in } L_2 \textit{ as } n \to \infty.$$

3.2 Birkhoff's pointwise ergodic theorem

We will now prove *the* fundamental result of ergodic theory, known alternately as **Birkhoff's pointwise ergodic theorem**; **Birkhoff's ergodic theorem**; the **pointwise ergodic theorem**; or just the **ergodic theorem**[5].

Theorem 3.3 (Birkhoff's ergodic theorem)**.** *Let $(\Omega, \mathcal{F}, \mathbf{P}, T)$ be a measure preserving system. Let $\mathcal{I}$ denote as usual the σ-algebra of T-invariant sets. For any random variable $X \in L_1(\Omega, \mathcal{F}, \mathbf{P})$, we have*

$$\frac{1}{n}\sum_{k=0}^{n-1} X \circ T^k \xrightarrow{a.s.} \mathbf{E}\,(X \,|\, \mathcal{I}) \quad \textit{as } n \to \infty. \tag{10}$$

When the system is ergodic, we have

$$\frac{1}{n}\sum_{k=0}^{n-1} X \circ T^k \xrightarrow{a.s.} \mathbf{E}(X) \quad \textit{as } n \to \infty. \tag{11}$$

For the proof, we start by proving a lemma, known as the **maximal ergodic inequality**.

Lemma 3.4. *With the same notation as above, denote also $S_0 = 0$, $S_n = \sum_{k=0}^{n-1} X \circ T^k$, and let $M_n = \max\{S_k \,:\, 0 \le k \le n\}$. For each $n \ge 1$ we have*

$$\mathbf{E}(X \mathbf{1}_{\{M_n > 0\}}) \ge 0.$$

[5]Incidentally, I've always found it strange that ergodic theory — unlike other areas of math — seems to be the only theory named after an adjective (as opposed to a noun, as in *the theory of numbers*, or as in the non-existent name *the theory of ergodicity*, which would perhaps have been a better name for ergodic theory). Similarly, the ergodic theorem is, as far as I know, the only theorem in math to be named after an adjective. (And what does the name mean, anyway? That the theorem has no nontrivial invariant sets...?) If you think of any counterexamples to this observation, please let me know!

Proof. For each $0 \le k \le n$ we have

$$S_{k+1} = X + S_k \circ T \le X + M_n \circ T,$$

or equivalently $X \ge S_{k+1} - M_n \circ T$. Since this is true for each $0 \le k \le n$, we get that

$$X \ge \max(S_1, \ldots, S_n) - M_n \circ T,$$

and therefore, noting that on the event $\{M_n > 0\}$, we have $M_n = \max(S_1, \ldots, S_n)$, we get that

$$\begin{aligned}
\mathbf{E}(X\mathbf{1}_{\{M_n>0\}}) &\ge \mathbf{E}\left[(\max(S_1, \ldots, S_n) - M_n \circ T)\mathbf{1}_{\{M_n>0\}}\right] \\
&= \mathbf{E}\left[(M_n - M_n \circ T)\mathbf{1}_{\{M_n>0\}}\right] \\
&= \mathbf{E}\left[(M_n - M_n \circ T)\right] - \mathbf{E}\left[(M_n - M_n \circ T)\mathbf{1}_{\{M_n>0\}^c}\right] \\
&= 0 - \mathbf{E}\left[(M_n - M_n \circ T)\mathbf{1}_{\{M_n=0\}}\right] \\
&= \mathbf{E}\left[(M_n \circ T)\mathbf{1}_{\{M_n=0\}}\right] \ge 0.
\end{aligned}$$

□

Proof of the ergodic theorem. $\mathbf{E}(X \,|\, \mathcal{I})$ is an invariant random variable, so by replacing X with $X - \mathbf{E}(X \,|\, \mathcal{I})$, we can assume without loss of generality that $\mathbf{E}(X \,|\, \mathcal{I}) = 0$; in this case, we need to prove that $S_n/n \to 0$ almost surely (where $S_n = \sum_{k=0}^{n-1} X \circ T^k$ as in the lemma above). Denote $\overline{X} = \limsup_{n\to\infty} S_n/n$. $\overline{X}$ is an invariant random variable, taking values in $\mathbb{R} \cup \{\pm\infty\}$. Fix $\epsilon > 0$, and consider the invariant event $A = \{\overline{X} > \epsilon\}$. We claim that $\mathbf{P}(A) = 0$. Once we prove this, since ϵ is arbitrary it will follow that $\overline{X} \le 0$ almost surely. By applying the same result to $-X$ instead of X the reverse inequality that almost surely $\liminf S_n/n \ge 0$ will also follow, and the theorem will be proved.

To prove the claim, define a new random variable $X^* = (X - \epsilon)\mathbf{1}_A$. Applying Lemma 3.4 to X^* we get that

$$\mathbf{E}\left[X^*\mathbf{1}_{\{M_n^*>0\}}\right] \ge 0,$$

where $M_n^* = \max(0, S_1^*, \ldots, S_n^*)$ and

$$S_k^* = \sum_{j=0}^{k-1} X^* \circ T^j = \sum_{j=0}^{k-1} \left((X - \epsilon) \circ T^{k-1}\right) \mathbf{1}_A = (S_k - k\epsilon)\mathbf{1}_A$$

(since A is an invariant event). Note that the events $\{M_n^* > 0\}$ are increasing, so $X^*\mathbf{1}_{\{M_n^*>0\}} \to X^*\mathbf{1}_B$ almost surely as $n \to \infty$, where the event B is defined by

$$B = \bigcup_{n=1}^{\infty} \{M_n^* > 0\} = \left\{ \sup_{n\geq 1} S_n^* > 0 \right\} = \left\{ \sup_{n\geq 1} S_n^*/n > 0 \right\}.$$

Furthermore, the convergence is dominated, since $\mathbf{E}|X^*| \leq \mathbf{E}|X| + \epsilon < \infty$, so the dominated convergence theorem implies that

$$\mathbf{E}(X^*\mathbf{1}_B) \geq 0.$$

Finally, observe that $A \subset B$, because

$$\begin{aligned} A &= \{\limsup_{n\to\infty} S_n/n > \epsilon\} \subseteq A \cap \{S_n > n\epsilon \text{ for some } n \geq 1\} \\ &= \bigcup_{n=1}^{\infty} \{(S_n - n\epsilon)\mathbf{1}_A > 0\} = \left\{ \sup_{n\geq 1} S_n^* > 0 \right\} = B. \end{aligned}$$

So we have shown that

$$\begin{aligned} 0 \leq \mathbf{E}(X^*\mathbf{1}_B) &= \mathbf{E}((X-\epsilon)\mathbf{1}_A\mathbf{1}_B) = \mathbf{E}((X-\epsilon)\mathbf{1}_{A\cap B}) = \mathbf{E}((X-\epsilon)\mathbf{1}_A) \\ &= \mathbf{E}(X\mathbf{1}_A) - \epsilon\mathbf{P}(A) = \mathbf{E}(\mathbf{E}\,(X\mathbf{1}_A \,|\, \mathcal{I})) - \epsilon\mathbf{P}(A) = \mathbf{E}(\mathbf{E}\,(X \,|\, \mathcal{I})\,\mathbf{1}_A) - \epsilon\mathbf{P}(A) = -\epsilon\mathbf{P}(A), \end{aligned}$$

which proves our claim that $\mathbf{P}(A) = 0$. □

3.3 The L_1 ergodic theorem

A trivial addendum to the previous proof shows that we also get L_1 convergence of the ergodic averages.

Theorem 3.5 (L_1 ergodic theorem)**.** *The convergence in* (10), (11) *is also in* L_1*.*

Proof. Fix $M > 0$, and write $X = Y_M + Z_M$ where $Y_M = X\mathbf{1}_{\{|X|\leq M\}}$ and $Z_M = X - Y_M = X\mathbf{1}_{\{|X|>M\}}$. The pointwise ergodic theorem implies that

$$\frac{1}{n}\sum_{k=0}^{n} Y_M \circ T^k \to \mathbf{E}\,(Y_M \,|\, \mathcal{I}) \quad \text{almost surely as } n \to \infty,$$

and since $|Y_M| \leq M$ the bounded convergence theorem implies also convergence in L_1, i.e.,

$$\left| \frac{1}{n}\sum_{k=0}^{n} Y_M \circ T^k - \mathbf{E}\,(Y_M \,|\, \mathcal{I}) \right| \to 0 \quad \text{as } n \to \infty. \tag{12}$$

Next, for Z_M we have the trivial estimates

$$\mathbf{E}\left|\frac{1}{n}\sum_{k=0}^{n-1} Z_M \circ T^k\right| \le \frac{1}{n}\sum_{k=0}^{n-1} \mathbf{E}|Z_M \circ T^k| = \mathbf{E}|Z_M|,$$
$$\mathbf{E}\,|\mathbf{E}\,(Z_M\,|\,\mathcal{I})| \le \mathbf{E}\Big[\mathbf{E}\,(|Z_M|\,|\,\mathcal{I})\Big] = \mathbf{E}|Z_M|,$$

so, combining this with (12), this shows that almost surely

$$\limsup_{n\to\infty} \mathbf{E}\left|\frac{1}{n}\sum_{k=0}^{n-1} X \circ T^k - \mathbf{E}\,(X\,|\,\mathcal{I})\right| \le 2\mathbf{E}|Z_M|.$$

Letting $M \to \infty$ finishes the proof, since $\limsup_{M\to\infty} \mathbf{E}|Z_M| = 0$ by the dominated convergence theorem. □

Exercise 3.6. *Prove that if X is in L_p for some $p > 1$ then the convergence in* (10) *is also in the L_p norm.*

3.4 Consequences of the ergodic theorem

In probability theory and many related fields, the ergodic theorem is an essential tool that is used frequently in concrete situations. Here are some of its consequences with regards to some of the examples we discussed before.

1. **The strong law of large numbers.** If $X_1, X_2, \ldots$ are i.i.d. with $\mathbf{E}|X_1| < \infty$, then if we think of the variables as being defined on the canonical product space $\mathbb{R}^{\mathbb{N}}$ (i.e., $X_n = \pi_n(\omega)$ is the nth coordinate function), then we have $X_n = X_1 \circ S^{n-1}$, where $S : \mathbb{R}^{\mathbb{N}} \to \mathbb{R}^{\mathbb{N}}$ is the shift map. Thus, the ergodic average $\frac{1}{n}\sum_{k=0}^{n-1} X_1 \circ S^k$ is the same as the familiar empirical average $\frac{1}{n}S_n = \frac{1}{n}\sum_{k=1}^{n} X_k$ for an i.i.d., sum, and Birkhoff's ergodic theorem implies the strong law of large numbers. (In fact, one can think of the ergodic theorem as a powerful and far-reaching generalization of the SLLN).

2. **Equidistribution of the fractional part of $n\alpha$.** A classical question in number theory concerns the statistical properties of the fractional part of the integer multiples of a number α, i.e., the sequence $\{n\alpha\}$ (sometimes written as $n\alpha$ mod 1), where $\{z\} = z - \lfloor z \rfloor$ denotes the fractional part of a real number z. If α is a rational number, it is easy to see that this sequence is periodic, and its range is the finite set of numbers

$\left\{\frac{k}{q} \,:\, k = 0, 1, \ldots, q-1\right\}$ (where q is the denominator in the representation of α as a reduced fraction p/q), so the question is trivial. In the case of irrational α something nice (though not too surprising, in hindsight) happens:

Theorem 3.7 (Equidistribution theorem)**.** *If $\alpha \in \mathbb{R} \setminus \mathbb{Q}$ then the sequence $(\{n\alpha\})_{n=1}^{\infty}$ is **equidistributed in** $[0,1]$[6]. More precisely, for any $0 < a < b < 1$ we have*

$$\frac{1}{n}\#\Big\{1 \le k \le n \,:\, \{n\alpha\} \in (a,b)\Big\} \to b - a \quad \text{as } n \to \infty.$$

To prove this, note that $\{n\alpha\}$ is simply $R_\alpha(0)$, where R_α is the circle rotation map discussed in previous sections. Since we proved that R_α is ergodic when α is irrational, the ergodic theorem implies that for almost every $x \in [0,1]$

$$\frac{1}{n}\#\Big\{1 \le k \le n \,:\, \{x+n\alpha\} \in (a,b)\Big\} = \frac{1}{n}\sum_{k=0}^{n-1}\left(\mathbf{1}_{(a,b)} \circ R_\alpha^k\right)(x) \to \int_0^1 \mathbf{1}_{(a,b)}(u)\,du = b-a$$

as $n \to \infty$. This would appear to be a weaker result, since it doesn't guarantee that the convergence occurs for the specific initial point $x = 0$. However, in the particular example of the irrational circle rotation map (and the particular observable of the form $\mathbf{1}_{(a,b)}$) a slightly unusual thing happens, which is that the ergodic theorem turns out to be true not just for almost every initial point x but for *all* x; in fact, it is easy to see that convergence for one value of x is equivalent to convergence for any other value of x (and in particular $x = 0$). This is left to the reader as an exercise.

Note. Theorem 3.7 was proved in 1909 and 1910 independently by Weyl, Sierpinski and Bohl. In 1916 Weyl showed that the sequence $\{n^2\alpha\}$ is equidistributed, and more generally that $\{p(n)\}$ is equidistributed if $p(x)$ is a polynomial with at least one irrational coefficient. Vinogradov proved in 1935 that if α is irrational then the sequence $\{p_n\alpha\}$ is equidistributed, where p_n is the nth prime number. Jean Bourgain (winner of a 1994 Fields Medal) proved similar statements in the more general setting of the pointwise ergodic theorem (i.e., the ergodic averages of the form $\frac{1}{n}\sum_{k=1}^{n} X \circ T^{k^2}$

[6]This is the terminology used in number theory — see for example the Wikipedia article `http://en.wikipedia.org/wiki/Equidistributed_sequence`. Note that in probability theory the word *equidistributed* means equal in distribution rather than uniformly distributed, so one should take care when using this term for the number theoretic meaning when talking to a probabilist.

and $\frac{1}{n}\sum_{k=1}^{n} X \circ T^{p_k}$ in a measure preserving system converge almost surely, under mild integrability conditions).

3. **Benford's law.** A beautiful variant of the circle rotation example above involves multiplication instead of addition (but one then has the luxury of multiplying by nice numbers such as rational numbers or integers, instead of adding irrational numbers). Consider for example the distribution of the first digit in the decimal expansion of the sequence of powers of 2, $(2^n)_{n=1}^{\infty}$. Should we expect all digits to appear equally frequently? No, a quick empirical test shows that small digits appear with higher frequency than large digits. To see why, note that this is related to the dynamical system $T : x \mapsto 2x \bmod (10^k)_{k=1}^{\infty}$ on the interval $[1, 10)$ (i.e., multiplication by 2 in the quotient group of all positive numbers with the multiplication operator quotiented by the cyclic group generated by the number 10). For example, starting from 1 and iterating the map we get the sequence

$$1 \mapsto 2 \mapsto 4 \mapsto 8 \mapsto 1.6 \mapsto 3.2 \mapsto 6.4 \mapsto 1.28 \mapsto \dots$$

It is easy to check that this map has the invariant measure

$$d\mu(x) = \frac{1}{\log 10}\frac{dx}{x} \qquad (0 < x < 1)$$

In fact, this is a thinly disguised version of the circle rotation map R_α with $\alpha = \log_{10} 2$; the two maps are conjugate by the mapping $C(x) = \log_{10} x$ (i.e., C maps $[1, 10)$ bijectively to $[0, 1)$ and the relation $T = C^{-1} \circ R_\alpha \circ C$ holds), and furthermore the measure μ defined above is the pull-back of Lebesgue measure on $[0, 1)$ with respect to the conjugation map C, which is why an experienced ergodicist will know immediately that μ is an invariant measure for T.[7]

With this setup, we can answer the question posed at the beginning of the example. For each $1 \leq d \leq 9$, the fraction of the first n powers of 2 whose decimal expansions

[7] We are skirting an important concept in ergodic theory here — in fact, the map C is an example of an **isomorphism** between two measure preserving systems. Isomorphisms play a central role in ergodic theory, and there's a lot more to say about them, but we will not go further into the subject due to lack of time.

start with a given digit d is given by

$$\frac{1}{n}\#\{0 \le k \le n-1 \,:\, T^k(1) \in [d, d+1)\} = \frac{1}{n}\sum_{k=0}^{n-1} \left(\mathbf{1}_{[d,d+1)} \circ T^k\right)(1)$$
$$\to \frac{1}{\log 10}\int_d^{d+1} \frac{dx}{x} = \log_{10}\frac{d+1}{d},$$

where the convergence follows by the equidistribution theorem (Theorem 3.7), the above comments and the exercise below. This probability distribution on the numbers $1, \ldots, 9$ is known as **Benford's law**. Note that the most common digit 1 appears more than 30% of the time, and the least frequent digit 9 only appears only 4.6% of the time.

Exercise 3.8. *Let $n < m$ be positive integers. Prove that if m is not an integer power of n then $\log_m n$ is an irrational number.*

Benford's law is indeed an amusing distribution. From the exercise it is apparent that the choice of 2 as the factor of multiplication of the dynamical system is not special, and any other number that is not a power of 10 will work. In fact, even this does not come close to describing the generality in which Benford's law holds empirically as the first-digit distribution of real-life datasets. The reason for this is the fact that the measure μ is invariant under all scaling transformations. Thus, one should expect to observe an approximation to Benford's law in any set of numbers which are more or less "scale-free", in the sense that the set contains samples that span a large number of orders of magnitude, and where the unit of measurement is arbitrary and not inherently tied to the data being measured. Examples include distances between points on a map, financial reports, heights of the world's tallest structures and many more; it has even been proposed in several studies that Benford's law can be applied to the problem of detecting tax evasion and various forms of financial fraud and possibly also election fraud. (Presumably, this will work under the assumption that the cheaters who fake financial and tax reports are themselves not aware of the importance of Benford's law!)

4. **Continued fractions.** The fact that the continued fraction map on $(0, 1)$ (together with the Gauss invariant measure) is ergodic has important consequences regarding the distribution of quotients in the continued fraction expansion of a number chosen

d	$\mathbb{P}_{\text{Benford}}(d) = \log_{10} \frac{d+1}{d}$	Graphical illustration
1	30.1%	
2	17.6%	
3	12.5%	
4	9.7%	
5	7.9%	
6	6.7%	
7	5.8%	
8	5.1%	
9	4.6%	

Table 1: The (approximate) digit frequencies in Benford's law

uniformly at random in $(0, 1)$. In contrast to the much more trivial case of the digits in a decimal (or base-b) expansion, which are simply i.i.d. random numbers chosen from $0, \ldots, 9$, the asymptotic distribution of successive continued fraction quotients is that they are identically distributed, but not quite independent. To see this, note that the marginal distribution of a single quotient can be computed using ergodic averages, as follows. For each $q \geq 1$, the set of numbers x whose first quotient $N(x) = \lfloor 1/x \rfloor$ is equal to q is exactly the interval $\left(\frac{1}{q+1}, \frac{1}{q}\right]$. Thus, for a given number x we can recover the proportion of the first n quotients equal to q as

$$\frac{1}{n} \sum_{k=0}^{n-1} \left(\mathbf{1}_{(1/(q+1),1/q]} \circ G^k\right)(x),$$

which by the ergodic theorem converges to

$$\frac{1}{\log 2} \int_{1/(q+1)}^{1/q} \frac{dx}{1+x} = \log_2 \left(\frac{(q+1)^2}{q(q+2)}\right) \tag{13}$$

for a set of x's that has measure 1 with respect to Gauss measure γ (and hence, also almost surely with respect to Lebesgue measure, since γ and Lebesgue are mutually absolutely continuous with respect to each other). Thus, the formula on the right-hand side of (13) (which is oddly reminiscent of Benford's law, though they are not related) represents the limiting distribution of the first quotient of a random number.

