Markov Chains and Mixing Times

David A. Levin

Yuval Peres

Elizabeth L. Wilmer

University of Oregon
E-mail address: dlevin@uoregon.edu
URL: http://www.uoregon.edu/~dlevin

Microsoft Research, University of Washington *and* UC Berkeley
E-mail address: peres@microsoft.com
URL: http://research.microsoft.com/~peres/

Oberlin College
E-mail address: elizabeth.wilmer@oberlin.edu
URL: http://www.oberlin.edu/math/faculty/wilmer.html

CHAPTER 1

Introduction to Finite Markov Chains

1.1. Finite Markov Chains

A finite Markov chain is a process which moves among the elements of a finite set Ω in the following manner: when at $x \in \Omega$, the next position is chosen according to a fixed probability distribution $P(x, \cdot)$. More precisely, a sequence of random variables $(X_0, X_1, \ldots)$ is a ***Markov chain with state space*** Ω ***and transition matrix*** P if for all $x, y \in \Omega$, all $t \geq 1$, and all events $H_{t-1} = \bigcap_{s=0}^{t-1} \{X_s = x_s\}$ satisfying $\mathbf{P}(H_{t-1} \cap \{X_t = x\}) > 0$, we have

$$\mathbf{P}\{X_{t+1} = y \mid H_{t-1} \cap \{X_t = x\}\} = \mathbf{P}\{X_{t+1} = y \mid X_t = x\} = P(x, y). \tag{1.1}$$

Equation (1.1), often called the ***Markov property***, means that the conditional probability of proceeding from state x to state y is the same, no matter what sequence $x_0, x_1, \ldots, x_{t-1}$ of states precedes the current state x. This is exactly why the $|\Omega| \times |\Omega|$ matrix P suffices to describe the transitions.

The x-th row of P is the distribution $P(x, \cdot)$. Thus P is ***stochastic***, that is, its entries are all non-negative and

$$\sum_{y \in \Omega} P(x, y) = 1 \qquad \text{for all } x \in \Omega.$$

EXAMPLE 1.1. A certain frog lives in a pond with two lily pads, *east* and *west*. A long time ago, he found two coins at the bottom of the pond and brought one up to each lily pad. Every morning, the frog decides whether to jump by tossing the current lily pad's coin. If the coin lands heads up, the frog jumps to the other lily pad. If the coin lands tails up, he remains where he is.

Let $\Omega = \{e, w\}$, and let $(X_0, X_1, \ldots)$ be the sequence of lily pads occupied by the frog on Sunday, Monday, ... Given the source of the coins, we should not assume that they are fair! Say the coin on the east pad has probability p of landing

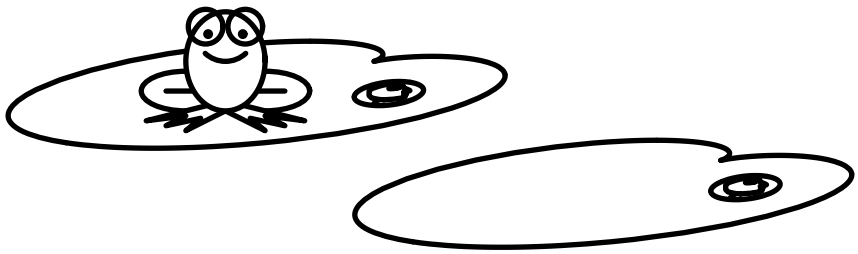
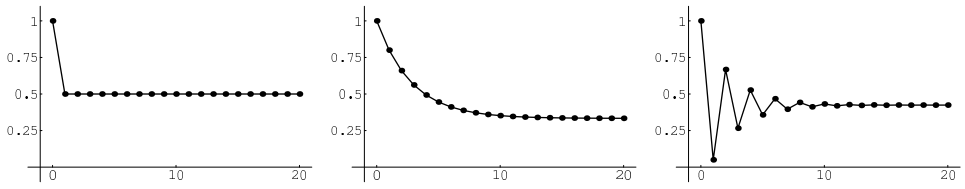


FIGURE 1.1. A randomly jumping frog. Whenever he tosses heads, he jumps to the other lily pad.