For example, the frequency of occurrence of the quotient 1 is $\log_2(4/3) \approx 41.5\%$ — more than 40% of the quotient are equal to 1! Note that this is an asymptotic result that pertains to the statistics of many quotients of a given number x, and not to the *first quotient of* x: if x is chosen uniformly in $[0,1]$, because Lebesgue measure is not invariant under the Gauss map G, the first quotient of x has a different distribution (clearly, the probability that the first quotient is q is exactly $1/q - 1/(q+1)$, the length of the interval $(1/(q+1), 1/q]$).

Exercise 3.9. *Compute the asymptotic probability that a pair of successive quotients of a randomly chosen* x *in* $[0,1]$ *is equal to* $(1,1)$ *and compare this to the square of the frequency of* 1*'s, to see why successive quotients are not independent of each other. Are two successive* 1*'s positively or negatively correlated?*

What other quantities of interest can one compute for the continued fraction expansion of random numbers? One can try computing the expected value of a quotient, but that turns out not to be very interesting — the average $\frac{1}{\log 2}\int_0^1 N(x)\frac{dx}{1+x}$ is infinite. The Russian probabilist Khinchin (known for his Law of Iterated Logarithm, a beautiful result on random walks and Brownian motion) derived an interesting limiting law for the *geometric* average of the quotients. He proved that for almost every $x \in [0,1]$, the geometric average $(q_1 \dots q_n)^{1/n}$ of the first n quotients of x converges to the constant

$$K = \prod_{k=1}^{\infty} \left(1 + \frac{1}{k(k+2)}\right)^{\log_2 k} \approx 2.68545$$

(known as **Khinchin's constant**).

Exercise 3.10. *Prove Khinchin's result.*

We mention one additional and very beautiful limiting result on continued fraction expansions. If $x \in (0,1)$ has an infinite continued fraction expansion

$$x = \cfrac{1}{n_1 + \cfrac{1}{n_2 + \cfrac{1}{n_3 + \cfrac{1}{n_4 + \cfrac{1}{\dots}}}}},$$

where $n_1, n_2, \ldots$ are the quotients in the expansion, it is interesting to consider the truncated expansion

$$\frac{P_k}{Q_k} = \cfrac{1}{n_1 + \cfrac{1}{n_2 + \cfrac{1}{n_3 + \cfrac{1}{\ddots + \cfrac{1}{n_k}}}}},$$

which are rational numbers that become better and better approximations to x. In fact, one reason why continued fraction expansions are so important in number theory is that it can be shown that the *best* rational approximation to x with denominator bounded by some bound N will always be the last truncated continued fraction P_k/Q_k for which $Q_k \leq N$, and furthermore, the inequalities

$$\frac{1}{Q_k(Q_k + Q_{k+1})} \leq \left| x - \frac{P_k}{Q_k} \right| \leq \frac{1}{Q_k Q_{k+1}} \tag{14}$$

hold. How fast should we expect this sequence of rational approximations to converge? The answer is given in the following theorem. For the proof (which is surprisingly not difficult), see [1, Sec. 1.4].

Theorem 3.11. *For almost every $x \in (0,1)$ we have*

$$\lim_{k\to\infty} \frac{1}{k} \log Q_k = \frac{\pi^2}{12 \log 2}, \tag{15}$$

$$\lim_{k\to\infty} \frac{1}{k} \log \left| x - \frac{P_k}{Q_k} \right| = \frac{\pi^2}{6 \log 2}, \tag{16}$$

$$\lim_{k\to\infty} \frac{1}{k} \log \mathrm{Leb}(\Delta_k(x)) = \frac{\pi^2}{6 \log 2}, \tag{17}$$

*where $\Delta_k(x) = \{y \in (0,1) : n_j(y) = n_j(x) \text{ for } 1 \leq j \leq k\}$ (this interval is sometimes called the k**th fundamental interval of** x), and $\mathrm{Leb}(\cdot)$ denotes Lebesgue measure (it is easy to see that the same statement is true if Gauss measure is used instead).*

Note that (16) and (17) follow easily by combining (15) with (14). The interesting constant $\frac{\pi^2}{6 \log 2}$ is sometimes referred to as the **entropy constant of the continued fraction map**.

Chapter 4: Entropy and information theory

4.1 Entropy and its basic properties

In this chapter we give an introduction to **information theory**, a beautiful theory at the intersection of ergodic theory, probability, statistics, computer science, and branches of engineering and physics.

Throughout the chapter, $X_1, X_2, X_3, \ldots$ will denote a stationary ergodic sequence of random variables taking values in a finite set $A = \{\alpha_1, \ldots, \alpha_d\}$. We think of the sequence as an **information source**, emitting successive symbols from the set A, which in this context will be referred to as the **alphabet**. Think of a long text in English or some other language[8]; a sequence of bits being transmitted from one computer to another over a network; data sampled by a scientific instrument over time, etc. — all of these are examples of information sources which in suitable circumstances are well-modeled by a stationary ergodic sequence over a finite alphabet.

A fundamental problem of information theory is to measure the information content of the source. This is a numerical quantity which has come to be known as **entropy**. We will define it and also try to explain what the number it gives means. E.g., if the entropy of a source is 3.5, what does that tell us regarding the difficulty of storing or communicating information coming from the source?

Let us start with the simplest case of an i.i.d. sequence. Denote $p_k = \mathbf{P}(X_1 = \alpha_k)$. The probability vector $(p_1, \ldots, p_d)$ gives the relative frequencies of occurrence of each of the symbols $\alpha_1, \ldots, \alpha_d$, and for an i.i.d. sequence completely characterizes the statistical properties of the sequence, so entropy will simply be a function of the numbers $p_1, \ldots, p_d$. We define it as

$$H(p_1, \ldots, p_d) = -\sum_{k=1}^{d} p_k \log_2(p_k),$$

[8]It may seem unusual to you that language is considered as a statistical source, but spoken and written language does exhibit very clear statistical characteristics. Note that information theory makes no attempt to address the *meaning* (or usefulness) of a string of text. Thus, the word "information" is used in a slightly different meaning in information theory versus how an ordinary person might use it. For example, a string of random unbiased binary bits might appear to contain very little information to a layperson, but in the information theory sense this kind of string has the highest possible information content for a binary string of given length.

with the convention that $0 \log 0 = 0$. The logarithm is traditionally taken to base 2, to reflect the importance of entropy in computer science and engineering, although in certain fields (notably thermodynamics and statistical physics) the natural base is used, and any other base may be used as long as it is used consistently in all formulas. If the base 2 is used, we say that entropy is measured in units of **bits**. The letter H used for the entropy function is actually a capital Greek *eta*, the first letter of the Greek word *entropia*[9].

Note that entropy can be regarded as the average of the quantity $-\log_2 p_k$ weighted by the probabilities p_k. Thus, sometimes we write

$$H(p_1, \ldots, p_k) = -\mathbf{E} \log_2 p(X),$$

where X is a random variable representing a source symbol (i.e., $\mathbf{P}(X = \alpha_k) = p_k$ for each $1 \leq k \leq d$, and $p(\alpha_k) = p_k$ represents the probability of each symbol. (It is a distinctive and somewhat curious feature of information theory that probabilities are often themselves regarded as random variables.)

Example 4.1. In the case of a 2-symbol alphabet ($d = 2$), the entropy function is usually written simply as a function of one variable, i.e.,

$$H(p) = -p \log_2 p - (1-p) \log_2 (1-p)$$

This function is concave, has the symmetry $H(p) = H(1-p)$, equal to 0 at $p = 0$ and $p = 1$, and takes the maximum value $H(1/2) = 1$ at $p = 1/2$ (see figure).

Lemma 4.2 (Gibbs's inequality)**.** *If $(p_1, \ldots, p_d)$ is a probability vector and $(q_1, \ldots, q_d)$ is a* sub-probability *vector, i.e., we have $p_k, q_k \geq 0$, $\sum_k p_k = 1$ and $\sum q_k \leq 1$, then*

$$-\sum_{i=1}^{d} p_i \log p_i \leq -\sum_{i=1}^{d} p_i \log q_i,$$

with equality holding if and only the two vectors are equal.

[9] See the Wikipedia article `http://en.wikipedia.org/wiki/History_of_entropy#Information_theory` for an amusing and often-told story about the origin of the term entropy in information theory.

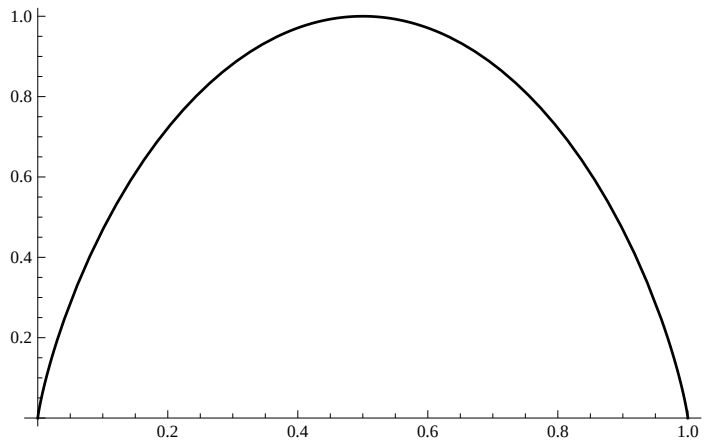


Figure 4: The entropy function $H(p)$ for a two-value distribution

Proof. The form of the inequality is unchanged by changing the logarithm basis, so we use the natural logarithm. Since $\log x \leq x - 1$ for all $x > 0$, we have

$$\begin{aligned} -\sum_{k=1}^{d} p_k(\log q_k - \log p_k) &= -\sum_{k=1}^{d} p_k \log\left(\frac{q_k}{p_k}\right) \geq -\sum_{k=1}^{d} p_k \left(\frac{q_k}{p_k} - 1\right) \\ &= -\sum_{k} q_k + \sum_{k} p_k \geq -1 + 1 = 0. \end{aligned}$$

□

Lemma 4.3 (Properties of the entropy function)**.** *The entropy function of d-dimensional probability vectors $(p_1, \ldots, p_d)$ satisfies:*

1. $0 \leq H(p_1, \ldots, p_d) \leq \log_2 d$
2. $H(p_1, \ldots, p_d) = 0$ *if and only if* $p_k = 1$ *for some* k *(and all the other* p_j*'s are 0).*
3. $H(p_1, \ldots, p_d) = \log_2 d$ *if and only if* $p_k = 1/d$ *for all* k*.*
4. $H(\mathbf{p} \otimes \mathbf{q}) = H(\mathbf{p}) + H(\mathbf{q})$*, where if* $\mathbf{p} = (p_1, \ldots, p_d)$ *and* $\mathbf{q} = (q_1, \ldots, q_\ell)$*, we use the notation* $\mathbf{p} \otimes \mathbf{q}$ *to denote the probability vector* $(p_i q_j)_{i,j}$ *on the product of two alphabets of sizes* d *and* ℓ*.*
5. $H(p_1, \ldots, p_d)$ *is a concave function.*

Exercise 4.4. *Prove Lemma 4.3.*

Exercise 4.5. *(a) Let $0 < p < 1$, and let $X_p \sim Geom(p)$ (a geometric r.v. with parameter p). Show that the entropy of X_p is given by*

$$H(X_p) = \frac{1}{p}H(p) = \frac{1}{p}\left(-p\log_2 p - (1-p)\log_2(1-p)\right).$$

Note. *The intuition for why we should expect such a formula to be true is that in order to sample from the $Geom(p)$ distribution, we toss a coin with bias p a random number of times that is on average $\mathbf{E}(X_p) = 1/p$. Each coin toss produces $H(p)$ bits of entropy, so the average amount of entropy produced is $\frac{1}{p}H(p)$. However, a calculation is still required to verify that this intuitive reasoning produces the correct answer; alternatively, as observed in the example on page 223 of [10], this fact can be seen to follow from a more conceptual result (see Corollary 1 on the preceding page of the same paper).*

(b) Let Y be a random variable taking values in the positive integers $\mathbb{N}$. Assume that $\mu := \mathbf{E}(Y) \in (1, \infty)$, and write $p = 1/\mu$. Prove that $H(Y) \leq H(X_p)$, where X_p is the geometric random variable defined above, with equality if and only if $Y \overset{\mathcal{D}}{=} X_p$.

Note. *This result gives a characterization of the geometric distribution with mean μ as the discrete distribution over the positive integers that has the maximal entropy, among all such distributions with a fixed mean — a conceptually important fact.*

4.2 The noiseless coding theorem

Our first interpretation of the entropy function will be in terms of the problem of **noiseless coding**. Recall that the information source emits symbols in the finite alphabet $A = \{\alpha_1, \ldots, \alpha_d\}$. To transmit the symbol over a digital communication channel or store it on a computer storage system (which is the same as transmission, except we're transmitting it to ourselves in the future rather than to a different physical location), we need to encode the symbols as binary bits. We assume that the storage system or communication system are **noiseless**, i.e., no corruption of our data is expected to occur.

What is a good way to encode the symbols as binary bits? A naive approach would be to allocate d distinct binary strings, one for each of the symbols. Since the strings need to be distinct so that the transmission can be decoded on the other end, obviously it is necessary

(and sufficient) for the strings to be of length $\lceil \log_2 d \rceil$. Thus, in terms of efficiency, this method uses the channel approximately $\log_2 d$ times for every symbol encoded.

But perhaps we can do better? For example, it is possible that some of the symbols occur more frequently than others. A more sophisticated approach would be to assign binary strings of *different* lengths to the different symbols, assigning the shorter strings to more frequently occurring symbols. One must be careful however to make sure that the transmission, which may consist of the concatenation of several of the strings used to encode a succession of source symbols, can be faithfully recovered. This leads to the idea of **codes**.

Definition 4.6. *Let $\{0,1\}^* = \cup_{n=1}^{\infty} \{0,1\}^n$ be the set of all finite binary sequences (which we will call **words** or **strings**). A **code** for the alphabet $A = \{\alpha_1, \ldots, \alpha_d\}$ is a collection $(w_k)_{k=1}^d$ of words in $\{0,1\}^*$. We say the code is **uniquely decodable** if any word formed as a concatenation $w_{j_1} w_{j_2} \ldots w_{j_m}$ of words in the code can be decoded in a unique way, i.e., it is not equal to any other concatenation of words from the same code. We say that the code is a **prefix code** if no word w_i in the code is a prefix of another word w_j.*

It is obvious that any prefix code is uniquely decodable, since, when reading a concatenation of words, we know immediately when a word terminates and the next word begins. Not all uniquely decodable codes are prefix codes, however (the code $0, 01, 011$ is an example). It is however true that uniquely decodable codes that are not prefix codes are in some sense pointless and for all practical purposes they may be ignored — see Corollary 4.11 at the end of this section to understand why. Prefix codes, on the other hand, are extremely useful in both theory and applications, and used by people (e.g., punctuation marks in language, the telephone directory), computers (innumerable examples) and even nature (the genetic code encodes amino acids used as building blocks of proteins as triplets of nucleotides in DNA).

For a word $w \in \{0,1\}^*$, denote its length by $\ell(w)$. Given a code $w_1, \ldots, w_d$ associated with an information source that emits a random symbol $\alpha_1, \ldots, \alpha_d$ with respective probabilities $p_1, \ldots, p_d$, denote by L the (random) word length, i.e., $L = \ell(w_k)$ with probability p_k for $k = 1, \ldots, d$. A crucial quantity that we are interested in is the **expected word length**

$$\mathbf{E}(L) = \sum_{k=1}^{d} p_k \ell(w_k).$$

By the law of large numbers, this quantity says how many bits we will need to transmit over the channel for every source symbol coded when encoding very long strings of source

symbols. How small can we make L? The following famous result answers this fundamental question.

Theorem 4.7 (Noiseless coding theorem)**.** *Let $(p_1, \ldots, p_d)$ be a probability vector. Then:*

1. *If $(w_1, \ldots, w_d) \in \{0,1\}^*$ is a prefix code for the source, then the expected word length satisfies*
$$\mathbf{E}(L) = \sum_{k=1}^{d} p_k \ell(w_k) \geq H(p_1, \ldots, p_d).$$
2. *A prefix code $(w_1, \ldots, w_d) \in \{0,1\}^*$ may be found for which the expected word length satisfies*
$$\mathbf{E}(L) = \sum_{k=1}^{d} p_k \ell(w_k) \leq H(p_1, \ldots, p_d) + 1.$$

To prove the theorem, we need an auxiliary result:

Theorem 4.8 (Kraft's inequality)**.**

1. *If $w_1, \ldots, w_d$ is a prefix code then $\sum_{k=1}^{d} 2^{-\ell(w_k)} \leq 1$.*
2. *Conversely, if $\ell_1, \ldots, \ell_d$ are positive integers satisfying $\sum_{k=1}^{d} 2^{-\ell_k} \leq 1$, then there exists a prefix code $w_1, \ldots, w_k$ with $\ell(w_k) = \ell_k$.*

Proof. For the first claim, for each word $w_k = a_1 \ldots a_{\ell(w_k)}$ define a real number x_k by

$$x_k = \sum_{j=1}^{\ell(w_k)} a_j 2^{-j} = (0.a_1 a_2 \ldots a_{\ell(w_k)})_{\text{binary}},$$

and consider the interval $I_k = [x_k, x_k + 2^{-\ell(w_k)})$ (a sub-interval of $[0,1]$; this is the set of numbers in $[0,1]$ whose binary expansion starts with the word w_k). The fact that the code is a prefix code is equivalent to the statement that the intervals I_k, $k = 1, \ldots, d$ are disjoint. It follows that $\sum_k |I_k| = \sum_k 2^{-\ell(w_k)} \leq 1$.

For the other direction, starting with the lengths $\ell_1, \ldots, \ell_d$, first assume without loss of generality that $\ell_1 \leq \ell_2 \leq \ldots \leq \ell_k$ (if not, relabel the indices). It is clear that we can inductively construct disjoint dyadic intervals $I_1, \ldots, I_k \subset [0,1]$ such that each I_k is of the form $[x_k, x_k + 2^{-\ell_k})$ where x_k has a binary expansion of length ℓ_k (take $x_1 = 0$ and let each

x_k for $k \geq 2$ be the rightmost endpoint of I_{k-1}; the construction will work because of the assumption that $\sum_{k=1}^{d} 2^{-\ell_k} \leq 1$, so the intervals never leave $[0,1]$, and the assumption that the lengths are increasing, which implies that the length of the binary expansion of x_k is at most ℓ_{k-1}). The code words $w_1, \ldots, w_k$ are then taken as the respective binary expansions of $x_1, \ldots, x_d$, where for each x_k, if the binary expansion is shorter than ℓ_k (as in the case of $x_1 = 0$), it is brought to the right length by padding it with zeros. □

Proof of the noiseless coding theorem. For the first part of the theorem, observe that since $\sum_{k=1}^{d} 2^{-\ell(w_k)} \leq 1$ by Kraft's inequality, we can apply Gibbs's inequality to the two vectors $(p_1, \ldots, p_k)$ and $(2^{-\ell(w_1)}, \ldots, 2^{-\ell(w_d)})$, to get that

$$\mathbf{E}(L) = \sum_{k=1}^{d} p_k \ell(w_k) = -\sum_{k=1}^{d} p_k \log_2 \left(2^{-\ell(w_k)}\right) \geq -\sum_{k=1}^{d} p_k \log_2 p_k = H(p_1, \ldots, p_d).$$

For the second part, for each $1 \leq k \leq d$ let $\ell_k = \lceil -\log_2 p_k \rceil$ (where $\lceil \cdot \rceil$ denotes the ceiling function), so that the inequality $2^{-\ell_k} \leq p_k < 2^{-\ell_k+1}$ holds. Then $\sum_k 2^{-\ell_k} \leq \sum_k p_k = 1$, so by Kraft's inequality we can find a prefix code $w_1, \ldots, w_k$ with word lengths $\ell_1, \ldots, \ell_k$. For this code, we have

$$\mathbf{E}(L) = \sum_{k=1}^{d} p_k \ell_k = -\sum_{k=1}^{d} p_k \log_2 \left(2^{-\ell_k}\right) \leq -\sum_{k=1}^{d} p_k \log_2(p_k/2) = H(p_1, \ldots, p_d) + 1.$$

□

While the noiseless coding theorem clearly indicates that the entropy $H = H(p_1, \ldots, p_d)$ is an interesting number, one might argue that the true minimal expected coding word length, which (by the theorem) lies somewhere in the interval $[H, H+1]$ (but which in practice may be hard to compute), is a more meaningful measure of the information content of a random symbol sampled from the distribution $p_1, \ldots, p_d$. For example, for a binary source information source with distribution $(p, 1-p)$ the "optimal expected word length" is exactly 1 bit per source symbol. However, in an asymptotic sense the entropy really is the more meaningful number; the trick is to cluster the source symbols into groups of fixed length and encode these longer strings, as the following reformulated version of the noiseless coding theorem explains.

Corollary 4.9 (Noiseless coding theorem, version 2)**.** *Let* $\mathbf{p} = (p_1, \ldots, p_d)$ *be a probability vector. Then:*

1. *Any prefix code for a source with distribution* $\mathbf{p}$ *has expected word length* $\geq H(\mathbf{p})$.

2. *For any* $\epsilon > 0$*, we can find an integer* N *large enough and a prefix code for a source with distribution* $\mathbf{p}^{\otimes N} = \mathbf{p} \otimes \ldots \otimes \mathbf{p}$ *(the distribution of a vector of* N *independent samples from* $\mathbf{p}$*) which has expected word length* $\leq N(H(\mathbf{p}) + \epsilon)$*; that is, the expected word length* per symbol coded *is at most* $H(\mathbf{p}) + \epsilon$.

Proof. For part 2, take $N = 1/\epsilon$ and apply the first version of the noiseless coding theorem to the distribution $\mathbf{p}^{\otimes N}$, making use of property 4 in Lemma 4.3. □

To summarize, this last formulation of the noiseless coding theorem gives a meaning to the entropy function as measuring precisely the difficulty of (noiselessly) coding the source, in an asymptotic sense: first, any code will require sending at least $H(p)$ binary bits over the communication channel; conversely, one can approach this lower bound asymptotically by coding for multiple symbols simultaneously.

The noiseless coding theorem solves the problem of the efficiency of coding over a noiseless channel using *prefix codes*. What about the more general class of uniquely decodable codes? The next lemma and the corollary that follows it show that the result is the same, and there is no real loss of generality in assuming that the codes we will be using are prefix codes.

Lemma 4.10 (Kraft's inequality, stronger version)**.** *If* $w_1, \ldots, w_d$ *is a uniquely decodable code then* $\sum_{k=1}^{d} 2^{-\ell(w_k)} \leq 1$.

Proof. For integers $k, m \geq 1$, let $A_{k,m}$ denote the number of ways in which m of the code words $w_1, \ldots, w_d$ can be concatenated to obtain a binary string of length k. Note that $A_{k,m}$ can be nonzero only if $k \leq \Lambda m$, where Λ denotes the maximal length of a code word. Moreover, the fact that the code is uniquely decodable implies that $A_{k,m} \leq 2^k$. Now observe that the definition of the numbers $A_{k,m}$'s can be expressed in an equaivalent algebraic form as an equality of generating functions. Specifically, we have the relation

$$\left(\sum_{j=1}^{d} x^{\ell(w_j)} \right)^m = \sum_{k \leq \Lambda m} A_{k,m} x^k.$$

In particular, setting $x = 1/2$ in this identity gives

$$\left(\sum_{j=1}^{d} 2^{-\ell(w_j)} \right)^m = \sum_{k \leq \Lambda m} A_{k,m} 2^{-k} \leq \sum_{k \leq \Lambda m} 2^k 2^{-k} = \Lambda m.$$

Taking mth roots, we obtain the inequality

$$\sum_{j=1}^{d} 2^{-\ell(w_j)} \leq (\Lambda m)^{1/m}.$$

Now take the limit as $m \to \infty$ to conclude that $\sum_{j=1}^{d} 2^{-\ell(w_j)} \leq 1$, as claimed. □

Corollary 4.11. *Any uniquely decodable code can be replaced by a prefix code with the same word lengths.*

4.3 The asymptotic equipartition property

A related way of thinking about entropy is in terms of data compression: given a string of source symbols of length n (which could itself be a binary string in the case of a 2-symbol alphabet), how much can we compress it, i.e., what is the typical length of a binary string we'll need to represent it? The noiseless coding theorem says that on the average we'll need around $nH(p_1, \ldots, p_d)$ bits; however, the theorem doesn't address the question of how many bits we'll need *typically* (that is, with probability close to 1)? Of course, these questions are in general not equivalent: for example, it may seem conceivable that the reason the average number of bits is around $nH(p_1, \ldots, p_d)$ is that around half the time we need a much smaller number of bits, and the other half of the time we need approximately twice as many. The following result, known as the **asymptotic equipartition property**, demonstrates that in fact in this case the typical behavior is the same as the average one.

Theorem 4.12 (Asymptotic equipartition property for an i.i.d. source)**.** *Let $X_1, X_2, \ldots$ be an i.i.d. information source over the alphabet $A = \{\alpha_1, \ldots, \alpha_d\}$, distributed according to the probability vector $\mathbf{p} = (p_1, \ldots, p_d)$ as before. Fix $\epsilon > 0$. There exists a large enough integer N such that the sequences A^N can be partitioned into a disjoint union of sequences of two types, namely,*

$$A^N = T \sqcup E,$$

*where the sequences in T and E are called the **typical** and **exceptional** sequences, respectively, such that the following properties hold:*

1. $\mathbf{P}((X_1, \ldots, X_N) \in E) < \epsilon$, *(i.e., the exceptional sequences are indeed exceptional).*

2. The probability of observing each typical sequence $(x_1, \ldots, x_N) \in T$ satisfies

$$2^{-N(H(\mathbf{p})+\epsilon)} \leq \mathbf{P}((X_1, \ldots, X_N) = (x_1, \ldots, x_N)) \leq 2^{-N(H(\mathbf{p})-\epsilon)}. \tag{18}$$

3. Consequently, assuming $\epsilon < 1/2$, the number of typical sequences satisfies

$$2^{N(H(\mathbf{p})-\epsilon)+1} \leq |T| \leq 2^{N(H(\mathbf{p})+\epsilon)}. \tag{19}$$

Proof. Define a sequence $Z_1, Z_2, \ldots$ of i.i.d. random variables by

$$Z_n = -\sum_{k=1}^{d} \log_2 p_k \mathbf{1}_{\{X_n = \alpha_k\}},$$

and denote $S_n = \sum_{j=1}^{n} Z_j$. By the weak law of large numbers we have that

$$\frac{1}{n} S_n \xrightarrow[n\to\infty]{\mathrm{P}} \mathbf{E}(Z_1) = H(\mathbf{p}),$$

and therefore for large enough N, we have

$$\mathbf{P}\left(\left|\frac{1}{N} S_N - H(\mathbf{p})\right| \leq \epsilon\right) \geq 1 - \epsilon. \tag{20}$$

Call the event on the left-hand side B. This is an event that depends on the r.v.'s $X_1, \ldots, X_N$, so it can be represented as a disjoint union of events of the form

$$B = \bigsqcup_{(x_1,\ldots,x_n)\in T} \{(X_1, \ldots, X_N) = (x_1, \ldots, x_n)\}$$

for some set $T \subset A^N$ of sequences. This will be our set of typical sequences; the exceptional sequences are defined as the complementary set $E = A^N \setminus T$.

We now claim that T and E satisfy the properties in the theorem. Property 1 holds automatically by (20). For property 2, observe that if $(x_1, \ldots, x_N) = (\alpha_{j_1}, \ldots, \alpha_{j_N}) \in T$ then by the definition of the event B we have

$$N\left(H(\mathbf{p}) - \epsilon\right) \leq -\sum_{n=1}^{N} \log_2 p_{j_n} \leq N\left(H(\mathbf{p}) + \epsilon\right),$$

or equivalently

$$2^{-N(H(\mathbf{p})+\epsilon)} \leq \prod_{n=1}^{N} p_{j_n} \leq 2^{-N(H(\mathbf{p})-\epsilon)}.$$

But $\prod_{n=1}^{N} p_{j_n}$ is exactly $\mathbf{P}((X_1, \ldots, X_N) = (x_1, \ldots, x_N))$, so we get (18). On the other hand, the total probability of observing *any* typical sequence is $\mathbf{P}(B)$, which is bounded between $1 - \epsilon$ and 1 (hence, between 1/2 and 1, if we assume $\epsilon < 1/2$). This implies (19). □

The implication of the theorem is that since the number of typical sequences is around $2^{n(H(\mathbf{p}) \pm \epsilon)}$, we can encode them using a binary string of length $\approx nH(\mathbf{p})$. How easy this is to do in practice is a different question (some very practical techniques exist that are not difficult to implement — for example, two well-known methods are known as Huffman coding and Lempel-Ziv coding).

Exercise 4.13. *Use the asymptotic equipartition property to give an alternate proof of the reformulated version of the noiseless coding theorem.*

4.4 Ergodic sources and the Shannon-McMillan-Breiman theorem

We are now ready to discuss the situation for a general stationary ergodic source $X_1, X_2, \ldots$. It turns out that a version of the asymptotic equipartition property is valid for such a source. To prove it, we first need to correctly define the entropy of the source, and to prove an important convergence result that replaces the (trivial) use of the law of large numbers in the case of an i.i.d. source.

For a sequence $(x_1, \ldots, x_n) \in A^n$, denote

$$p(x_1, \ldots, x_n) = \mathbf{P}((X_1, \ldots, X_n) = (x_1, \ldots, x_n)), \tag{21}$$

$$p(x_n \,|\, x_1, \ldots, x_{n-1}) = \mathbf{P}(X_n = x_n \,|\, (X_1, \ldots, X_{n-1}) = (x_1, \ldots, x_{n-1})), \tag{22}$$

$$H_n = -\mathbf{E}(\log_2 p(X_n \,|\, X_1, \ldots, X_{n-1})). \tag{23}$$

In information theory the quantity H_n is often denoted by $H(X_n \,|\, X_1, \ldots, X_{n-1})$; it is a special case of a **conditional entropy**.

Lemma 4.14. *$(H_n)_{n=1}^{\infty}$ is a weakly monotone decreasing sequence, hence converges to a limit*

$$H \equiv \lim_{n \to \infty} H_n \geq 0. \tag{24}$$

The proof follows by induction by applying the result of the following exercise.

Exercise 4.15. *Let $A = \{\alpha_1, \ldots, \alpha_d\}$ and $B = \{\beta_1, \ldots, \beta_m\}$ be two finite sets. If X, Y are two random variables such that $\mathbf{P}(X \in A, Y \in B) = 1$, the conditional entropy $H(X \,|\, Y)$ is defined by*

$$\begin{aligned} H(X \,|\, Y) &= -\sum_{j=1}^{m} \sum_{k=1}^{d} \mathbf{P}(X = \alpha_k, Y = \beta_j) \log_2 \mathbf{P}(X = \alpha_k \,|\, Y = \beta_j) \\ &= \sum_{j=1}^{m} \mathbf{P}(Y = \beta_j) H(X \,|\, Y = \beta_j). \end{aligned}$$

I.e., $H(X \,|\, Y)$ is the average of the entropies of the conditional distributions of X given the outcome of Y. Prove that $H(X \,|\, Y) \leq H(X)$, with equality if and only if X and Y are independent. Deduce also that $H(X \,|\, Y, Z) \leq H(X \,|\, Z)$ if Z is another random variable, and explain why this implies Lemma 4.14.

We refer to H in (24) as **the entropy of the source** $(X_n)_{n=1}^{\infty}$. There is an equivalent way to define it which is also interesting. Since $H_n \to H$, the Cesàro averages of $(H_n)_n$ also converge to H, i.e.,

$$\frac{1}{n}(H_1 + \ldots + H_n) \to H \text{ as } n \to \infty.$$

The average on the left-hand side can be written as

$$\begin{aligned} &-\frac{1}{n}\mathbf{E}\Big[\log_2 p(X_1) + \log_2 p(X_2 \,|\, X_1) + \log_2 p(X_3 \,|\, X_1, X_2) + \ldots + p(X_n \,|\, X_1, \ldots, X_{n-1}) \Big] \\ &\qquad = -\frac{1}{n}\mathbf{E}\left[\log_2 p(X_1, \ldots, X_n)\right] = \frac{1}{n} H(X_1, \ldots, X_n). \end{aligned}$$

(Here, $H(X_1, \ldots, X_n)$ refers to the entropy of the discrete vector random variable $(X_1, \ldots, X_n)$, which takes values in the finite set A^n.) Thus, H may be interpreted as the limit of $\frac{1}{n} H(X_1, \ldots, X_n)$, i.e., the asymptotic entropy per symbol in a long string of symbols sampled from the source.

The importance of H is explained by the following fundamental result, sometimes referred to as "the individual ergodic theorem of information theory".

Theorem 4.16 (Shannon-McMillan-Breiman theorem)**.** *We have the almost sure convergence*

$$-\frac{1}{n} \log_2(p(X_1, \ldots, X_n)) \xrightarrow[n \to \infty]{a.s.} H \tag{25}$$

Lemma 4.17. *If $(Z_n)_n$ is a sequence of nonnegative random variables such that $\mathbf{E}(Z_n) \leq 1$ for all n, then*

$$\mathbf{P}\left(\limsup_{n\to\infty} \frac{1}{n} \log Z_n \leq 0\right) = 1. \tag{26}$$

Proof. Fix $\epsilon > 0$. By Markov's inequality, we have

$$\mathbf{P}(n^{-1} \log Z_n \geq \epsilon) = \mathbf{P}(Z_n \geq e^{n\epsilon}) \leq e^{-n\epsilon}.$$

Since $\sum_n e^{-n\epsilon} < \infty$, the first Borel-Cantelli implies that $\mathbf{P}(n^{-1} \log Z_n \geq \epsilon \text{ i.o.}) = 0$. This is true for any $\epsilon > 0$, so taking a union of these events over $\epsilon = 1/k, k = 1, 2, \ldots$ gives (26). □

Proof of Theorem 4.16. As explained in Section 2.2, we may assume without loss of generality that the sequence $(X_n)_n$ is actually a two-sided ergodic stationary sequence $(X_n)_{n=-\infty}^{\infty}$. We start by giving yet another, more subtle, interpretation of the source entropy H. By stationarity, we may rewrite H_n as

$$H_n = -\mathbf{E}(\log_2 p(X_0 \,|\, X_{-n+1}, \ldots, X_{-1})) = -\sum_{j=1}^{d} \mathbf{E}\Big[L\big(\mathbf{E}\left(\mathbf{1}_{\{X_0=\alpha_j\}} \,|\, \mathcal{G}_{-n+1}^{-1}\right)\big)\Big],$$

where we denote $L(p) = p \log_2 p$ and $\mathcal{G}_s^t = \sigma(X_m\,;\, s \leq m \leq t)$. Note that for each j, the expression $\mathbf{E}\left(\mathbf{1}_{\{X_0=\alpha_j\}} \,|\, \mathcal{G}_{-n+1}^{-1}\right)\big)$ inside the conditional expectation above forms a martingale (as a function of n) taking values in $[0, 1]$. By Lévy's martingale convergence theorem (Theorem 3.27 in [11]), we have

$$\mathbf{E}\left(\mathbf{1}_{\{X_0=\alpha_j\}} \,|\, \mathcal{G}_{-n+1}^{-1}\right)\big) \to \mathbf{E}\left(\mathbf{1}_{\{X_0=\alpha_j\}} \,|\, \mathcal{G}_{-\infty}^{-1}\right) \quad \text{a.s. as } n \to \infty.$$

Since $L(\cdot)$ is a bounded continuous function on $[0, 1]$, using the bounded convergence theorem we therefore get also that

$$H_n \xrightarrow[n\to\infty]{} \mathbf{E}\left[-\sum_{j=1}^{d} \mathbf{E}\left(\mathbf{1}_{\{X_0=\alpha_j\}} \,|\, \mathcal{G}_{-\infty}^{-1}\right) \log_2 \mathbf{E}\left(\mathbf{1}_{\{X_0=\alpha_j\}} \,|\, \mathcal{G}_{-\infty}^{-1}\right)\right]$$

Of course, the limit of H_n is H, so we have derived another formula

$$H = \mathbf{E}\left[-\sum_{j=1}^{d} \mathbf{E}\left(\mathbf{1}_{\{X_0=\alpha_j\}} \,|\, \mathcal{G}_{-\infty}^{-1}\right) \log_2 \mathbf{E}\left(\mathbf{1}_{\{X_0=\alpha_j\}} \,|\, \mathcal{G}_{-\infty}^{-1}\right)\right] \tag{27}$$

for the source entropy. Furthermore, this expression can be rewritten in the simpler form

$$H = -\mathbf{E} \log_2 p(X_0 \,|\, \mathcal{G}_{-\infty}^{-1}), \tag{28}$$

where we adopt the notation (in the same vein as (21) and (22))

$$p(x \,|\, \mathcal{G}_s^t) = \mathbf{P}\left(X_{t+1} = x \,|\, \mathcal{G}_s^t\right). \tag{29}$$

To see why, note that

$$p(X_0 \,|\, \mathcal{G}_{-\infty}^{-1}) = \sum_{j=1}^{d} \mathbf{1}_{\{X_0 = \alpha_j\}} p(\alpha_j \,|\, \mathcal{G}_{-\infty}^{-1}),$$

and use this to write the right-hand side of (28) as

$$\begin{aligned}
-\mathbf{E} \log_2 p(X_0 \,|\, \mathcal{G}_{-\infty}^{-1}) &= -\sum_{j=1}^{d} \mathbf{E}\Big[\mathbf{E}\left(\mathbf{1}_{\{X_0=\alpha_j\}} \log_2 p(\alpha_j \,|\, \mathcal{G}_{-\infty}^{-1}) \,|\, \mathcal{G}_{-\infty}^{-1}\right)\Big] \\
&= -\sum_{j=1}^{d} \mathbf{E}\Big[\log_2 p(\alpha_j \,|\, \mathcal{G}_{-\infty}^{-1}) \mathbf{E}\left(\mathbf{1}_{\{X_0=\alpha_j\}} \,|\, \mathcal{G}_{-\infty}^{-1}\right)\Big] \\
&= -\sum_{j=1}^{d} \mathbf{E}\Big[p(\alpha_j \,|\, \mathcal{G}_{-\infty}^{-1}) \log_2 p(\alpha_j \,|\, \mathcal{G}_{-\infty}^{-1})\Big],
\end{aligned}$$

which is the same as the right-hand side of (27).