(a) (b) (c)

FIGURE 1.2. The probability of being on the east pad (started from the east pad) plotted versus time for (a) $p = q = 1/2$, (b) $p = 0.2$ and $q = 0.1$, (c) $p = 0.95$ and $q = 0.7$. The long-term limiting probabilities are $1/2$, $1/3$, and $14/33 \approx 0.42$, respectively.

heads up, while the coin on the west pad has probability q of landing heads up. The frog's rules for jumping imply that if we set

$$P = \begin{pmatrix} P(e,e) & P(e,w) \\ P(w,e) & P(w,w) \end{pmatrix} = \begin{pmatrix} 1-p & p \\ q & 1-q \end{pmatrix}, \tag{1.2}$$

then $(X_0, X_1, \dots)$ is a Markov chain with transition matrix P. Note that the first row of P is the conditional distribution of X_{t+1} given that $X_t = e$, while the second row is the conditional distribution of X_{t+1} given that $X_t = w$.

Assume that the frog spends Sunday on the east pad. When he awakens Monday, he has probability p of moving to the west pad and probability $1-p$ of staying on the east pad. That is,

$$\mathbf{P}\{X_1 = e \mid X_0 = e\} = 1 - p, \qquad \mathbf{P}\{X_1 = w \mid X_0 = e\} = p. \tag{1.3}$$

What happens Tuesday? By considering the two possibilities for X_1, we see that

$$\mathbf{P}\{X_2 = e \mid X_0 = e\} = (1-p)(1-p) + pq \tag{1.4}$$

and

$$\mathbf{P}\{X_2 = w \mid X_0 = e\} = (1-p)p + p(1-q). \tag{1.5}$$

While we could keep writing out formulas like (1.4) and (1.5), there is a more systematic approach. We can store our distribution information in a row vector

$$\mu_t := \left(\mathbf{P}\{X_t = e \mid X_0 = e\}, \mathbf{P}\{X_t = w \mid X_0 = e\}\right).$$

Our assumption that the frog starts on the east pad can now be written as $\mu_0 = (1, 0)$, while (1.3) becomes $\mu_1 = \mu_0 P$.

Multiplying by P on the right updates the distribution by another step:

$$\mu_t = \mu_{t-1} P \quad \text{for all } t \geq 1. \tag{1.6}$$

Indeed, for any initial distribution μ_0,

$$\mu_t = \mu_0 P^t \quad \text{for all } t \geq 0. \tag{1.7}$$

How does the distribution μ_t behave in the long term? Figure 1.2 suggests that μ_t has a limit π (whose value depends on p and q) as $t \to \infty$. Any such limit distribution π must satisfy

$$\pi = \pi P,$$

which implies (after a little algebra) that

$$\pi(e) = \frac{q}{p+q}, \qquad \pi(w) = \frac{p}{p+q}.$$

If we define

$$\Delta_t = \mu_t(e) - \frac{q}{p+q} \quad \text{for all } t \geq 0,$$

then by the definition of μ_{t+1} the sequence (Δ_t) satisfies

$$\Delta_{t+1} = \mu_t(e)(1-p) + (1-\mu_t(e))(q) - \frac{q}{p+q} = (1-p-q)\Delta_t. \tag{1.8}$$

We conclude that when $0 < p < 1$ and $0 < q < 1$,

$$\lim_{t\to\infty} \mu_t(e) = \frac{q}{p+q} \quad \text{and} \quad \lim_{t\to\infty} \mu_t(w) = \frac{p}{p+q} \tag{1.9}$$

for any initial distribution μ_0. As we suspected, μ_t approaches π as $t \to \infty$.