Having derived the representation (28) for the source entropy, we now apply another piece of heavy machinery, the ergodic theorem, which implies that

$$-\frac{1}{n} \sum_{k=0}^{n-1} \log_2 p(X_k \,|\, \mathcal{G}_{-\infty}^{k-1}) \xrightarrow[n\to\infty]{\text{a.s.}} H.$$

Furthermore, one may verify without much difficulty that this ergodic average can be rewritten in the form

$$-\frac{1}{n} \sum_{k=0}^{n-1} \log_2 p(X_k \,|\, \mathcal{G}_{-\infty}^{k-1}) = -\frac{1}{n} \log_2 p(X_0, \ldots, X_{n-1} \,|\, \mathcal{G}_{-\infty}^{-1}). \tag{30}$$

(where the notation $p(x_0, \ldots, x_{n-1} \,|\, \mathcal{G}_s^t)$ is defined as an obvious generalization of (29)). So we conclude that

$$-\frac{1}{n} \log_2 p(X_0, \ldots, X_{n-1} \,|\, \mathcal{G}_{-\infty}^{-1}) \to H \quad \text{a.s. as } n \to \infty.$$

This fact bears some resemblance to the claim (25) that we are trying to prove, and indeed, we can deduce "half" of our result from it — a one-sided asymptotic bound — using Lemma 4.17, as follows. Define a sequence of random variables $(Z_n)_{n=1}^{\infty}$ by $Z_n = \frac{p(X_0,\ldots,X_{n-1})}{p(X_0,\ldots,X_{n-1}\,|\,\mathcal{G}_{-\infty}^{-1})}$. We have

$$\begin{aligned}
\mathbf{E}(Z_n) &= \mathbf{E}\left[\mathbf{E}\left(Z_n \,|\, \mathcal{G}_{-\infty}^{-1}\right)\right] \\
&= \mathbf{E}\left[\sum_{x_0,\ldots,x_{n-1}\in A} \mathbf{E}\left(\frac{p(x_0,\ldots,x_{n-1})}{p(x_0,\ldots,x_{n-1}\,|\,\mathcal{G}_{-\infty}^{-1})} p(x_0,\ldots,x_{n-1}\,|\,\mathcal{G}_{-\infty}^{-1}) \,|\, \mathcal{G}_{-\infty}^{-1}\right)\right] \\
&= \sum_{x_0,\ldots,x_{n-1}\in A} p(x_0,\ldots,x_{n-1}) = 1. \qquad (31)
\end{aligned}$$

So, we are in the situation described in Lemma 4.17, and we conclude that almost surely we have the inequality

$$\begin{aligned}
0 \le -\limsup_{n\to\infty} \frac{1}{n}\log_2 Z_n &= \liminf_{n\to\infty}\left(-\frac{1}{n}\log_2 Z_n\right) \\
&= \liminf_{n\to\infty}\left[-\frac{1}{n}\log_2 p(X_0,\ldots,X_{n-1}) + \frac{1}{n}\log_2 p(X_0,\ldots,X_{n-1}\,|\,\mathcal{G}_{-\infty}^{-1})\right] \\
&= \liminf_{n\to\infty}\left[-\frac{1}{n}\log_2 p(X_0,\ldots,X_{n-1})\right] + \lim_{n\to\infty}\frac{1}{n}\log_2 p(X_0,\ldots,X_{n-1}\,|\,\mathcal{G}_{-\infty}^{-1}) \\
&= \liminf_{n\to\infty}\left[-\frac{1}{n}\log_2 p(X_0,\ldots,X_{n-1})\right] - H.
\end{aligned}$$

That is, we have proved that the inequality

$$\liminf_{n\to\infty}\left[-\frac{1}{n}\log_2 p(X_0,\ldots,X_{n-1})\right] \ge H. \qquad (32)$$

holds with probability 1.

To finish the proof, we will now prove an asymptotically matching upper bound; more precisely, we claim that for each fixed $k \ge 1$, almost surely the inequality

$$-\limsup_{n\to\infty}\frac{1}{n}\log_2 p(X_0,\ldots,X_{n-1}) \le H_k \qquad (33)$$

holds. Since $H_k \searrow H$, the inequalities (32) and (33) together imply (25). To this end, for each $k \ge 1$ we define the "kth order Markov approximation" to the function $p(x_1\ldots,x_n)$ by

$$\begin{aligned}
p_k(x_1,\ldots,x_n) &= p(x_1,\ldots,x_k)p(x_{k+1}\,|\,x_1,\ldots,x_k)p(x_{k+2}\,|\,x_2,\ldots,x_{k+1})\cdots p(x_n\,|\,x_{n-k},\ldots,x_{n-1}) \\
&= p(x_1,\ldots,x_k)\prod_{j=k+1}^{n} p(x_j\,|\,x_{j-k},\ldots,x_{j-1}) \qquad (n\ge k).
\end{aligned}$$

The idea in this definition is that $p_k(x_1, \ldots, x_k)$ is the symbol distribution of a modified source process $(X_n^{(k)})_{n=1}^{\infty}$ in which the conditional distribution of observing a symbol x_n given the past symbols $x_1, \ldots, x_{n-1}$ is computed from the symbol distribution of the original process by "forgetting" all the symbols before x_{n-k}, i.e., using only the information in the past k symbols. This modified source is a generalized type of Markov chain known as a **Markov chain of order** k or **Markov chain with memory** k.

Now observe that we have an expansion analogous to (30), namely

$$-\frac{1}{n}\log_2 p_k(X_0, \ldots, X_{n-1}) = -\frac{1}{n}\log_2 p(X_0, \ldots, X_{k-1}) - \frac{1}{n}\sum_{j=k}^{n-1} \log_2 p(X_j \mid X_{j-k}, \ldots, X_{j-1}).$$

Combining it with an application of the ergodic theorem, we deduce that

$$-\frac{1}{n}\log_2 p_k(X_0, \ldots, X_{n-1}) \xrightarrow[n\to\infty]{\text{a.s.}} -\mathbf{E}\big(\log_2 p(X_k \mid X_0, \ldots, X_{k-1})\big) = H_k.$$

This again bears a resemblance to (25), and we can relate the two using the lemma. Define a sequence $(Y_n)_{n=k}^{\infty}$ of random variables by $Y_n = \frac{p_k(X_0,\ldots,X_{n-1})}{p(X_0,\ldots,X_{n-1})}$. A short computation similar to (31), which we leave to the reader to verify, shows that $\mathbf{E}(Y_n) \le 1$ for all $n \ge k$, so from Lemma 4.17 we get that almost surely,

$$\begin{aligned} 0 &\ge \limsup_{n\to\infty} \frac{1}{n}\log_2 Y_n \\ &= \limsup_{n\to\infty} \left[\frac{1}{n}\log_2 p_k(X_0, \ldots, X_{n-1}) - \frac{1}{n}\log_2 p(X_0, \ldots, X_{n-1})\right] \\ &= -H_k + \limsup_{n\to\infty} \left(-\frac{1}{n}\log_2 p(X_0, \ldots, X_{n-1})\right), \end{aligned}$$

which proves (33) and thus finishes the proof. □

Theorem 4.18 (Asymptotic equipartition property for an ergodic source)**.** *Let $X_1, X_2, \ldots$ be a stationary ergodic information source over the alphabet $A = \{\alpha_1, \ldots, \alpha_d\}$. Fix $\epsilon > 0$. There exists a large enough integer N such that the sequences A^N can be partitioned into a disjoint union of typical and exceptional sequences, namely, $A^N = T \sqcup E$, such that we have:*

1. $\mathbf{P}((X_1, \ldots, X_N) \in E) < \epsilon$.
2. $2^{-N(H+\epsilon)} \le \mathbf{P}((X_1, \ldots, X_N) = (x_1, \ldots, x_N)) \le 2^{-N(H-\epsilon)}$ *for each typical sequence* $(x_1, \ldots, x_N) \in T$.

3. Consequently, assuming $\epsilon < 1/2$, the number of typical sequences satisfies

$$2^{N(H-\epsilon)+1} \leq |T| \leq 2^{N(H+\epsilon)}.$$

Proof. The proof is completely analogous to the proof of the i.i.d. case from the previous section; the random variable S_n is redefined as $-\log p(X_1, \ldots, X_n)$, and the use of the weak law of large numbers is replaced by the Shannon-McMillan-Breiman theorem. □

We conclude this chapter with some examples of stationary ergodic sequences and their entropies.

1. **i.i.d. source.** If $X_1, X_2, \ldots$ is an i.i.d. source whose distribution is described by the probability vector $(p_1, \ldots, p_d)$ then $H_n = H(X_n \,|\, X_1, \ldots, X_{n-1}) = H(p_1, \ldots, p_d)$, so the entropy is the usual entropy we discussed before. For example, if X is a Bernoulli random variable satisfying $\mathbf{P}(X = 1) = 1/3 = 1 - \mathbf{P}(X = 0)$ then $H = -\frac{1}{3}\log_2 \frac{1}{3} - \frac{2}{3}\log_2 \frac{2}{3} = 0.91829\ldots$ bits.

2. **Markov source.** If $X_1, X_2, \ldots$ is a stationary Markov chain, then the "n-step" conditional entropy H_n is given by $H_n = H(X_n \,|\, X_1, \ldots, X_{n-1}) = H(X_n \,|\, X_{n-1}) = H(X_2 \,|\, X_1)$ by the Markov property, so it is enough to compute this "1-step" conditional entropy. It is easy to see that this is simply an average with respect to the stationary probabilities of the entropies of each of the rows of the transition matrix. For example, if the Markov chain has the transition matrix $\begin{pmatrix} \frac{1}{2} & \frac{1}{2} \\ 1 & 0 \end{pmatrix}$, then it is easy to check that $(\frac{2}{3}, \frac{1}{3})$ is a stationary probability vector for the chain. The entropy is therefore given by

$$H = H(X_2 \,|\, X_1) = \tfrac{2}{3}H(\tfrac{1}{2}, \tfrac{1}{2}) + \tfrac{1}{3}H(1, 0) = \tfrac{2}{3} \cdot 1 = \tfrac{2}{3} = 0.6666\ldots \text{ bits.}$$

   Note that this stationary Markov chain has the same one-dimensional marginals as the i.i.d. source discussed above. Nonetheless, the entropy is lower, since it measures the incremental amount of information gained by examining a symbol once all the previous symbols are known, which is lower in the case where there is dependence.

3. **Continued fractions.** From the results discussed in the previous chapter, the entropy of the sequence of quotients $(N \circ G^k)_{k=0}^{\infty}$ in the continued fraction expansion of a number chosen according to Gauss measure γ is equal to $\pi^2/6\log 2$, *when measured in the natural*

base. If we want to adhere to the information theory convention and measure this entropy in bits, we must divide by a further factor of $\log 2$, giving an entropy of

$$\frac{\pi^2}{6(\log 2)^2} = 3.423714\ldots \text{ bits.}$$

One way of interpreting this fact is that, as we examine more and more of the continued fraction quotients of a number x chosen uniformly at random from $(0,1)$, on the average each additional quotients will increase our knowledge of the *binary* expansion of x by about 3.42 additional digits. Incidentally, while preparing these notes I discovered the curious fact (which I have not seen mentioned anywhere) that if we measure the entropy in base 10, we get

$$\frac{\pi^2}{6\log(2)\log(10)} = 1.03064\ldots,$$

i.e., on the average each continued fraction quotient adds an amount of information almost precisely equal to one decimal expansion digit.

4. **Rotations of the circle.** Let $\alpha \in (0,1)$ be irrational, let X be a random variable taking finitely many values on $((0,1), \mathcal{B}, \text{Leb})$, and let $X_n = X \circ R_\alpha^{n-1}$. Then $(X_n)_{n=1}^\infty$ is a stationary ergodic sequence.

   Exercise 4.19. *Prove that the entropy of this sequence is* 0.

Chapter 5: Brownian motion

Brownian motion started its life as a topic of scientific study as the physical phenomenon of the random motion of particles suspended in a fluid, famously observed through a microscope in 1827 by the botanist Robert Brown. Albert Einstein in 1905 showed through a brilliant analysis that this behavior can be explained as arising out of random collisions of the particle with molecules of the surrounding fluid. This explanation is considered an important early source of support for the atomic theory of matter.

Separately from the study of "physical" Brownian motion, mathematicians starting with Thiele in 1880 studied a continuous-time stochastic process that we today also call Brownian motion (or, in certain contexts, the **Wiener process**, in honor of Norbert Wiener, an early pioneer of the subject), and that is now understood to be a ubiquituous and stunningly successful model for many phenomena in nature, economics, science and engineering, and more. Our goal in this chapter is to develop the rigorous mathematical theory of this process; physics will not play any role in the discussion.

5.1 Preliminaries (1): multivariate normal distribution

Definition 5.1. *Let $\boldsymbol{\mu} = (\mu_1, \ldots, \mu_d) \in \mathbb{R}^d$, and let $\Sigma = (\sigma_{j,k})_{j,k=1}^d$ be a symmetric, non-negative definite matrix of real numbers. We say that a random vector $\mathbf{X} = (X_1, \ldots, X_d)$ has the d-dimensional multivariate Gaussian (or multinormal) distribution with mean $\boldsymbol{\mu}$ and covariance matrix Σ, and denote this as $\mathbf{X} \sim N(\boldsymbol{\mu}, \Sigma)$, if any of the following equivalent conditions are satisfied:*

1. *$\mathbf{E}(X_j) = \mu_j$, $Cov(X_j, X_k) = \Sigma_{j,k}$ for all $1 \le j, k \le d$, and every linear combination $\sum_{j=1}^d a_j X_j$ is a normal r.v. with distribution $N\left(\sum_j a_j \mu_j, \sum_{j,k} \sigma_{j,k} a_j a_k\right)$.*
2. *The vector $\mathbf{X}$ can be represented in the form*

$$\mathbf{X}^\top = A\mathbf{Z}^\top + \boldsymbol{\mu}^\top,$$

where $\mathbf{Z} = (Z_1, \ldots, Z_m)$ is a vector of i.i.d. $N(0,1)$ random variables, $m = \operatorname{rank}(\Sigma)$, and A is a $d \times m$ matrix. (In this case it also necessarily follows easily that $\Sigma = AA^\top$.)

3. *The characteristic function* $\varphi_{\mathbf{X}}(\mathbf{u}) = \mathbf{E}\Big[\exp\left(i\langle \mathbf{u}, \mathbf{X}\rangle\right)\Big] = \mathbf{E}\Big[\exp\left(i\sum_{j=1}^{d} u_j X_j\right)\Big]$ *is given by the formula*

$$\varphi_{\mathbf{X}}(\mathbf{u}) = \exp\left(i\langle \mathbf{u}, \boldsymbol{\mu}\rangle - \frac{1}{2}\langle \Sigma\mathbf{u}^\top, \mathbf{u}\rangle\right) = \exp\left(i\sum_{j=1}^{d}\mu_j u_j - \frac{1}{2}\sum_{j,k=1}^{d}\sigma_{j,k}u_j u_k\right)$$

In the case when Σ *is not just nonnegative-definite but actually* positive-*definite, another condition that is equivalent to the above conditions is:*

4. $\mathbf{X}$ *is an absolutely continuous random vector with* d*-dimensional p.d.f. given by*

$$f_{\mathbf{X}}(\mathbf{x}) = \frac{1}{(2\pi)^{d/2}|\Sigma|^{1/2}}\exp\left(-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^\top \Sigma^{-1}(\mathbf{x}-\boldsymbol{\mu})\right).$$

In the case when $\boldsymbol{\mu} = (0, \ldots, 0)$ *and* $\Sigma = I_d$ *is the identity matrix of dimension* d*, we say that* $\mathbf{X} \sim N(\mathbf{0}, I_d)$ *has the* ***standard*** d*-dimensional Gaussian distribution. This just means that the coordinates of* $\mathbf{X}$ *are i.i.d.* $N(0,1)$ *random variables.*

Note that this definition is actually a theorem in disguise, since it needs to be proved that the above conditions are equivalent. The details are standard; see Wikipedia.

5.2 Preliminaries (2): Gaussian processes

Definition 5.2. *A stochastic process* $(X_i)_{i\in I}$*, indexed by elements of a set* I*, is called a* ***Gaussian process*** *if for any finite set of indices* $i_1, \ldots, i_d \in I$*, the finite-dimensional vector* $(X_{i_1}, \ldots, X_{i_d})$*, is a multivariate normal vector.*

Given a Gaussian process $(X_i)_{i\in I}$, we denote functions $\mu : I \to \mathbb{R}$ and $\Sigma : I \times I \to \mathbb{R}$ by

$$\mu(i) = \mathbf{E}(X_i),$$
$$\Sigma(i,j) = \mathrm{Cov}(X_i, X_j).$$

The function $\mu(\cdot)$ is the **mean** of the process, and the function $\Sigma(\cdot,\cdot)$ is called the **covariance kernel** of the process. Because the distribution of a stochastic process is determined uniquely by its finite-dimensional marginals, and because multivariate normal vectors are determined uniquely by their mean and covariance matrix, if follows that the distribution of the process is determined uniquely by those two functions.

It is also immediate that the covariance kernel is symmetric, that is, satisfies $\Sigma(i,j) = \Sigma(j,i)$, and is a **positive-semidefinite kernel**; that is, for any $a_1, \ldots, a_d \in \mathbb{R}$ and $i_1, \ldots, i_d \in I$, we have

$$\sum_{j,k=1}^{d} a_j a_k \Sigma(i_j, i_k) \geq 0,$$

since the quantity on the left-hand side is the variance $\mathbf{V}\left(\sum_{j=1}^{d} a_j X_{i_j}\right)$.

Conversely, given a mean function $\mu : I \to \mathbb{R}$ and a kernel $\Sigma : I \to I \to \mathbb{R}$ satisfying the symmetry and positive-definiteness conditions, one can construct a Gaussian process having μ and Σ as its respective mean and covariance kernel. This involves the standard approach of constructing a stochastic process with given finite-dimensional marginals on the product space $\mathbb{R}^I$ using the Kolmogorov extension theorem. These finite-dimensional marginals are, of course, multivariate normal random vectors.

In the next section we'll define **Brownian motion**, which is undoubtedly the most important of all Gaussian processes. However, the topic of Gaussian processes is quite extensive and involves many other interesting stochastic processes that one encounters in applied probability and statistics, such as the **Brownian bridge**; **Ornstein-Uhlenbeck process**; **white noise**; and many others.

5.3 Definition and basic properties of Brownian motion

Definition 5.3. *A stochastic process $(B_t)_{t\geq 0}$ is called a **Brownian motion** (abbrev. **BM**) if it has the following properties:*

(a) If $0 \leq t_0 < t_1 < \ldots < t_n$ then $B_{t_0}, B_{t_1} - B_{t_0}, B_{t_2} - B_{t_1}, \ldots, B_{t_n} - B_{t_{n-1}}$ are independent r.v.'s.[10]

(b) For all $s, t \geq 0$, $B_{s+t} - B_s \sim N(0,t)$.

(c) There is an event E such that $\mathbf{P}(E) = 1$ and

$$\Big\{\omega \in \Omega : \text{the function } (t \mapsto B_t(\omega))_{t\geq 0} \text{ is a continuous function}\Big\} \supseteq E.$$

If $(B_t)_{t\geq 0}$ is a Brownian motion and in addition we have

[10] A continuous-time stochastic process with this property is said to have **independent increments**.

(d) $B_0 \equiv 0$

then we say that $(B_t)_{t\geq 0}$ *is a* ***standard*** *Brownian motion.*

If $(B_t)_{t\geq 0}$ is a standard Brownian motion, it is easy to see that it is a Gaussian process, with a covariance kernel that can be easily computed: for $0 \leq s \leq t$ we have

$$\mathrm{Cov}(B_s, B_t) = \mathrm{Cov}\Big(B_s, B_s + (B_t - B_s)\Big) = \mathrm{Cov}(B_s, B_s) + 0 = s \quad \text{(the variance of a } N(0,s) \text{ r.v.)}.$$

Therefore for general $s, t \geq 0$ the covariance kernel is given by the formula

$$\mathrm{Cov}(B_s, B_t) = \min(s, t).$$

As always when defining any nontrivial mathematical object, the first important goal is to convince ourselves that the object actually exists. In the next section we will prove:

Theorem 5.4. *A standard Brownian motion exists. Moreover, for any distribution* F*, a Brownian motion exists with* $B_0 \sim F$*.*

We conclude this section with a few easy lemmas.

Lemma 5.5. *The kernel* $\Sigma : \mathbb{R}_+ \times \mathbb{R}_+ \to \mathbb{R}$ *defined by*

$$\Sigma(t, s) := \min(t, s) \qquad (t, s \geq 0)$$

is symmetric and positive-definite.

Exercise 5.6. *Prove Lemma 5.5*

Lemma 5.7 (standardization)**.** *If* $(B_t)_{t\geq 0}$ *is a BM then* $(B_t - B_0)_{t\geq 0}$ *is a standard Brownian motion that is independent of* B_0*.*

Proof. Denote $C_t = B_t - B_0$, and let $(D_t)_{t\geq 0}$ denote a standard BM that is independent of the process $(B_t)_{t\geq 0}$. Clearly $C_0 \equiv 0$. For any $0 \leq t_0 < t_1 < \ldots < t_n$ and Borel sets $A, E_0, E_2, \ldots, E_n \subset \mathbb{R}$, we have

$$\begin{aligned}
&\mathbf{P}\Big(B_0 \in A, C_{t_0} \in E_0, C_{t_1} - C_{t_0} \in E_1, \ldots, C_{t_n} - C_{t_{n-1}} \in E_n\Big) \\
&= \mathbf{P}\Big(B_0 \in A, B_{t_0} - B_0 \in E_0, B_{t_1} - B_{t_0} \in E_1, \ldots, B_{t_n} - B_{t_{n-1}} \in E_n\Big) \\
&= \mathbf{P}(B_0 \in A)\mathbf{P}(B_{t_0} - B_0 \in E_0)\mathbf{P}(B_{t_1} - B_{t_0} \in E_1) \cdots \mathbf{P}(B_{t_n} - B_{t_{n-1}} \in E_n) \\
&= \mathbf{P}(B_0 \in A)\mathbf{P}(D_{t_0} - D_0 \in E_0)\mathbf{P}(D_{t_1} - D_{t_0} \in E_1) \cdots \mathbf{P}(D_{t_n} - D_{t_{n-1}} \in E_n) \\
&= \mathbf{P}\Big(B_0 \in A, D_{t_0} \in E_0, D_{t_1} - D_{t_0} \in E_1, \ldots, D_{t_n} - D_{t_{n-1}} \in E_n\Big).
\end{aligned}$$

This implies that the family of random variables $\{B_0\} \cup \{C_t \,:\, t \geq 0\}$ has the same joint distribution as the the family $\{B_0\} \cup \{D_t \,:\, t \geq 0\}$. □

Lemma 5.8 (scaling)**.** *If $(B_t)_{t\geq 0}$ is a standard BM, then we have the equalities in distribution:*

$$\{B_{at} \,:\, t \geq 0\} \stackrel{\mathcal{D}}{=} \{a^{1/2} B_t \,:\, t \geq 0\} \qquad (\textit{for any } a > 0),$$
$$\{tB_{1/t} \,:\, t > 0\} \stackrel{\mathcal{D}}{=} \{B_t \,:\, t > 0\}.$$

Proof. The first equality is immediate from standard scaling properties of Gaussian random variables. For the second equality, note that $\{tB_{1/t} \,:\, t > 0\}$ is clearly a Gaussian process with mean 0. Its covariance kernel is given by

$$\mathrm{Cov}\Big(sB_{1/s}, tB_{1/t}\Big) = st \cdot \mathrm{Cov}\Big(B_{1/s}, B_{1/t}\Big) = st \cdot \min\left(\frac{1}{s}, \frac{1}{t}\right) = \frac{st}{\max(1/s, 1/t)} = \min(s,t),$$

which coincides with the covariance kernel of the original standard BM $(B_t)_{t\geq 0}$. □

5.4 Construction of Brownian motion

Our goal in this section is to prove Theorem 5.4. We start by constructing a process that satisfies conditions (a), (b), and (d) in the definition. To this end, let $(B_t)_{t\geq 0}$ be a Gaussian process with mean $\mu(t) = 0$ and

$$\mathrm{Cov}(B_t, B_s) = \Sigma(t,s) := \min(t,s) \qquad (t, s \geq 0).$$

That such a process exists is ensured by the Lemma 5.5.

The most nontrivial part about the construction is the proof that the paths of our Gaussian process are almost surely continuous. One technical difficulty that is a common difficulty in probability theory when dealing with continuous-time stochastic processes, comes from the fact that path properties such as continuity involve looking at the value of the process at all times $t \geq 0$. This is an *uncountably infinite* number of times; this means that a set of points of our sample space Ω such as "$(t \mapsto B_t \,:\, t \geq 0)$ is a continuous function", which

naively would be written as

$$\Big\{\omega \in \Omega \,:\, \text{for all } s \geq 0 \text{ and } \epsilon > 0, \text{ there exists } \delta > 0 \text{ such that for all } t \geq 0, \text{ if } |t-s| < \delta \text{ then } |B_s(\omega) - B_t(\omega)| < \epsilon\Big\}$$
$$= \bigcap_{s \geq 0} \bigcap_{\epsilon > 0} \bigcup_{\delta > 0} \bigcap_{t \geq 0} \Big\{|B_t - B_s| < \epsilon\Big\}$$

is not obviously an *event* (that is, it does not immediately follow from the axioms of a σ-algebra that this set is in the σ-algebra $\mathcal{F}$ of events in our probability space). And even if it is an event, how does one go about showing that it has probability 1?

The answer is to try as much as possible to restrict our attention to a *countably* infinite set of times such that the properties of the process at that set of times give sufficient insight into the behavior of the process at *all* times. Fortunately, in the case of Brownian motion this is not terribly difficult to do. Let $Q_2 = \left\{\frac{m}{2^n} \,:\, n, m \in \mathbb{Z},\ m, n \geq 0\right\}$ be the set of dyadic rationals. Below we prove a number of estimates involving the restriction of the path functions $(t \mapsto B_t \,:\, t \geq 0)$ to Q_2, and then use them to prove the continuity property.

The first theorem involves a general continuous-time stochastic process satisfying certain continuity-in-the-mean bounds; Brownian motion is the obvious example, but the result has broader applicability (see also the next section) and the proof in the general case is no more difficult than the special case.

Theorem 5.9. *Let $(X_t)_{t \geq 0}$ be a stochastic process such that for some constants $K, \alpha, \beta > 0$, the inequality*

$$\mathbf{E}|X_s - X_t|^\beta \leq K|s-t|^{1+\alpha} \quad \textit{for all } s, t \in Q_2 \cap [0,1].$$

Then for any $\gamma < \alpha/\beta$, the event

$$\Big\{\omega \in \Omega \,:\, \textit{there exists } C = C(\omega) > 0 \textit{ s.t. for all } q, r \in Q_2 \cap [0,1],\ \ |X_q - X_r| \leq C|q-r|^\gamma\Big\}$$

has probability 1.

Proof. In the proof below, we use the notation $X(t)$ instead of X_t for convenience (since the indices we will be evaluating X_t at get a bit messy). Fix some small $\delta > 0$ that will be specified below. Denote

$$I_n = \left\{(i,j) \,:\, 0 \leq i < j \leq 2^n,\ j - i \leq 2^{\delta n}\right\}.$$

Consider the event

$$G_n = \left\{ \left| X\left(\frac{j}{2^n}\right) - X\left(\frac{i}{2^n}\right) \right| \leq \left(\frac{j-i}{2^n}\right)^\gamma \text{ for all } (i,j) \in I_n \right\}$$

We have

$$\begin{aligned}\mathbf{P}(G_n^c) &\leq \sum_{(i,j)\in I_n} \mathbf{P}\left(\left| X\left(\frac{j}{2^n}\right) - X\left(\frac{i}{2^n}\right) \right| > \left(\frac{j-i}{2^n}\right)^\gamma \right) \\ &= \sum_{(i,j)\in I_n} \mathbf{P}\left(\left| X\left(\frac{j}{2^n}\right) - X\left(\frac{i}{2^n}\right) \right|^\beta > \left(\frac{j-i}{2^n}\right)^{\gamma\beta} \right) \leq \sum_{(i,j)\in I_n} \frac{\mathbf{E}\left| X\left(\frac{j}{2^n}\right) - X\left(\frac{i}{2^n}\right) \right|^\beta}{\left(\frac{j-i}{2^n}\right)^{\gamma\beta}},\end{aligned}$$

by Markov's inequality. Using the inequality the process was assumed to satisfy, this last sum is

$$\leq K \sum_{(i,j)\in I_n} \left(\frac{j-i}{2^n}\right)^{1+\alpha-\gamma\beta} \leq K \cdot 2^n \cdot 2^{\delta n} \left(2^{\delta n} 2^{-n}\right)^{1+\alpha-\gamma\beta} = K \cdot 2^{-n\lambda},$$

where we define

$$\lambda = (1-\delta)(1+\alpha-\beta\gamma) - (1+\delta) = (1-\delta)(\alpha-\beta\gamma) - 2\delta.$$

Note that $\alpha - \beta\gamma > 0$. Now remembering that we left ourselves the freedom to choose the value of $\delta > 0$, we see that if we take δ positive but small enough we can guarantee that $\lambda > 0$. From now on, fix δ to be such a value. (We also want the condition $1 - \delta > 0$ to hold; note that this happens automatically with the requirement we imposed on δ.)

Next, denote $A = 3 \cdot 2^{(1-\delta)\gamma}/(1 - 2^{-\lambda})$ (a positive number), and $H_N = \cap_{n=N}^\infty G_n$. Since $\mathbf{P}(G_n^c) \to 0$ as $n \to \infty$ exponentially fast by the above estimate, the same is true for $\mathbf{P}(H_N^c)$, since

$$\mathbf{P}(H_N^c) \leq \sum_{n=N}^\infty \mathbf{P}(G_n^c) \leq K \sum_{n=N}^\infty 2^{-\lambda n} = \frac{K}{1-2^{-\lambda}} 2^{-\lambda N} \xrightarrow[N\to\infty]{} 0.$$

We claim that on the event H_N, the inequality

$$|X(q) - X(r)| \leq A|q-r|^\gamma$$

holds for all $q, r \in Q_2 \cap [0,1]$ satisfying $|q - r| < 2^{-(1-\delta)N}$. Before proving this, let us check that this is enough ti imply the claim of the theorem. Indeed, assuming the claim, the event

$$H = \{H_N \text{ eventually}\} = \bigcup_{M=1}^\infty \bigcap_{N=M}^\infty H_N = \{H_N^c \text{ infinitely often}\}^c$$

has probability 1 by the Borel-Cantelli lemma. Define a random variable L_0 as the minimal integer ℓ for which H_n occurred for all $n \geq \ell$. This r.v. is finite on the event H, that is, it is almost surely finite. Fix $\ell \geq 1$. On the event $\{L_0 = \ell\}$, the inequality $|X(q) - X(r)| \leq A|q-r|^\gamma$ holds for all $q, r \in Q_2 \cap [0,1]$ satisfying $|q-r| < 2^{-(1-\delta)\ell}$, and for the pairs $q, r \in Q_2 \cap [0,1]$ satisfying the reverse inequality $|q-r| \geq 2^{-(1-\delta)\ell}$, we have instead that

$$|X(q) - X(r)| \leq R_\ell |q-r|^\gamma,$$

where R_ℓ is a random variable defined as

$$R_\ell = 2^{(1-\delta)\gamma\ell} \sup\Big\{|X(q) - X(r)| \,:\, q, r \in [0,1] \cap Q_2,\ |q-r| \geq 2^{-(1-\delta)\ell}\Big\}.$$

on the event $\{L_0 = \ell\}$, and as 0 elsewhere. It is not hard to see that R_ℓ is finite. Thus, we can conclude that the inequality $|X(q) - X(r)| \leq D_\ell |q-r|^\gamma$ is satisfied for all $q \in [0,1] \cap Q_2$, where $D_\ell = \max(A, R_\ell)$, on the event $\{L_0 = \ell\}$. Now gluing the D_ℓ's together for the different values of $\ell \geq 1$, we obtain a single, a.s. finite random variable D for which the inequality $|X(q) - X(r)| \leq D_\ell |q-r|^\gamma$ is satisfied for all $q \in [0,1] \cap Q_2$, which was the result to prove.

Now to prove the claim about H_N, take some $q, r \in Q_2 \cap [0,1]$ satisfying $q < r$ and $|q-r| < 2^{-(1-\delta)N}$. There is an $m \geq N$ such that $2^{-(m+1)(1-\delta)} \leq r - q < 2^{-m(1-\delta)}$. Furthermore, we can write

$$r = \frac{j}{2^m} + 2^{-r_1} + \ldots + 2^{-r_\ell}, \quad q = \frac{i}{2^m} - 2^{-q_1} - \ldots - 2^{-q_k},$$

for some integers i, j and exponents $m < r_1 < r_2 < \ldots < r_\ell$, $m < q_1 < q_2 < \ldots < q_k$. It follows that $j - i \leq 2^m(r-q) < 2^{\delta m}$. By the definition of the event H_N, on that event we have

$$\left|X\left(\frac{i}{2^m}\right) - X\left(\frac{j}{2^m}\right)\right| \leq \left(\frac{j-i}{2^m}\right)^\gamma \leq \left(2^{\delta m} 2^{-m}\right)^\gamma.$$

We also get on H_N that

$$\left|X(q) - X\left(\frac{i}{2^m}\right)\right| \leq \sum_{h=1}^{k} 2^{-q_h\gamma} \leq \sum_{d=m}^{\infty} 2^{-\gamma d} = \frac{1}{1-2^{-\gamma}} 2^{-\gamma m}, \quad \text{and, similarly,}$$
$$\left|X(r) - X\left(\frac{j}{2^m}\right)\right| \leq \frac{1}{1-2^{-\gamma}} 2^{-\gamma m}.$$

Combining these bounds gives

$$|X(q) - X(r)| \leq 2^{-\gamma m(1-\delta)} + 2\frac{1}{1-2^{-\gamma}}2^{-\gamma m} \leq 3\frac{1}{1-2^{-\gamma}} \cdot 2^{-\gamma m(1-\delta)}.$$

Moreover, observe that $2^{-(1-\delta)\gamma}2^{-m(1-\delta)\gamma} \leq |r-q|^\gamma$, so that we obtain the bound

$$|X(q) - X(r)| \leq \frac{3}{1-2^{-\gamma}}2^{(1-\delta)\gamma}|r-q|^\gamma,$$

as was to be shown. □

Theorem 5.10. *Let Q_2 be as before the set of dyadic rationals, and let $(B_t)_{t\geq 0}$ be the Gaussian process we were working with earlier. For any $T > 0$ there is an event $E = E_T$ such that $\mathbf{P}(E) = 1$ and*

$$E \subseteq \left\{\omega \in \Omega : \big(\omega \mapsto B_t(\omega)\big)_{t\in Q_2\cap[0,T]} \text{ is a uniformly continuous function}\right\}.$$

Proof. By Lemma 5.8, it is enough to prove this with $T = 1$. The idea now is to apply Theorem 5.9, where in our particular case we have $X_t = B_t$, $\beta = 4$, $\alpha = 1$, $\gamma \in (0, 1/4)$. Such a choice of parameters works (that is, satisfies the assumptions of Theorem 5.9), since

$$\mathbf{E}|B_t - B_s|^4 = \kappa_4(t-s)^2, \qquad \text{where } \kappa_4 = \text{the fourth moment of a } N(0,1) \text{ r.v.}$$

The conclusion is that on the probability-1 event guaranteed to exist by Theorem 5.9, there exists a constant $C > 0$ such that $|B_t - B_s| \leq C|t-s|^\gamma$ for all $s, t \in Q_2 \cap [0,1]$, which immediately implies uniform continuity. □

Exercise 5.11. *Let $D \subseteq [a,b]$ be a dense subset of $[a,b]$. Prove that if $f : D \to \mathbb{R}$ is uniformly continuous then it has a unique extension $\overline{f} : [a,b] \to \mathbb{R}$.*

Proof of Theorem 5.4. Let $(B_t)_{t\geq 0}$ be the Gaussian process we constructed earlier, defined on a probability space $(\Omega, \mathcal{F}, \mathbf{P})$, where we now interpret the probability space to be the canonical product space $\mathbb{R}^{[0,\infty)}$, with the random variable B_t being defined as the coordinate function mapping $\omega \in \mathbb{R}^{[0,\infty)}$ to its t-coordinate $B_t(\omega) = \omega(t)$.

We modify the construction of this Gaussian process as follows. Instead of the canonical product space $\mathbb{R}^{[0,\infty)}$, we use only the part $\mathbb{R}^{Q_2}$ — call this space Ω, with the associated product σ-algebra $\mathcal{F}$ and measure $\mathbf{P}$; in other words, we discard from our earlier process the random variables B_t for $t \notin Q_2$. Theorem 5.10 can still be applied to this reduced process,

since it (and Theorem 5.9 that it relies on) only make reference to the values of the stochastic process on dyadic rationals. Let $E_\infty = \cup_{T=1}^\infty E_T$, where E_T are the events from Theorem 5.10. The event E is a probability 1 event on which Brownian paths $(t \mapsto B_t)_{t\geq 0, t\in Q_2}$ are defined and are uniformly continuous on $Q_2 \cap [0,T]$ for any $T > 0$.

Define a new probability space $(\Omega', \mathcal{F}', \mathbf{P}')$ by setting

$$\Omega' = C[0,\infty) = \text{continuous functions on } [0,\infty),$$
$$\mathcal{F}' = \text{the } \sigma\text{-algebra generated by the coordinate functions } \omega \mapsto \omega(t),\ t \geq 0,$$
$$\mathbf{P}' = \mathbf{P} \circ T^{-1},$$

where $T : \Omega \to \Omega'$ is the mapping that takes a function $f \in \Omega$ and returns its unique continuous extension $\overline{f} : [0,\infty) \to \mathbb{R}$. (Exercise: check that T is measurable.)

Finally, we have constructed a probability space $(\Omega', \mathcal{F}', \mathbf{P}')$ where the sample space is the space of continuous functions on $[0,\infty)$, so property (c) in the definition of BM is automatically satisfied. Since the mapping T might have modified the paths on $[0,\infty) \setminus Q_2$, now we only know that properties (a) and (b) of the definition are satisfied for values $s, t, t_0, t_1, \ldots, t_n$ that are in Q_2; but since we are now working with continuous paths, it is easy to check by a limiting argument that they must hold also for arbitrary values in $[0,\infty)$. $\square$

Note. The probability space $(\Omega', \mathcal{F}', \mathbf{P}')$ is the natural probability space on which Brownian motion is defined. It is sometimes called **Wiener space**, after Norbert Wiener.