REMARK 1.2. The traditional theory of finite Markov chains is concerned with convergence statements of the type seen in (1.9), that is, with the rate of convergence as $t \to \infty$ for a *fixed chain*. Note that $1-p-q$ is an eigenvalue of the frog's transition matrix P. Note also that this eigenvalue determines the rate of convergence in (1.9), since by (1.8) we have

$$\Delta_t = (1-p-q)^t \Delta_0.$$

The computations we just did for a two-state chain generalize to any finite Markov chain. In particular, the distribution at time t can be found by matrix multiplication. Let $(X_0, X_1, \dots)$ be a finite Markov chain with state space Ω and transition matrix P, and let the row vector μ_t be the distribution of X_t:

$$\mu_t(x) = \mathbf{P}\{X_t = x\} \quad \text{for all } x \in \Omega.$$

By conditioning on the possible predecessors of the $(t+1)$-st state, we see that

$$\mu_{t+1}(y) = \sum_{x\in\Omega} \mathbf{P}\{X_t = x\}P(x,y) = \sum_{x\in\Omega} \mu_t(x)P(x,y) \quad \text{for all } y \in \Omega.$$

Rewriting this in vector form gives

$$\mu_{t+1} = \mu_t P \quad \text{for } t \geq 0$$

and hence

$$\mu_t = \mu_0 P^t \quad \text{for } t \geq 0. \tag{1.10}$$

Since we will often consider Markov chains with the same transition matrix but different starting distributions, we introduce the notation $\mathbf{P}_\mu$ and $\mathbf{E}_\mu$ for probabilities and expectations given that $\mu_0 = \mu$. Most often, the initial distribution will be concentrated at a single definite starting state x. We denote this distribution by δ_x:

$$\delta_x(y) = \begin{cases} 1 & \text{if } y = x, \\ 0 & \text{if } y \neq x. \end{cases}$$

We write simply $\mathbf{P}_x$ and $\mathbf{E}_x$ for $\mathbf{P}_{\delta_x}$ and $\mathbf{E}_{\delta_x}$, respectively.

These definitions and (1.10) together imply that

$$\mathbf{P}_x\{X_t = y\} = (\delta_x P^t)(y) = P^t(x,y).$$

FIGURE 1.3. Random walk on $\mathbb{Z}_{10}$ is periodic, since every step goes from an even state to an odd state, or vice-versa. Random walk on $\mathbb{Z}_9$ is aperiodic.

That is, the probability of moving in t steps from x to y is given by the (x, y)-th entry of P^t. We call these entries the ***t-step transition probabilities***.

NOTATION. A probability distribution μ on Ω will be identified with a row vector. For any event $A \subset \Omega$, we write

$$\pi(A) = \sum_{x \in A} \mu(x).$$

For $x \in \Omega$, the row of P indexed by x will be denoted by $P(x, \cdot)$.

REMARK 1.3. The way we constructed the matrix P has forced us to treat distributions as row vectors. In general, if the chain has distribution μ at time t, then it has distribution μP at time $t + 1$. *Multiplying a row vector by P on the right takes you from today's distribution to tomorrow's distribution.*

What if we multiply a column vector f by P on the left? Think of f as a function on the state space Ω (for the frog of Example 1.1, we might take $f(x)$ to be the area of the lily pad x). Consider the x-th entry of the resulting vector:

$$Pf(x) = \sum_y P(x, y) f(y) = \sum_y f(y) \mathbf{P}_x\{X_1 = y\} = \mathbf{E}_x(f(X_1)).$$

That is, the x-th entry of Pf tells us the expected value of the function f at tomorrow's state, given that we are at state x today. *Multiplying a column vector by P on the left takes us from a function on the state space to the expected value of that function tomorrow.*

1.2. Random Mapping Representation

We begin this section with an example.

EXAMPLE 1.4 (Random walk on the n-cycle). Let $\Omega = \mathbb{Z}_n = \{0, 1, \ldots, n-1\}$, the set of remainders modulo n. Consider the transition matrix

$$P(j, k) = \begin{cases} 1/2 & \text{if } k \equiv j + 1 \pmod{n}, \\ 1/2 & \text{if } k \equiv j - 1 \pmod{n}, \\ 0 & \text{otherwise.} \end{cases} \tag{1.11}$$

The associated Markov chain (X_t) is called ***random walk on the n-cycle***. The states can be envisioned as equally spaced dots arranged in a circle (see Figure 1.3).