The construction used in the proof above was somewhat messy, and it is easy to get lost in the technical details. Conceptually, a key high-level idea to take away is that it is beneficial to have a realization of Brownian motion (that is, a concrete probability space on which we define a stochastic process that is shown to satisfy the axioms of Brownian motion) that uses only countably many random variables as its "source of randomness". This allows getting a handle on many subtle events whose definition seems to require inspecting the values of B_t for a continuum range of values of t. (The meta-principle here is that working with continuum-sized families of random variables is something we try as to avoid as much as possible in the theory of stochastic processes, so as to avoid having to deal with delicate issues of measurability.) The specific countable family of random variables that we chose to work with in the current construction are the variables B_t for $t \in Q_2$, but that is a technical detail and by no means a unique or canonical choice.

As it turns out, there are other ways to construct Brownian motion using a countable family of random variables, including some that are in many ways more natural than the representation in terms of the values of B_t on dyadic rationals. The Fourier series representation considered in the following exercise, discovered by Wiener, is perhaps the most elegant representation of this type.

Exercise 5.12 (Fourier series representation of Brownian motion on an interval)**.** *Let $Z_0, Z_1, Z_2, \ldots$ be an i.i.d. sequence of $N(0,1)$ random variable. Define a stochastic process $(X_t)_{0\le t\le 1}$ by the random series*

$$X_t = Z_0 t + \sqrt{2}\sum_{n=1}^{\infty} Z_n \frac{\sin(\pi n t)}{\pi n}.$$

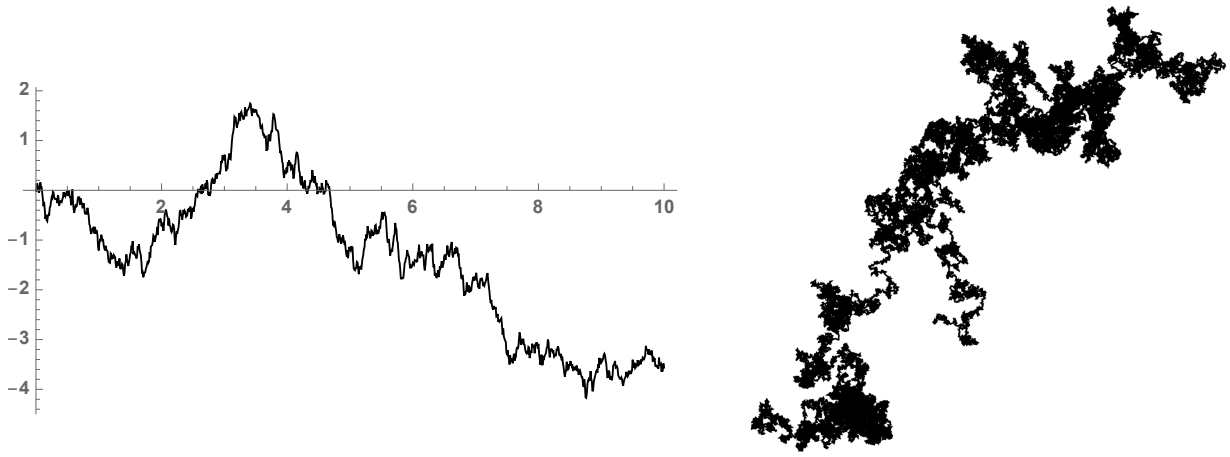
1. *Prove that this infinite series converges absolutely almost surely for all $0 \le t \le 1$, hence $(X_t)_{0\le t\le 1}$ is a well-defined stochastic process. (Hint: Kolmogorov three-series theorem.)*

2. *Prove that the process $(X_t)_{0\le t\le 1}$ is (the restriction to $[0,1]$ of) a standard Brownian motion. That is, if $(B_t)_{t\ge 0}$ a standard BM then we have the equality in distribution*

$$(X_t)_{0\le t\le 1} \stackrel{\mathcal{D}}{=} (B_t)_{0\le t\le 1}.$$

5.5 Hölder-continuity, nondifferentiability of BM

One of the fascinating aspects of Brownian motion is that it is a naturally occurring random **fractal**, that is, a set with fractional dimension. This is intuitively rather obvious when looking at simulated Brownian motion paths (Fig. 5(a)–(b)). Mathematically it is quite a subtle phenomenon however, discussed extensively in the literature (see for example [9, Ch. 4] for the proof that the graph of a 1-dimensional Brownian motion almost surely has Hausdorff dimension 3/2, and [7] for the proof of the much more difficult result that the boundary of planar Brownian motion is 4/3). Here, we give a small taste of some of fractal-like properties of the Brownian motion paths by discussing the Lipschitz continuity, differentiability, and Hölder continuity of the BM paths. (Recall that a real-valued function f defined on some real interval is called **Lipschitz-continuous** if it satisfies an inequality of the form $|f(x)-f(y)| \le C|x-y|$ for some constant C and all x, y; more generally, f is called **Hölder-continuous** with exponent α if satisfies the inequality $|f(x)-f(y)| \le C|x-y|^\alpha$ for some C and all x, y.) We will prove the following two results.



(a) (b)

Figure 5: (a) the simulated graph of a 1-dimensional Brownian motion; (b) a planar BM sample path

Theorem 5.13. *The Brownian motion paths are almost surely Hölder-continuous with exponent γ, for any $\gamma < 1/2$.*

Theorem 5.14. *Almost surely, Brownian motion is not Lipschitz-continuous at any point, and in particular it is not differentiable at any point.*

Proof of Theorem 5.13. In the previous section, we applied Theorem 5.9 to deduce that for any $\gamma < 1/4$, almost surely the inequality $|B_t - B_s| \leq C|t-s|^\gamma$ is satisfed uniformly on any compact interval. That is, BM is almost surely Hölder-continuous with exponent γ for any $\gamma < 1/4$. Now note the following way in which this estimate can be improved: for any $m \geq 1$, if we denote $Z \sim N(0,1)$, then

$$\mathbf{E}|B_t - B_s|^{2m} = \mathbf{E}\left(\sqrt{|t-s|}Z\right)^{2m} = |t-s|^m \mathbf{E}(Z^{2m}) = C_m|t-s|^m.$$

So in fact we can also apply Theorem 5.9 with $\alpha = m-1$, $\beta = 2m$, and any γ satisfying $\gamma < \alpha/\beta = \frac{1}{2} - \frac{1}{2m}$, to get that Brownian motion is almost surely Hölder-continuous with exponent γ. Since this is true for any $m \geq 1$, it follows that in fact for any $\gamma < 1/2$ we have a.s. Hölder-continuity with exponent γ, as claimed. □

Proof of Theorem 5.14. Fix an arbitrary constant $C > 0$, and define a sequence of events

$$A_n = \Big\{ \omega \in \Omega \,:\, \text{there exists } s \in [0,1] \text{ s.t. } |B_s(\omega) - B_t(\omega)| \le C|t-s| \\ \text{for any } t \in [0,1] \text{ satisfying } |t-s| \le \frac{3}{n} \Big\}$$

If we can show that $\mathbf{P}(A_n) = 0$ for any $n \ge 1$ and any $C > 0$, that would imply the result.

Define random variables

$$Y_{k,n} = \max\left\{ \left| B\left(\frac{k+j}{n}\right) - B\left(\frac{k+j-1}{n}\right) \right| \,:\, j = 0,1,2 \right\}, \qquad 1 \le k \le n-2,$$

and events

$$E_n = \left\{ \text{at least one } Y_{k,n} \text{ for some } 1 \le k \le n-2 \text{ is } \le \frac{6C}{n} \right\}.$$

It is not difficult to see that $A_n \subseteq E_n$: indeed if $s \in [0,1]$ is such that $|B_t - B_s| \le C|t-s|$ for t satisfying $|t-s| \le \frac{3}{n}$, take $1 \le k \le n-2$ such that $\frac{k-1}{n} \le s \le \frac{k+2}{n}$, and note that, for any $j = 0, 1, 2$,

$$\left| B\left(\frac{k+j}{n}\right) - B\left(\frac{k+j-1}{n}\right) \right| \le \left| B\left(\frac{k+j}{n}\right) - B(s) \right| + \left| B(s) - B\left(\frac{k+j-1}{n}\right) \right| \\ \le \left(\frac{3}{n} + \frac{3}{n}\right) C = \frac{6C}{n}.$$

It follows that

$$\mathbf{P}(A_n) \le \mathbf{P}(E_n) \le n\mathbf{P}\left(|B(1/n)| \le \frac{6C}{n} \right)^3 \le n\mathbf{P}\left(|B_1| \le \frac{6C}{\sqrt{n}} \right)^3 \\ \le n\left(\frac{1}{\sqrt{2\pi}} \cdot \frac{12C}{\sqrt{n}} \right)^3 \le \frac{\text{const}}{\sqrt{n}} \xrightarrow[n\to\infty]{} 0.$$

Since $A_n \subseteq A_{n+1}$, this implies that $\mathbf{P}(A_n) = 0$ for all n, as claimed, and the proof is complete. □

5.6 The Markov property and its consequences

Let us fix some notation. The measure space on which BM lives is (C, Σ), where

$C = C[0, \infty)$,

$\Sigma = \sigma\Big(\omega \mapsto \omega_t \,:\, t \ge 0 \Big)$ (the σ-algebra generated by the coordinate functions).

For a probability measure, we equip this space with a *family* of probability measures $\mathbf{P}_x$, where $x \in \mathbb{R}$, and for each $x \in \mathbb{R}$ the measure $\mathbf{P}_x$ is the probability measure under which the coordinate functions $\Big(B_t(\omega) = \omega(t)\Big)_{t\geq 0}$ become a Brownian motion started at x (that is, we have $\mathbf{P}_x(B_0 = x) = 1$). The notation $\mathbf{E}_x$ will denote the expectation operator relative to the measure $\mathbf{P}_x$.

For $s \geq 0$, define also σ-algebras

$$\mathcal{F}_s^\circ = \sigma\Big(B_s \,:\, t \leq s\Big),$$
$$\mathcal{F}_s^+ = \bigcap_{t>s} \mathcal{F}_t^\circ.$$

Each of the families of σ-algebras $(\mathcal{F}_s^\circ)_{s\geq 0}$ and $(\mathcal{F}_s^+)_{s\geq 0}$ is a **filtration** (that is, an increasing family of σ-algebras indexed by an integer- or real-valued parameter). Conceptually speaking, $\mathcal{F}_s^\circ$ represents the information known about the behavior of our BM up to time s, and $\mathcal{F}_s^\circ$ represents the information known about the behavior of the BM up to time "$s+$". Both filtrations are interesting and natural. The $\mathcal{F}_s^+$ are right-continuous, i.e., satisfy $\mathcal{F}_s^+ = \bigcap_{t>s} \mathcal{F}_t^+$. The filtrations are not identical, e.g., in the sense that, one can define r.v.s that are measurable w.r.t. $\mathcal{F}_s^+$ but not $\mathcal{F}_s^\circ$ (for example: $X_s = \limsup_{t\searrow s} \frac{B_t - B_s}{f(t-s)}$ is such a random variable for any measurable function $f : [0,\infty) \to [0,\infty)$). But it turns out that such r.v.s are a.s. constant, in other words, $\mathcal{F}_s^\circ$ and $\mathcal{F}_s^+$ differ only in sets that have measure 0 with respect to any of the measures $\mathbf{P}_x$.

For $s \geq 0$, let $\theta_s : C \to C$ denote the shift transformation

$$(\theta_s\omega)(t) = \omega(s+t).$$

For $t \geq 0$ and $x, y \in \mathbb{R}$, denote

$$p_t(x,y) = \frac{1}{\sqrt{2\pi t}} \exp\left(-\frac{(y-x)^2}{2t}\right).$$

The function $p_t(x,y)$ is called the **transition kernel** of Brownian motion,[11] and plays a role in the theory analogous to the matrix powers A^n, where A is the transition matrix

[11] In a mathematical analysis context this function is called the **heat kernel**, and plays a fundamental role in the theory of the heat equation (aka the diffusion equation). Indeed, there are many connections between Brownian motion and the heat equation, and more generally partial differential equations, which have been exploited to great effect.

of a discrete-time, discrete-space Markov chain. Intuitively, $p_t(x,y)$ can be thought of as "the probability density at y of B_t conditioned on $B_0 = x$". With this interpretation, the following lemma seems like a fairly intuitive statement, especially when considering it again as an analogue to a statement about discrete-time, discrete-space Markov chains.

Lemma 5.15. *If $0 < u_1 < \ldots < u_n$ are real numbers and $g_1, \ldots, g_n : \mathbb{R} \to \mathbb{R}$ are bounded measurable functions, then*

$$\mathbf{E}_x\left(\prod_{j=1}^{n} g_j(B_{u_j})\right) = \int_{\mathbb{R}} p_{u_1}(x,z_1)g_1(z_1)\,dz_1 \int_{\mathbb{R}} p_{u_2-u_1}(z_1,z_2)g_2(z_2)\,dz_2 \\ \cdots \int_{\mathbb{R}} p_{u_n-u_{n-1}}(z_{n-1},z_n)g_n(z_n)\,dz_n$$

Proof. Let $X_1, \ldots, X_n$ denote independent random variables distributed as follows: $X_1 \sim N(x, u_1)$, $X_j \sim N(0, u_j - u_{j-1})$ for $2 \le j \le n$. Define random variables $Y_1, \ldots, Y_n$ by

$$Y_k = \sum_{j=1}^{k} X_j.$$

By the definition of BM, the random vector $(Y_1, \ldots, Y_n)$ is equal in distribution to the random vector $(B_{u_1}, \ldots, B_{u_n})$ under the measure $\mathbf{P}_x$.

Moreover, by the standard formula for the transformation of a vector r.v., the joint density of $(Y_1, \ldots, Y_n)$ can be expressed in terms of the joint density of $(X_1, \ldots, X_n)$ as

$$f_{Y_1,\ldots,Y_n}(y_1,\ldots,y_n) = f_{X_1,\ldots,X_n}(y_1, y_2-y_1, \ldots, y_n - y_{n-1}) = f_{X_1}(y_1)\prod_{j=2}^{n} f_{X_j}(y_j - y_{j-1}) \\ = p_{u_1}(x,y_1)\prod_{j=2}^{n} p_{u_j-u_{j-1}}(y_{j-1},y_j).$$

Therefore we get that

$$\mathbf{E}_x\left(\prod_{j=1}^{n} g_j(B_{u_j})\right) = \int\ldots\int_{\mathbb{R}^n} \prod_{j=1}^{n} g_j(z_j)\cdot p_{u_1}(x,z_1)\cdot \prod_{j=2}^{n} p_{u_j-u_{j-1}}(z_{j-1},z_j)\,dz_1\ldots dz_n,$$

which is the same as the identity in the lemma. □

Theorem 5.16 (The Markov property)**.** *If $s \ge 0$ and Y is a r.v. on (C,Σ), then for any $x \in \mathbb{R}$, we have the identity*

$$\mathbf{E}_x\Big(Y \circ \theta_s \Big| \mathcal{F}_s^+\Big) = \mathbf{E}_{B_s} Y,$$

where $\mathbf{E}_{B_s} Y = \mathbf{E}_z(Y)\big|_{z=B_s}$, and $(B_t)_{t\ge 0}$ is the BM defined on the space (C,Σ) as above.

Proof. The random variable $\mathbf{E}_{B_s} Y$ is $\mathcal{F}_s^+$-measurable (being a function of B_s). We need to prove that for any $A \in \mathcal{F}_s^+$, the relation

$$\mathbf{E}_x\Big((Y \circ \theta_s)\mathbf{1}_A\Big) = \mathbf{E}_x\Big((\mathbf{E}_{B_s} Y)\mathbf{1}_A\Big)$$

holds. First, we consider Y and A of a special form. Fix numbers

$$0 < t_1 < \ldots < t_n, \quad 0 < h < t_1, \quad 0 < s_1 < \ldots < s_k \le s + h.$$

Let $Y = \prod_{m=1}^n f_m(B_{t_m})$, where $f_1, \ldots, f_n : \mathbb{R} \to \mathbb{R}$ are bounded measurable functions. Let

$$A = \Big\{ B_{s_j} \in A_j \text{ for } j = 1, \ldots, k \Big\},$$

where $A_1, \ldots, A_k$ are Borel sets in $\mathbb{R}$. In this case, clearly $\mathbf{1}_A = \prod_{j=1}^k \mathbf{1}_{\{B_{s_j} \in A_j\}}$. So, by applying Lemma 5.15 with a carefully chosen set of parameters, we can write

$$\begin{aligned}
\mathbf{E}_x\Big((Y \circ \theta_s)\mathbf{1}_A\Big) &= \mathbf{E}_x \left(\prod_{j=1}^k \mathbf{1}_{\{B_{s_j} \in A_j\}} \prod_{m=1}^n f_m(B_{t_m+s}) \right) \\
&= \mathbf{E}_x \left(\prod_{j=1}^k \mathbf{1}_{\{B_{s_j} \in A_j\}} \prod_{m=1}^n f_m(B_{t_m+s}) \cdot \mathbf{1}_{\{B_{s+h} \in \mathbb{R}\}} \right) \\
&= \int_{A_1} p_{s_1}(x, x_1)\, dx_1 \int_{A_2} p_{s_2 - s_1}(x_1, x_2)\, dx_2 \cdots \int_{A_k} p_{s_k - s_{k-1}}(x_{k-1}, x_k)\, dx_k \\
&\qquad \times \int_{\mathbb{R}} p_{s+h-s_k}(y) \cdot 1\, dy \\
&\qquad \times \int_{\mathbb{R}} p_{t_1 - h}(y, y_1) f_1(y_1)\, dy_1 \int_{\mathbb{R}} p_{t_2 - t_1}(y_1, y_2) f_2(y_2)\, dy_2 \cdots \int_{\mathbb{R}} p_{t_n - t_{n-1}}(y_{n-1}, y_n) f_n(y_n)\, dy_n \\
&= \int_{A_1} p_{s_1}(x, x_1)\, dx_1 \int_{A_2} p_{s_2 - s_1}(x_1, x_2)\, dx_2 \cdots \int_{A_k} p_{s_k - s_{k-1}}(x_{k-1}, x_k)\, dx_k \\
&\qquad \times \int_{\mathbb{R}} p_{s+h-s_k}(y) \cdot \varphi(y, h)\, dy,
\end{aligned}$$

where we denote

$$\varphi(y, h) = \int_{\mathbb{R}} p_{t_1 - h}(y, y_1) f_1(y_1)\, dy_1 \int_{\mathbb{R}} p_{t_2 - t_1}(y_1, y_2) f_2(y_2)\, dy_2 \cdots \int_{\mathbb{R}} p_{t_n - t_{n-1}}(y_{n-1}, y_n) f_n(y_n)\, dy_n.$$

By another application of Lemma 5.15, this is equal to $\mathbf{E}_x\Big(\varphi(B_{s+h}, h)\mathbf{1}_A\Big)$. This equality holds for all finite-dimensional cylinder sets $A \in \mathcal{F}^\circ_{s+h}$. By the π-λ theorem, it must therefore hold for all $A \in \mathcal{F}^\circ_{s+h}$, and in particular for $A \in \mathcal{F}_s^+ \subset \mathcal{F}^\circ_{s+h}$.