Rather than writing down the transition matrix in (1.11), this chain can be specified simply in words: at each step, a coin is tossed. If the coin lands heads up, the walk moves one step clockwise. If the coin lands tails up, the walk moves one step counterclockwise.

More precisely, suppose that Z is a random variable which is equally likely to take on the values -1 and $+1$. If the current state of the chain is $j \in \mathbb{Z}_n$, then the next state is $j + Z \bmod n$. For any $k \in \mathbb{Z}_n$,

$$\mathbf{P}\{(j + Z) \bmod n = k\} = P(j, k).$$

In other words, the distribution of $(j + Z) \bmod n$ equals $P(j, \cdot)$.

A ***random mapping representation*** of a transition matrix P on state space Ω is a function $f : \Omega \times \Lambda \to \Omega$, along with a Λ-valued random variable Z, satisfying

$$\mathbf{P}\{f(x, Z) = y\} = P(x, y).$$

The reader should check that if $Z_1, Z_2, \ldots$ is a sequence of independent random variables, each having the same distribution as Z, and X_0 has distribution μ, then the sequence $(X_0, X_1, \ldots)$ defined by

$$X_n = f(X_{n-1}, Z_n) \quad \text{for } n \geq 1$$

is a Markov chain with transition matrix P and initial distribution μ.

For the example of the simple random walk on the cycle, setting $\Lambda = \{1, -1\}$, each Z_i uniform on Λ, and $f(x, z) = x + z \bmod n$ yields a random mapping representation.

PROPOSITION 1.5. *Every transition matrix on a finite state space has a random mapping representation.*

PROOF. Let P be the transition matrix of a Markov chain with state space $\Omega = \{x_1, \ldots, x_n\}$. Take $\Lambda = [0, 1]$; our auxiliary random variables $Z, Z_1, Z_2, \ldots$ will be uniformly chosen in this interval. Set $F_{j,k} = \sum_{i=1}^{k} P(x_j, x_i)$ and define

$$f(x_j, z) := x_k \text{ when } F_{j,k-1} < z \leq F_{j,k}.$$

We have

$$\mathbf{P}\{f(x_j, Z) = x_k\} = \mathbf{P}\{F_{j,k-1} < Z \leq F_{j,k}\} = P(x_j, x_k).$$

■

Note that, unlike transition matrices, random mapping representations are far from unique. For instance, replacing the function $f(x, z)$ in the proof of Proposition 1.5 with $f(x, 1 - z)$ yields a different representation of the same transition matrix.

Random mapping representations are crucial for simulating large chains. They can also be the most convenient way to describe a chain. We will often give rules for how a chain proceeds from state to state, using some extra randomness to determine where to go next; such discussions are implicit random mapping representations. Finally, random mapping representations provide a way to coordinate two (or more) chain trajectories, as we can simply use the same sequence of auxiliary random variables to determine updates. This technique will be exploited in Chapter 5, on coupling Markov chain trajectories, and elsewhere.

1.3. Irreducibility and Aperiodicity

We now make note of two simple properties possessed by most interesting chains. Both will turn out to be necessary for the Convergence Theorem (Theorem 4.9) to be true.

A chain P is called ***irreducible*** if for any two states $x, y \in \Omega$ there exists an integer t (possibly depending on x and y) such that $P^t(x,y) > 0$. This means that it is possible to get from any state to any other state using only transitions of positive probability. We will generally assume that the chains under discussion are irreducible. (Checking that specific chains are irreducible can be quite interesting; see, for instance, Section 2.6 and Example B.5. See Section 1.7 for a discussion of all the ways in which a Markov chain can fail to be irreducible.)

Let $\mathcal{T}(x) := \{t \geq 1 : P^t(x,x) > 0\}$ be the set of times when it is possible for the chain to return to starting position x. The ***period*** of state x is defined to be the greatest common divisor of $\mathcal{T}(x)$.

LEMMA 1.6. *If P is irreducible, then $\gcd \mathcal{T}(x) = \gcd \mathcal{T}(y)$ for all $x, y \in \Omega$.*

PROOF. Fix two states x and y. There exist non-negative integers r and ℓ such that $P^r(x,y) > 0$ and $P^\ell(y,x) > 0$. Letting $m = r+\ell$, we have $m \in \mathcal{T}(x) \cap \mathcal{T}(y)$ and $\mathcal{T}(x) \subset \mathcal{T}(y) - m$, whence $\gcd \mathcal{T}(y)$ divides all elements of $\mathcal{T}(x)$. We conclude that $\gcd \mathcal{T}(y) \leq \gcd \mathcal{T}(x)$. By an entirely parallel argument, $\gcd \mathcal{T}(x) \leq \gcd \mathcal{T}(y)$. ∎

For an irreducible chain, the period of the chain is defined to be the period which is common to all states. The chain will be called ***aperiodic*** if all states have period 1. If a chain is not aperiodic, we call it ***periodic***.

PROPOSITION 1.7. *If P is aperiodic and irreducible, then there is an integer r such that $P^r(x,y) > 0$ for all $x, y \in \Omega$.*

PROOF. We use the following number-theoretic fact: any set of non-negative integers which is closed under addition and which has greatest common divisor 1 must contain all but finitely many of the non-negative integers. (See Lemma 1.27 in the Notes of this chapter for a proof.) For $x \in \Omega$, recall that $\mathcal{T}(x) = \{t \geq 1 : P^t(x,x) > 0\}$. Since the chain is aperiodic, the gcd of $\mathcal{T}(x)$ is 1. The set $\mathcal{T}(x)$ is closed under addition: if $s, t \in \mathcal{T}(x)$, then $P^{s+t}(x,x) \geq P^s(x,x)P^t(x,x) > 0$, and hence $s + t \in \mathcal{T}(x)$. Therefore there exists a $t(x)$ such that $t \geq t(x)$ implies $t \in \mathcal{T}(x)$. By irreducibility we know that for any $y \in \Omega$ there exists $r = r(x,y)$ such that $P^r(x,y) > 0$. Therefore, for $t \geq t(x) + r$,

$$P^t(x,y) \geq P^{t-r}(x,x)P^r(x,y) > 0.$$

For $t \geq t'(x) := t(x) + \max_{y \in \Omega} r(x,y)$, we have $P^t(x,y) > 0$ for all $y \in \Omega$. Finally, if $t \geq \max_{x \in \Omega} t'(x)$, then $P^t(x,y) > 0$ for all $x, y \in \Omega$. ∎

Suppose that a chain is irreducible with period two, e.g. the simple random walk on a cycle of even length (see Figure 1.3). The state space Ω can be partitioned into two classes, say ***even*** and ***odd***, such that the chain makes transitions only between states in complementary classes. (Exercise 1.6 examines chains with period b.)

Let P have period two, and suppose that x_0 is an even state. The probability distribution of the chain after $2t$ steps, $P^{2t}(x_0, \cdot)$, is supported on even states, while the distribution of the chain after $2t+1$ steps is supported on odd states. It is evident that we cannot expect the distribution $P^t(x_0, \cdot)$ to converge as $t \to \infty$.

Fortunately, a simple modification can repair periodicity problems. Given an arbitrary transition matrix P, let $Q = \frac{I+P}{2}$ (here I is the $|\Omega| \times |\Omega|$ identity matrix). (One can imagine simulating Q as follows: at each time step, flip a fair coin. If it comes up heads, take a step in P; if tails, then stay at the current state.) Since $Q(x,x) > 0$ for all $x \in \Omega$, the transition matrix Q is aperiodic. We call Q a ***lazy version of*** P. It will often be convenient to analyze lazy versions of chains.

EXAMPLE 1.8 (The n-cycle, revisited). Recall random walk on the n-cycle, defined in Example 1.4. For every $n \geq 1$, random walk on the n-cycle is irreducible.