Now, we want to justify the idea of setting $h = 0$ in this equation. Note that $\varphi(y, h)$ can be written in the form

$$\varphi(y, h) = \int p_{t_1-h}(y, y_1)\psi(y_1)\, dy_1,$$

where

$$\psi(y_1) = f_1(y_1) \int p_{t_2-t_1}(y_1, y_2) f_2(y_2)\, dy_2 \cdots \int p_{t_n-t_{n-1}}(y_{n-1}, y_n) f_n(y_n)\, dy_n.$$

It is easy to check that ψ is bounded and measurable. It follows by the dominated convergence theorem that if $h \searrow 0$ and $y = y(h) \to y(0)$ then $\varphi(y(h), h) \to \varphi(y(0), 0)$. Taking $y(h) = B_{s+h}$, we get that

$$\varphi(B_{s+h}, h) \xrightarrow[h\searrow 0]{} \varphi(B_s, 0) \quad \text{a.s..}$$

Applying the bounded convergence theorem, we get that

$$\mathbf{E}_x\Big((Y \circ \theta_s)\mathbf{1}_A\Big) = \mathbf{E}_x\Big(\varphi(B_{s+h}, h)\mathbf{1}_A\Big) \xrightarrow[h\searrow 0]{} \mathbf{E}_x\Big(\varphi(B_s, 0)\mathbf{1}_A\Big).$$

Note that $\varphi(y, 0)$ can be interpreted, again by Lemma 5.15, as $\mathbf{E}_y(Y)$. So we have shown that

$$\mathbf{E}_x\Big(Y \circ \theta_s \mathbf{1}_A\Big) = \mathbf{E}_x\mathbf{E}_{B_s}(Y) \cdot \mathbf{1}_A\Big),$$

which was the claim. We proved this for all $A \in \mathcal{F}_s^+$ and Y of the special form $\prod_{j=1}^n f_j(B_{t_j})$. The extension to general Y now follows from the Monotone Class Theorem (Theorem 5.2.2 on page 275 of [4]). □

The Markov property allows us to prove "obvious" statements, such as the following:

Exercise 5.17. *Define random variables*

$$T_0 = \inf\Big\{s > 0 \,:\, B_s = 0\Big\},$$
$$R = \inf\Big\{t > 1 \,:\, B_t = 0\Big\},$$
$$L = \sup\Big\{t \le 1 \,:\, B_t = 0\Big\}.$$

Prove that the following relations hold:

$$\mathbf{P}_x(R > 1 + t) = \int_{\mathbb{R}} p_1(x, y)\mathbf{P}_y(T_0 > t)\, dy,$$
$$\mathbf{P}_0(L \le t) = \int_{\mathbb{R}} p_t(0, y)\mathbf{P}_y(T_0 > 1 - t)\, dy.$$

Note that $\mathbf{E}_x(Y \circ \theta_s \,|\, \mathcal{F}_s^+) = \mathbf{E}_{B_s} Y$ ends up being $\mathcal{F}_s^\circ$-measurable (being a function of B_s). It follows that $\mathbf{E}_x(Y \circ \theta_s \,|\, \mathcal{F}_s^+) = \mathbf{E}_x(Y \circ \theta_s \,|\, \mathcal{F}_s^\circ)$. Therefore also $\mathbf{E}_x(X \cdot (Y \circ \theta_s) \,|\, \mathcal{F}_s^+) = \mathbf{E}_x(X \cdot (Y \circ \theta_s) \,|\, \mathcal{F}_s^\circ)$ for any r.v. X that is integrable and $\mathcal{F}_s^\circ$-measurable. This shows that $\mathbf{E}_x(Z \,|\, \mathcal{F}_s^+) = \mathbf{E}_x(Z \,|\, \mathcal{F}_s^\circ)$ whenever Z is of the form $Z = \prod_{j=1}^n f_j(B_{t_j})$ for some real numbers $0 < t_1 < \ldots < t_n$ and bounded and measurable functions $f_1, \ldots, f_n : \mathbb{R} \to \mathbb{R}$. By the monotone class theorem, this extends to arbitrary bounded random variables Z. We have proved:

Theorem 5.18. *For any bounded r.v. Z on (C, Σ), any $s \geq 0$ and any $x \in \mathbb{R}$, we have*

$$\mathbf{E}_x\Big(Z \,|\, \mathcal{F}_s^\circ\Big) = \mathbf{E}_x\Big(Z \,|\, \mathcal{F}_s^+\Big) \quad a.s.$$

Theorem 5.19 (Blumenthal's 0-1 law)**.** *If $A \in \mathcal{F}_0^+$ then for any $x \in \mathbb{R}$, $\mathbf{P}_x(A) = 0$ or 1. That is, the field $\mathcal{F}_0^+$, known as the **germ field**, is trivial w.r.t. any of the measures $\mathbf{P}_x$.*

Proof.

$$\mathbf{1}_A = \mathbf{E}_x\Big(\mathbf{1}_A \,|\, \mathcal{F}_0^+\Big) \overset{\text{a.s.}}{=} \mathbf{E}_x\Big(\mathbf{1}_A \,|\, \mathcal{F}_0^\circ\Big) = \mathbf{E}_x\Big(\mathbf{1}_A \,|\, B_0\Big) = \mathbf{P}_x(A) \quad \mathbf{P}_x\text{-a.s.}$$

So $\mathbf{P}_x(A) = 0$ or 1. □

Theorem 5.20. *Let $\tau = \inf\Big\{t \geq 0 \,:\, B_t > 0\Big\}$. Then $\mathbf{P}_0(\tau = 0) = 1$.*

Proof. The event $\{\tau = 0\}$ is in $\mathcal{F}_0^+$, so by Theorem 5.19, $\mathbf{P}_0(\tau = 0)$ is 0 or 1. Furthermore,

$$\mathbf{P}_0(\tau = 0) = \lim_{t \downarrow 0} \mathbf{P}_0(\tau \leq t) \geq \limsup_{t \downarrow 0} \mathbf{P}_0(B_t > 0) = \frac{1}{2}.$$

So $\mathbf{P}_0(\tau = 0)$ must be 1. □

Corollary 5.21. *If $T_0 = \inf\Big\{t > 0 \,:\, B_t = 0\Big\}$, then $\mathbf{P}_0(T_0 = 0) = 1$.*

Exercise 5.22. *Use Corollary 5.21 to prove that for any $a < b$, BM a.s. has a local maximum in (a, b). That is, the set of local maxima is a.s. dense.*

Theorem 5.23. *If $(B_t)_{t \geq 0}$ is the standard BM, then the process $(X_t)_{t \geq 0}$ defined by $X_0 = 0$, $X_t = tB_{1/t}$, is also a standard BM.*

Proof. We already proved this for $t > 0$ in the scaling lemma (Lemma 5.8), so it remains to verify that $(X_t)_{t\geq 0}$ is a.s. continuous at $t = 0$. To this end, note first that $\frac{1}{n}B_n \xrightarrow[n\to\infty]{} 0$ a.s. by the strong law of large numbers. Furthermore, by Kolmogorov's maximal inequality (Theorem 2.5.5 on page 84 of [4]) we can write for each $n, m \geq 1$ that

$$\mathbf{P}\left(\max_{0\leq k\leq 2^m}\left|B\left(n+\frac{k}{2^m}\right)-B_n\right| > n^{2/3}\right) \leq n^{-4/3}\mathbf{E}(B_{n+1}-B_n)^2 = n^{-4/3}.$$

Since the bound is independent of m, we then get that

$$\begin{aligned}\mathbf{P}\left(\sup_{n\leq u\leq n+1}|B_u-B_n| > n^{2/3}\right) &= \mathbf{P}\left(\max_{m\geq 1,\ 0\leq k\leq 2^m}\left|B\left(n+\frac{k}{2^m}\right)-B_n\right| > n^{2/3}\right)\\ &= \lim_{m\to\infty}\mathbf{P}\left(\max_{0\leq k\leq 2^m}\left|B\left(n+\frac{k}{2^m}\right)-B_n\right| > n^{2/3}\right) \leq n^{-4/3}.\end{aligned}$$

Since $\sum_n n^{-4/3} < \infty$, we get using the Borel-Cantelli lemma that

$$\begin{aligned}\mathbf{P}\left(\lim_{t\to\infty}X_t = 0\right) &= \mathbf{P}\left(\lim_{u\to\infty}\frac{B_u}{u} = 0\right)\\ &\geq \mathbf{P}\left(\left\{\lim_{n\to\infty}\frac{1}{n}B_n = 0\right\}\bigcap\left\{\sup_{n\leq u\leq n+1}|B_u-B_n| > n^{2/3}\text{ i.o.}\right\}^{\mathrm{c}}\right) = 1. \quad \square\end{aligned}$$

Define the family of **future** σ-algebras for BM by $\mathcal{F}'_t = \sigma(B_s : s \geq t)$. The **tail σ-algebra** is $\mathcal{T} = \cap_{t\geq 0}\mathcal{F}'_s$.

Theorem 5.24. *If $\mathcal{A} \in \mathcal{T}$ then $\mathbf{P}_x(A)$ is either 0 or 1, and has the same value for all values of $x \in \mathbb{R}$.*

Proof. The tail σ-algebra of B_t is exactly the same as the germ σ-algebra of $X_t = tB_{1/t}$. So $\mathbf{P}_0(A) \in \{0, 1\}$. Next, since $A \in \mathcal{F}'_1$, we can write $A = \theta_1^{-1}(D)$, i.e., $\mathbf{1}_A = \mathbf{1}_D \circ \theta_1$, for some $D \in \Sigma$. From the Markov property we therefore get that

$$\begin{aligned}\mathbf{P}_x(A) = \mathbf{E}_x(\mathbf{1}_D\circ\theta_1) = \mathbf{E}_x\Big[\mathbf{E}_x(\mathbf{1}_D\circ\theta_1\,|\,\mathcal{F}_1^\circ)\Big] = \mathbf{E}_x\Big[\mathbf{E}_{B_1}(\mathbf{1}_D)\Big] = \mathbf{E}_x\Big[\mathbf{P}_y(D)\Big|_{y=B_1}\Big]\\ = \int_{\mathbb{R}} p_1(x,y)\mathbf{P}_y(D)\,dy = \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{\infty} e^{-(y-x)^2/2}\mathbf{P}_y(D)\,dy.\end{aligned}$$

From this representation we see that if $\mathbf{P}_0(A) = 0$ then necessarily $\mathbf{P}_y(D) = 0$ for Lebesgue-almost every $y \in \mathbb{R}$, and therefore $\mathbf{P}_x(A) = 0$ for all $x \in \mathbb{R}$. Similarly, if $\mathbf{P}_0(A) = 1$ then the same argument applied to A^c instead of A leads to the conclusion that $\mathbf{P}_x(A) = 1$ for all $x \in \mathbb{R}$. $\square$

Theorem 5.25. *Brownian motion is recurrent. That is, for any $x \in \mathbb{R}$, we have*

$$\mathbf{P}_x\left(\bigcap_{n\geq 1}\{B_t = 0 \text{ for some } t \geq n\}\right) = 1.$$

Proof. Fix $K > 0$. We have

$$\mathbf{P}_0\left(\frac{B_n}{\sqrt{n}} \geq K \text{ infinitely often}\right) \geq \limsup_{n\to\infty} \mathbf{P}_0\left(B_n \geq K\sqrt{n}\right) = \mathbf{P}_0(B_1 \geq K) > 0.$$

By Theorem 5.24, it follows that actually $\mathbf{P}_0\left(\frac{B_n}{\sqrt{n}} \geq K \text{ infinitely often}\right) = 1$, since this is a tail event. Symmetrically, we also have $\mathbf{P}_0\left(\frac{B_n}{\sqrt{n}} \leq -K \text{ infinitely often}\right) = 1$. Since this is true for arbitrary $K > 0$, we conclude that

$$\mathbf{P}_0\left(\limsup_{t\to\infty} \frac{B_t}{\sqrt{t}} = \infty, \quad \liminf_{t\to\infty} \frac{B_t}{\sqrt{t}} = -\infty\right) = 1.$$

The event $A := \bigcap_{n\geq 1}\{B_t = 0 \text{ for some } t \geq n\}$ is an even bigger event, so $\mathbf{P}_0(A) = 1$. Moreover, A is a tail event, so, again from Theorem 5.24 we see that $\mathbf{P}_x(A) = 1$ for all x. □

5.7 Stopping times and the strong Markov property

Some more notation. We replace the σ-algebras $\mathcal{F}_s^+$ and $\mathcal{F}_s^\circ$ with σ-algebras that reflect our new understanding from the 0-1 law. Define

$$\begin{aligned}
\mathcal{N}_x &= \{A \subset C \,:\, \exists D \in \Sigma \text{ s.t. } A \subseteq D \text{ and } \mathbf{P}_x(D) = 0\} \\
\mathcal{F}_s^x &= \sigma(\mathcal{F}_s^+ \vee \mathcal{N}_x) \\
\mathcal{F}_s &= \bigcap_{x\in\mathbb{R}} \mathcal{F}_s^x.
\end{aligned}$$

Exercise 5.26. *Check that $(\mathcal{F}_s)_{s\geq 0}$ is a right-continuous filtration.*

We now define the important concept of a **stopping time**. A r.v. $S : C \to [0, \infty]$ (note: the value ∞ is allowed!) is called a stopping time for BM if for all $t \geq 0$, the event $\{S < t\}$ is in the σ-algebra $\mathcal{F}_t$. An equivalent condition is that for all $t \geq 0$, the event $\{S \leq t\} \in \mathcal{F}_t$.

Lemma 5.27. *The above two conditions are indeed equivalent.*

Proof. This follows immediately from the relations

$$\{S < t\} = \bigcup_{n=1}^{\infty} \left\{ S \le t - \frac{1}{n} \right\},$$
$$\{S \le t\} = \bigcap_{n=1}^{\infty} \left\{ S < t + \frac{1}{n} \right\},$$

(where we also use the fact that the family of σ-algebras $\{\mathcal{F}_t : t \ge 0\}$ is increasing and right-continuous). □

Theorem 5.28. *1. If $G \subset \mathbb{R}$ is an open set, then $T = \inf\{t \ge 0 : B_t \in G\}$ is a stopping time.*

2. If $(T_n)_{n\ge1}$ is a sequence of stopping times and $T_n \downarrow T$ a.s., then T is a stopping time.

3. If $(T_n)_{n\ge1}$ is a sequence of stopping times and $T_n \uparrow T$ a.s., then T is a stopping time.

4. If $K \subset \mathbb{R}$ is a closed set, then $T = \inf\{t \ge 0 : B_t \in K\}$ is a stopping time.

5. If S, T are stopping times, then so are $S \vee T$, $S \wedge T$, $S + T$.

6. If S is a stopping time and $t \ge 0$, then $S \vee t$, $S \wedge t$, $S + t$ are also stopping times.

If $(T_n)_{n\ge1}$ is a sequence of stopping times, then $\sup_n T_n$, $\inf_n T_n$, $\limsup_n T_n$ and $\liminf_n T_n$ are all stopping times.

Proof. Proof of 1. By a.s. continuity of $(B_t)_{t\ge0}$, we have $\{T < t\} = \bigcup_{q<t, q\in\mathbb{Q}} \{B_q \in G\}$.

Proof of 2. $\{T < t\} = \bigcup_{n\ge1} \{T_n < t\}$.

Proof of 3. $\{T \le t\} = \bigcap_{n\ge1} \{T_n < t\}$.

Proof of 4. Define sets G_n, $n = 1, 2, \ldots$, by

$$G_n = \left\{ y \in \mathbb{R} : |y - x| < \frac{1}{n} \text{ for some } x \in K \right\}$$

(the $1/n$-dilation of K). G_n is an open set It is easy to check that $T_n \uparrow T$, where $T = \inf\{t \ge 0 : B_t \in G_n\}$, so the claim follows from part 3 of the theorem.

Proof of 5, 6, 7. Left as an exercise. □

Given a stopping time S, we can associate with it a "shift by the random amount S" operator $\theta_S : C \to C \cup \{\Delta\}$, defined by

$$\theta_S(\omega)(t) = \begin{cases} \omega(S(\omega) + t) & \text{if } S(\omega) < \infty, \\ \Delta & \text{if } S(\omega) = \infty. \end{cases}$$

Here, Δ is an extra symbol we add to the sample space C, corresponding to a kind of "undefined" value. We also define a σ-algebra $\mathcal{F}_S$ of "information about BM available at time S", by

$$\mathcal{F}_S = \{A \in \Sigma \,:\, A \cap \{S \le t\} \in \mathcal{F}_t \text{ for all } t \ge 0\}.$$

Theorem 5.29. *1. $\mathcal{F}_S$ is a σ-algebra, and in its definition it does not matter if we write $\{S \le t\}$ or $\{S < t\}$.*

2. The stopping time S is $\mathcal{F}_S$-measurable.

3. If S, T are stopping times then the events $\{S < t\}$, $\{S > t\}$, $\{S = t\}$ are in $\mathcal{F}_S$, and the events $\{S < T\}$, $\{S > T\}$, $\{S = T\}$ are in $\mathcal{F}_S \cap \mathcal{F}_T$.

4. If S, T are stopping times and $S \le T$, then $\mathcal{F}_S \subseteq \mathcal{F}_T$.

5. If $(T_n)_{n\ge1}$ are stopping times and $T_n \downarrow T$ a.s., then $\mathcal{F}_T = \bigcap_{n\ge1} \mathcal{F}_{T_n}$.

6. If S is a stopping time, then B_S is $\mathcal{F}_S$-measurable.

Proof. Parts 1 and 2 are left as an exercise.

Proof of 3. If $A \in \mathcal{F}_S$ then for any $t \ge 0$,

$$A \cap \{T \le t\} = (A \cap \{S \le t\}) \cap \{T \le t\} \in \mathcal{F}_t.$$

$(A \cap \{S \le t\}) \in \mathcal{F}_t$ since $A \in \mathcal{F}_S$, and $\{T \le t\} \in \mathcal{F}_t$ since T is a stopping time.) So $A \in \mathcal{F}_T$.

Proof of 4. Since $T \le T_n$, by part 3 above we have $\mathcal{F}_T \subset \mathcal{F}_{T_n}$ for all $n \ge 1$, so $\mathcal{F}_T \subseteq \bigcap_{n\ge1} \mathcal{F}_{T_n}$. On the other hand, if $A \in \bigcap_{n\ge1} \mathcal{F}_{T_n}$, then $A \cap \{T_n < t\} \in \mathcal{F}_t$ for any $n \ge 1$, so also $A \cap \{T < t\} = \bigcap_{n\ge1}(A \cap \{T_n < t\}) \in \mathcal{F}_t$. So $A \in \mathcal{F}_T$.

Proof of 5. Define discrete approximations $(S_n)_{n=1}^\infty$ to S by $S_n = \frac{1}{2^n}(\lfloor 2^n S \rfloor + 1)$. Then S_n are stopping times, and $S_n \downarrow S$ a.s. Now we leave it as an exercise to show that B_{S_n} is $\mathcal{F}_{S_n}$-measurable for all n. This is enough, since then $B_S = \lim_{n\to\infty} B_{S_n}$ is measurable with respect to $\bigcap_{n\ge1} \mathcal{F}_{S_n}$, which (by part 4 above) is equal to $\mathcal{F}_S$. □

Theorem 5.30 (The strong Markov property)**.** *Let $(s,\omega) \to Y_s(\omega)$ be a function from $[0,\infty) \times C to \mathbb{R}$ which is bounded and measurable (with respect to the product σ-algebra $\mathcal{B} \times \Sigma$, where $\mathcal{B}$ is the Borel σ-algebra on $\mathbb{R}$). Let S be a stopping time. For any $x \in \mathbb{R}$ we have*

$$\mathbf{E}_x\Big(T_S \circ \theta_S \,\Big|\, \mathcal{F}_S\Big) = \mathbf{E}_{B_S}(Y_S) \quad a.s.\ on\ \{S < \infty\}.$$

Here, the quantity on the right-hand side is to be interpreted as the function $\varphi(x,t) = \mathbf{E}_x Y_t$ evaluated at $x = B_S$, $t = S$.