Random walk on any even-length cycle is periodic, since $\gcd\{t : P^t(x,x) > 0\} = 2$ (see Figure 1.3). Random walk on an odd-length cycle is aperiodic.

The transition matrix Q for lazy random walk on the n-cycle is

$$Q(j,k) = \begin{cases} 1/4 & \text{if } k \equiv j+1 \pmod{n}, \\ 1/2 & \text{if } k \equiv j \pmod{n}, \\ 1/4 & \text{if } k \equiv j-1 \pmod{n}, \\ 0 & \text{otherwise.} \end{cases} \tag{1.12}$$

Lazy random walk on the n-cycle is both irreducible and aperiodic for every n.

REMARK 1.9. Establishing that a Markov chain is irreducible is not always trivial; see Example B.5, and also Thurston (1990).

1.4. Random Walks on Graphs

Random walk on the n-cycle, which is shown in Figure 1.3, is a simple case of an important type of Markov chain.

A ***graph*** $G = (V, E)$ consists of a ***vertex set*** V and an ***edge set*** E, where the elements of E are unordered pairs of vertices: $E \subset \{\{x,y\} : x, y \in V, x \neq y\}$. We can think of V as a set of dots, where two dots x and y are joined by a line if and only if $\{x,y\}$ is an element of the edge set. When $\{x,y\} \in E$, we write $x \sim y$ and say that y is a ***neighbor*** of x (and also that x is a neighbor of y). The ***degree*** $\deg(x)$ of a vertex x is the number of neighbors of x.

Given a graph $G = (V, E)$, we can define ***simple random walk on*** G to be the Markov chain with state space V and transition matrix

$$P(x,y) = \begin{cases} \frac{1}{\deg(x)} & \text{if } y \sim x, \\ 0 & \text{otherwise.} \end{cases} \tag{1.13}$$

That is to say, when the chain is at vertex x, it examines all the neighbors of x, picks one uniformly at random, and moves to the chosen vertex.

EXAMPLE 1.10. Consider the graph G shown in Figure 1.4. The transition matrix of simple random walk on G is

$$P = \begin{pmatrix} 0 & \frac{1}{2} & \frac{1}{2} & 0 & 0 \\ \frac{1}{3} & 0 & \frac{1}{3} & \frac{1}{3} & 0 \\ \frac{1}{4} & \frac{1}{4} & 0 & \frac{1}{4} & \frac{1}{4} \\ 0 & \frac{1}{2} & \frac{1}{2} & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 \end{pmatrix}.$$

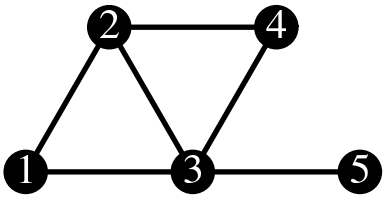


FIGURE 1.4. An example of a graph with vertex set $\{1, 2, 3, 4, 5\}$ and 6 edges.

REMARK 1.11. We have chosen a narrow definition of "graph" for simplicity. It is sometimes useful to allow edges connecting a vertex to itself, called ***loops***. It is also sometimes useful to allow multiple edges connecting a single pair of vertices. Loops and multiple edges both contribute to the degree of a vertex and are counted as options when a simple random walk chooses a direction. See Section 6.5.1 for an example.

We will have much more to say about random walks on graphs throughout this book—but especially in Chapter 9.

1.5. Stationary Distributions

1.5.1. Definition. We saw in Example 1.1 that a distribution π on Ω satisfying

$$\pi = \pi P \tag{1.14}$$

can have another interesting property: in that case, π was the long-term limiting distribution of the chain. We call a probability π satisfying (1.14) a ***stationary distribution*** of the Markov chain. Clearly, if π is a stationary distribution and $\mu_0 = \pi$ (i.e. the chain is started in a stationary distribution), then $\mu_t = \pi$ for all $t \geq 0$.

Note that we can also write (1.14) elementwise. An equivalent formulation is

$$\pi(y) = \sum_{x \in \Omega} \pi(x) P(x, y) \quad \text{for all } y \in \Omega. \tag{1.15}$$

EXAMPLE 1.12. Consider simple random walk on a graph $G = (V, E)$. For any vertex $y \in V$,

$$\sum_{x \in V} \deg(x) P(x, y) = \sum_{x \sim y} \frac{\deg(x)}{\deg(x)} = \deg(y). \tag{1.16}$$

To get a probability, we simply normalize by $\sum_{y \in V} \deg(y) = 2|E|$ (a fact the reader should check). We conclude that the probability measure

$$\pi(y) = \frac{\deg(y)}{2|E|} \quad \text{for all } y \in \Omega,$$

which is proportional to the degrees, is always a stationary distribution for the walk. For the graph in Figure 1.4,

$$\pi = \left(\tfrac{2}{12}, \tfrac{3}{12}, \tfrac{4}{12}, \tfrac{2}{12}, \tfrac{1}{12}\right).$$

If G has the property that every vertex has the same degree d, we call G ***d-regular***. In this case $2|E| = d|V|$ and the uniform distribution $\pi(y) = 1/|V|$ for every $y \in V$ is stationary.

A central goal of this chapter and of Chapter 4 is to prove a general yet precise version of the statement that "finite Markov chains converge to their stationary distributions." Before we can analyze the time required to be close to stationarity, we must be sure that it is finite! In this section we show that, under mild restrictions, stationary distributions exist and are unique. Our strategy of building a candidate distribution, then verifying that it has the necessary properties, may seem cumbersome. However, the tools we construct here will be applied in many other places. In Section 4.3, we will show that irreducible and aperiodic chains do, in fact, converge to their stationary distributions in a precise sense.

1.5.2. Hitting and first return times. Throughout this section, we assume that the Markov chain $(X_0, X_1, \dots)$ under discussion has finite state space Ω and transition matrix P. For $x \in \Omega$, define the ***hitting time*** for x to be

$$\tau_x := \min\{t \geq 0 : X_t = x\},$$

the first time at which the chain visits state x. For situations where only a visit to x at a positive time will do, we also define

$$\tau_x^+ := \min\{t \geq 1 \,:\, X_t = x\}.$$

When $X_0 = x$, we call τ_x^+ the ***first return time***.

LEMMA 1.13. *For any states x and y of an irreducible chain, $\mathbf{E}_x(\tau_y^+) < \infty$.*

PROOF. The definition of irreducibility implies that there exist an integer $r > 0$ and a real $\varepsilon > 0$ with the following property: for any states $z, w \in \Omega$, there exists a $j \leq r$ with $P^j(z, w) > \varepsilon$. Thus for any value of X_t, the probability of hitting state y at a time between t and $t + r$ is at least ε. Hence for $k > 0$ we have

$$\mathbf{P}_x\{\tau_y^+ > kr\} \leq (1 - \varepsilon)\mathbf{P}_x\{\tau_y^+ > (k-1)r\}. \tag{1.17}$$

Repeated application of (1.17) yields

$$\mathbf{P}_x\{\tau_y^+ > kr\} \leq (1 - \varepsilon)^k. \tag{1.18}$$

Recall that when Y is a non-negative integer-valued random variable, we have

$$\mathbf{E}(Y) = \sum_{t \geq 0} \mathbf{P}\{Y > t\}.$$

Since $\mathbf{P}_x\{\tau_y^+ > t\}$ is a decreasing function of t, (1.18) suffices to bound all terms of the corresponding expression for $\mathbf{E}_x(\tau_y^+)$:

$$\mathbf{E}_x(\tau_y^+) = \sum_{t \geq 0} \mathbf{P}_x\{\tau_y^+ > t\} \leq \sum_{k \geq 0} r\mathbf{P}_x\{\tau_y^+ > kr\} \leq r \sum_{k \geq 0} (1 - \varepsilon)^k < \infty.$$

■

1.5.3. Existence of a stationary distribution. The Convergence Theorem (Theorem 4.9 below) implies that the "long-term" fractions of time a finite irreducible aperiodic Markov chain spends in each state coincide with the chain's stationary distribution. However, we have not yet demonstrated that stationary distributions exist! To build a candidate distribution, we consider a sojourn of the chain from some arbitrary state z back to z. Since visits to z break up the trajectory of the chain into identically distributed segments, it should not be surprising that the average fraction of time per segment spent in each state y coincides with the "long-term" fraction of time spent in y.