Proof. The random variable $\varphi(B_S, S) = \mathbf{E}_{B_S} Y_S$ is $\mathcal{F}_S$-measurable. So what we need to show is that for any $A \in \mathcal{F}_S$ and $x \in \mathbb{R}$, the equation

$$\mathbf{E}_x\Big(Y_S \circ \theta_S \cdot \mathbf{1}_{A \cap \{S<\infty\}}\Big) = \mathbf{E}_x\Big(\left(\mathbf{E}_{B_S} Y_S\right) \mathbf{1}_{A \cap \{S<\infty\}}\Big)$$

holds. We first prove the result in the case when S is a discrete r.v. taking values in some discrete set $\{t_1, t_2, \ldots\} \cup \{\infty\}$. Denote $Z_n = Y_{t_n}$. If $A \in \mathcal{F}_S$, then

$$\mathbf{E}_x\Big(Y_S \circ \theta_S \cdot \mathbf{1}_{A \cap \{S<\infty\}}\Big) = \sum_{n=1}^{\infty} \mathbf{E}_x\Big(Z_n \circ \theta_{t_n} \cdot \mathbf{1}_{A \cap \{S=t_n\}}\Big)$$

Note that $A \cap \{T = t_n\} \in \mathcal{F}_{t_n}$. Therefore by the Markov property (Theorem 5.16), this is equal to

$$\sum_{n=1}^{\infty} \mathbf{E}_x\Big(\left(\mathbf{E}_{B_{t_n}} Z_n\right) \mathbf{1}_{A \cap \{S=t_n\}}\Big) = \mathbf{E}_x\Big(\left(\mathbf{E}_{B_S} Y_S\right) \mathbf{1}_{A \cap \{S<\infty\}}\Big),$$

which is what we wanted.

Next, for the case of a general stopping time S, define discrete approximations S_n to S by

$$S = \frac{1}{2^n}\left(\lfloor 2^n S \rfloor + 1\right).$$

We then have that $S_n \downarrow S$ almost surely as $n \to \infty$. Consider functions $Y_s(\omega)$ which are of the form

$$Y_s(\omega) = f_0(s) \prod_{m=1}^{n} f_m(\omega(t_m)),$$

where $0 < t_1 < \ldots < t_n$ and $f_0, \ldots, f_m : \mathbb{R} \to \mathbb{R}$ are bounded continuous functions. In this case, we have by Lemma 5.15 that

$$\varphi(x,s) = \mathbf{E}_x Y_s = f_0(s) \int_{\mathbb{R}} p_{t_1}(x, y_1) f_1(y_1)\, dy_1 \int_{\mathbb{R}} p_{t_2 - t_1}(t_1, y_2) f_2(y_2)\, dy_2 \\ \cdots \int_{\mathbb{R}} p_{t_n - t_{n-1}}(y_{n-1}, y_n) f_n(y_n)\, dy_n.$$

Clearly $\varphi(x,s)$ is bounded and continuous. Let $A \in \mathcal{F}_S$. Since $S \leq S_n$, the event A is also in $\mathcal{F}_{S_n}$. Applying the special case of discrete stopping times proved above, we have

$$\mathbf{E}_x\Big(Y_{S_n} \circ \theta_{S_n} \cdot \mathbf{1}_{A\cap\{S<\infty\}}\Big) = \mathbf{E}_x\Big(\varphi(B_{S_n}, S_n) \cdot \mathbf{1}_{A\cap\{S<\infty\}}\Big).$$

(Note that $S < \infty$ if and only if $S_n < \infty$.) Letting $n \to \infty$, we have that $B_{S_n} \to B_S$, $S_n \downarrow S$, $\varphi(B_{S_n}, S_n) \to \varphi(B_S, S)$, and $Y_{S_n} \circ \theta_{S_n} \to Y_S \circ \theta_S$ on the event $\{S < \infty\}$. So, by the bounded convergence theorem, we get that

$$\mathbf{E}_x\Big(Y_S \circ \theta_S \cdot \mathbf{1}_{A\cap\{S<\infty\}}\Big) = \mathbf{E}_x\Big(\varphi(B_S, S) \cdot \mathbf{1}_{A\cap\{S<\infty\}}\Big),$$

as claimed.

It remains to extend the validity of the result to general measurable bounded functions $Y_s(\omega)$. This follows from an application of the monotone class theorem (see [4, p. ?] for the details). □

5.8 Applications of the strong Markov property

Corollary 5.31. *Let S be a stopping time with $\mathbf{P}_X(S < \infty) = 1$. Under the measure $\mathbf{P}_x$, the process $(D_t)_{t\geq 0} := (B_{S+t} - B_S)_{t\geq 0}$ is a standard BM which is independent of $\mathcal{F}_S$.*

Proof. Let $g : \mathbb{R}^n \to \mathbb{R}$ be bounded and measurable, and let $0 < t_1 < \ldots < t_n$. Then we have

$$g(D_{t_1}, \ldots, D_{t_n}) = g(B_{t_1} - B_0, B_{t_2} - B_0, \ldots, B_{t_n} - B_0) \circ \theta_S =: Y \circ \theta_S.$$

By the strong Markov property, for any $A \in \mathcal{F}_S$,

$$\mathbf{E}_x\Big(g(D_{t_1}, \ldots, D_{t_n})\mathbf{1}_A\Big) = \mathbf{E}_x\Big[\mathbf{E}_z\Big(g(B_{t_1} - B_0, \ldots, B_{t_n} - B_0)\Big|_{z=B_S} \cdot \mathbf{1}_A\Big)\Big].$$

But $\mathbf{E}_z\Big(g(B_{t_1} - B_0, \ldots, B_{t_n} - B_0)$ is independent of z, since under $|prob_z$, $(B_t - B_0)_{t\geq 0}$ is a standard BM (Lemma 5.7). Thus, we have

$$\mathbf{E}_x\Big(g(D_{t_1}, \ldots, D_{t_n})\mathbf{1}_A\Big) = \mathbf{E}_0\Big(g(D_{t_1}, \ldots, D_{t_n})\mathbf{1}_A\Big) = \mathbf{E}_0\Big(g(D_{t_1}, \ldots, D_{t_n})\Big)\mathbf{P}_x(B_0 \in A).$$

The claims follow. □

Next, we consider the **hitting times**

$$T_a = \inf\{t \geq 0 \,:\, B_t = a\} \qquad (a \in \mathbb{R}),$$

under the measure $\mathbf{P}_0$. We already saw that T_a is an a.s. finite stopping time. It turns out to be very interesting to consider $(T_a)_{a\geq 0}$ as a stochastic process, with the parameter a playing the role of "time". Note that under $\mathbf{P}_0$, $T_0 = 0$ almost surely — the process starts from 0 — and the process is increasing in a.

Theorem 5.32. *Under $\mathbf{P}_0$, the process $(T_a)_{a\geq 0}$ has stationary independent increments. That is, for any $0 \leq a_1 < a_2 < \ldots < a_n$, the r.v.s*

$$T_{a_1}, T_{a_2} - Ta_1, \ldots, T_{a_n} - T_{a_{n-1}}$$

are independent, and for any $0 \leq a < b$,

$$T_b - T_a \stackrel{\mathcal{D}}{=} T_{b-a}$$

(that is, the distribution of an increment depends only on the length of the interval between the two times for which the incremenet is measured).

Proof. For $0 < a < b$, we have $T_b \circ \theta_{T_a} = T_b - T_a$. (Exercise: check the more general fact that for any a.s. finite stopping times $S \leq T$, the relation $T \circ \theta_S = T - S$ holds.) If $f : \mathbb{R} \to \mathbb{R}$ is bounded and measurable, by the strong Markov property we have:

$$\mathbf{E}_0\Big(f(T_b - T_a)\Big|\mathcal{F}_{T_a}\Big) = \mathbf{E}_0\Big(f(T_b)\circ\theta_{T_a}\Big|\mathcal{F}_{T_a}\Big) = \mathbf{E}_{B_{T_a}} f(T_b) = \mathbf{E}_a f(T_b) = \mathbf{E}_0 f(T_{b-a}).$$

This shows that $T_b - T_a$ is independent of $\mathcal{F}_{T_a}$ (see the proof of Corollary 5.31 above — the idea is similar), and equal in distribution to T_{b-a}. It follows easily by induction that for any numbers $0 < a_1 < \ldots < a_n$ and bounded and measurable functions $f_1, \ldots, f_n : \mathbb{R} \to \mathbb{R}$, we have

$$\mathbf{E}_0\left(\prod_{j=1}^n f_j(T_{a_j} - T_{a_{j-1}})\right) = \prod_{j=1}^n \mathbf{E}_0 f_j(T_{a_j - a_{j-1}})$$

(where we denote $a_0 = 0$). This proves the claim. □

Lemma 5.33 (scaling)**.** *We have the equalities in distribution*

$$T_a \stackrel{\mathcal{D}}{=} a^2 T_1 \qquad \text{for any } a > 0,$$
$$(T_{ca})_{a\geq 0} \stackrel{\mathcal{D}}{=} (c^2 T_a)_{a\geq 0} \qquad \text{for any } c > 0.$$

Proof. Define $D_t = s^{-1/2} B_{st}$, $t \geq 0$ for some fixed $s > 0$. By the scaling relation of Brownian motion (Lemma 5.8), $(D_t)_{t \geq 0} \stackrel{\mathcal{D}}{=} (B_t)_{t \geq 0}$ is a standard BM. So, if we define

$$U_a = \inf\{t \geq 0 \,:\, D_t = a\},$$

then $(U_a)_{a \geq 0} \stackrel{\mathcal{D}}{=} (T_a)_{a \geq 0}$. On the other hand,

$$U_a = \inf\{t \geq 0 \,:\, B_{st} = s^{1/2} a\} = \frac{1}{s} \inf\{u \geq 0 \,:\, B_u = s^{1/2} a\} = \frac{1}{s} T_{s^{1/2} a}.$$

Taking $c = s^{1/2}$ we get that $(T_a)_{a \geq 0} = (U_a)_{a \geq 0} = (c^{-2} T_{ca})_{a \geq 0}$, which proves the second claim. The first claim follows by taking $c = 1/a$. □

Note that for integer $n \geq 1$, if we let $X_1, X_2, \ldots$ be i.i.d. copies of T_1, then by Theorem 5.32, we have the equality in distribution $T_n \stackrel{\mathcal{D}}{=} \sum_{k=1}^n X_k$. On the other hand, by Lemma 5.33, $T_n \stackrel{\mathcal{D}}{=} n^2 T_1$. So we see that the distribution of T_1 has the property that a sum of n i.i.d. copies of it is distributed as n^2 times the original random variable:

$$\sum_{k=1}^n X_k \stackrel{\mathcal{D}}{=} n^2 X_1.$$

A distribution with this property is called **a stable distribution with index** 2. More generally, a distribution with the property that

$$\sum_{k=1}^n Y_k \stackrel{\mathcal{D}}{=} n^\alpha Y_1,$$

where $Y_1, Y_2, \ldots$ is an i.i.d. sequence with that distribution, is called **a stable distribution with index** α. (The two most elementary examples of this property are the standard normal distribution, which is stable with index $1/2$, and the Cauchy distribution, which is stable distribution with index 1.)

The random variables T_a have a simple explicit formula for their density functions.

Proposition 5.34. *The probability density function of T_a is given by*

$$\frac{a}{\sqrt{2\pi t^3}} e^{-a^2/2t}, \qquad t \in \mathbb{R}.$$

The distribution of the r.v. T_a is called the **Lévy distribution**.

Proposition 5.34 is an easy corollary of the following result.

Theorem 5.35 (The reflection principle)**.** *For $a > 0$ and $t \geq 0$, we have the relation*

$$\mathbf{P}_0(T_a < t) = 2\mathbf{P}_0(B_t > a).$$

Proof.

$$\begin{aligned}
\mathbf{P}_0(B_t > a) &= \mathbf{P}_0(B_t > a, T_a < t) = \mathbf{P}_0(T_a < t)\mathbf{P}_0(B_t > a \,|\, T_a < t) \\
&= \mathbf{P}_0(T_a < t)\mathbf{P}_0(B_{T_a} + (B_t - B_{T_a}) > a \,|\, T_a < t) \\
&= \mathbf{P}_0(T_a < t)\mathbf{P}_0(a + B_t - B_{T_a} > a \,|\, T_a < t) \\
&= \mathbf{P}_0(T_a < t)\mathbf{P}_0(B_t - B_{T_a} > 0 \,|\, T_a < t)
\end{aligned}$$

Thus the claim is equivalent to the statement that $\mathbf{P}_0(B_t - B_{T_a} > 0 \,|\, T_a < t) = \frac{1}{2}$. This is at least intuitively obvious. To prove it rigorously, we use the strong Markov property. Let

$$S = \begin{cases} \inf\left\{0 \leq s < t \,:\, B_s = a\right\} & \text{if } T_a \leq t, \\ \infty & \text{otherwise.} \end{cases}$$

and define further

$$Y_s(\omega) = \begin{cases} 1 & \text{if } s \leq t \text{ and } \omega(t-s) > a, \\ 0 & \text{otherwise.} \end{cases}$$

Then we have that

$$(Y_S \circ \theta_S)(\omega) = \begin{cases} 1 & \text{if } S < t \text{ and } B_t > a, \\ 0 & \text{otherwise} \end{cases} = \mathbf{1}_{\{S<t, B_t>a\}}.$$

The strong Markov property therefore gives

$$\mathbf{E}_0\Big(Y_S \circ \theta_S \,\Big|\, \mathcal{F}_S\Big) = \varphi(B_S, S) \quad \text{on } \{S < \infty\} = \{T_a \leq t\},$$

where $\varphi(a, s) = \mathbf{P}_a(B_{t-s} > a) = \frac{1}{2}$. But note that on $\{S < \infty\}$, $S < t$ and $B_S = a$, so that $\varphi(B_S, S) = \frac{1}{2}$. Thus, we have shown that

$$\mathbf{P}_0(T_a < t, B_t > a) = \mathbf{E}_0\Big(Y_S \circ \theta_S \mathbf{1}_{\{S<\infty\}}\Big) = \mathbf{E}_0\Big(\frac{1}{2}\mathbf{1}_{\{S<\infty\}}\Big) = \frac{1}{2}\mathbf{P}_0(S < \infty) = \frac{1}{2}\mathbf{P}_0(T_a < t),$$

as claimed. □

Proof of Proposition 5.34.

$$\begin{aligned}\mathbf{P}_0(T_a \le t) &= 2\mathbf{P}_0(B_t \ge a) = 2\int_a^\infty \frac{1}{\sqrt{2\pi t}} e^{-x^2/2t}\,dx \\ &= 2\int_t^0 \frac{1}{\sqrt{2\pi t}} \exp\left(-\frac{a^2}{2s}\right)\left(-\frac{1}{2}\frac{\sqrt{t}a}{s^{3/2}}\right) ds = \int_0^t \frac{a}{\sqrt{2\pi s^3}} \exp\left(-\frac{a^2}{2s}\right) ds.\end{aligned}$$

□

Exercise 5.36. *1. Generalize the proof of Theorem 5.35 to conclude that if $u \le v \le a$ then the relation*

$$\mathbf{P}_0(T_a < t, u < B_t < v) = \mathbf{P}_0(2a - v < B_t < 2a - u)$$

holds.

2. Let $M_t = \max_{0\le s\le t} B_s$. Use the above result to show that the joint density of M_t, B_t is given by the formula

$$f_{M_t,B_t}(a,x) = \frac{2(2a-x)}{\sqrt{2\pi x^3}} \exp\left(-\frac{(2a-x)^2}{2t}\right).$$

Theorem 5.37 (The arcsine law)**.** *Define $L = \sup\{0 \le t \le 1 : B_t = 0\}$ (the last return time of BM to 0 in $[0,1]$). The random variable has the arcsine distribution, that is, its density is given by*

$$f_L(t) = \frac{1}{\pi}\frac{1}{\sqrt{t(1-t)}}.$$

(Equivalently, $L \sim \text{Beta}\left(\frac{1}{2},\frac{1}{2}\right)$, or $F_L(x) = \frac{1}{\pi}\arcsin(\sqrt{x})$.)

Proof. Let $T_0 = \inf\{s > 0 : B_s = 0\}$. We have (using the exercise on page ?)

$$\begin{aligned}\mathbf{P}_0(L \le s) &= \int_{-\infty}^\infty p_s(0,x)\mathbf{P}_x(T_0 > 1-s)\,dx \\ &= 2\int_0^\infty \frac{1}{\sqrt{2\pi s}} \exp\left(-\frac{x^2}{2s}\right)\int_{1-s}^\infty \frac{1}{\sqrt{2\pi r^2}} x \exp\left(-\frac{x^2}{2r}\right) dr\,dx \\ &= \frac{1}{\pi}\int_{1-s}^\infty \frac{1}{\sqrt{sr^3}}\int_0^\infty x\exp\left(-x^2(r+s)(2rs)\right) dx\,dr \\ &= \frac{1}{\pi}\int_{1-s}^\infty \frac{1}{\sqrt{sr^3}}\frac{rs}{r+s}\,dr = \frac{1}{\pi}\int_{1-s}^\infty \left(\frac{(r+s)^2}{rs}\right)^{1/2}\frac{s}{(r+s)^2}\,dr.\end{aligned}$$

Applying the change of variables $t = s/(r+s)$ in this last integral, we arrive at the formula $\frac{1}{\pi}\int_0^s \frac{1}{\sqrt{t(1-t)}}\,dt$ for $\mathbf{P}_0(L \le s)$, which was the claim. □

Exercise 5.38. *Let* $R = \inf\{t > 1 : B_t = 0\}$. *Show that under* $\mathbf{P}_0$, *the r.v.* R *has probability density*

$$f_R(t) = \frac{1}{\sqrt{\pi t(t-1)}} \qquad (t > 1).$$

****TO BE CONTINUED****

References

[1] P. Billingsley. Ergodic Theory and Information. John Wiley & Sons, 1965.

[2] R. Bowen. Invariant measures for Markov maps of the interval. *Commun. Math. Phys.* 69 (1979).

[3] G. D. Birkhoff. What is the ergodic theorem? *American Math. Monthly* 49 (1942), 222–226.

[4] R. Durrett. Probability: Theory and Examples, 5th Ed. Cambridge University Press, 2019.

[5] D. E. Knuth. The Art of Computer Programming, Vol. II: Seminumerical Algorithms, 3rd. Ed. Addison-Wesley Professional, 1997.

[6] J. C. Lagarias. The $3x + 1$ problem and its generalizations. *American Math. Monthly* 92 (1985), 3–23.

[7] G. F. Lawler, O. Schramm, W. Werner. The dimension of the planar Brownian frontier is 4/3. *Math. Res. Lett.* 8 (2001), 401–411.

[8] K. R. Matthews, A. M. Watts. A generalization of Hasse's generalization of the Syracuse algorithm. *Acta Arith.* 43 (1983), 75–83.

[9] P. Mörters, Y. Peres. Brownian Motion. Cambridge University Press, 2010.

[10] D. Romik. Sharp entropy bounds for discrete statistical simulation. *Stat. Probab. Lett.* 42 (1999), 219–227.

[11] D. Romik. Math 235B — Probability Theory Lecture Notes, Winter 2011. Available at `https://www.math.ucdavis.edu/~romik/data/uploads/notes/lecturenotes235b.pdf`. Online resource, accessed March 23, 2022.