PROPOSITION 1.14. *Let P be the transition matrix of an irreducible Markov chain. Then*

(i) *there exists a probability distribution π on Ω such that $\pi = \pi P$ and $\pi(x) > 0$ for all $x \in \Omega$, and moreover,*
(ii) $\pi(x) = \frac{1}{\mathbf{E}_x(\tau_x^+)}$.

REMARK 1.15. We will see in Section 1.7 that existence of π does not need irreducibility, but positivity does.

PROOF. Let $z \in \Omega$ be an arbitrary state of the Markov chain. We will closely examine the time the chain spends, on average, at each state in between visits to z. Hence define

$$\begin{aligned}\tilde{\pi}(y) &:= \mathbf{E}_z(\text{number of visits to } y \text{ before returning to } z) \\ &= \sum_{t=0}^{\infty} \mathbf{P}_z\{X_t = y, \tau_z^+ > t\}.\end{aligned} \tag{1.19}$$

For any state y, we have $\tilde{\pi}(y) \le \mathbf{E}_z\tau_z^+$. Hence Lemma 1.13 ensures that $\tilde{\pi}(y) < \infty$ for all $y \in \Omega$. We check that $\tilde{\pi}$ is stationary, starting from the definition:

$$\sum_{x\in\Omega} \tilde{\pi}(x)P(x,y) = \sum_{x\in\Omega}\sum_{t=0}^{\infty} \mathbf{P}_z\{X_t = x, \tau_z^+ > t\}P(x,y). \tag{1.20}$$

Because the event $\{\tau_z^+ \ge t+1\} = \{\tau_z^+ > t\}$ is determined by $X_0, \dots, X_t$,

$$\mathbf{P}_z\{X_t = x,\, X_{t+1} = y,\, \tau_z^+ \ge t+1\} = \mathbf{P}_z\{X_t = x,\, \tau_z^+ \ge t+1\}P(x,y). \tag{1.21}$$

Reversing the order of summation in (1.20) and using the identity (1.21) shows that

$$\begin{aligned}\sum_{x\in\Omega} \tilde{\pi}(x)P(x,y) &= \sum_{t=0}^{\infty} \mathbf{P}_z\{X_{t+1} = y, \tau_z^+ \ge t+1\} \\ &= \sum_{t=1}^{\infty} \mathbf{P}_z\{X_t = y, \tau_z^+ \ge t\}.\end{aligned} \tag{1.22}$$

The expression in (1.22) is very similar to (1.19), so we are almost done. In fact,

$$\begin{aligned}
&\sum_{t=1}^{\infty}\mathbf{P}_z\{X_t = y, \tau_z^+ \geq t\} \\
&\quad = \tilde{\pi}(y) - \mathbf{P}_z\{X_0 = y, \tau_z^+ > 0\} + \sum_{t=1}^{\infty}\mathbf{P}_z\{X_t = y, \tau_z^+ = t\} \\
&\quad = \tilde{\pi}(y) - \mathbf{P}_z\{X_0 = y\} + \mathbf{P}_z\{X_{\tau_z^+} = y\}. \qquad (1.23) \\
&\quad = \tilde{\pi}(y). \qquad (1.24)
\end{aligned}$$

The equality (1.24) follows by considering two cases:

$y = z$: Since $X_0 = z$ and $X_{\tau_z^+} = z$, the last two terms of (1.23) are both 1, and they cancel each other out.

$y \neq z$: Here both terms of (1.23) are 0.

Therefore, combining (1.22) with (1.24) shows that $\tilde{\pi} = \tilde{\pi}P$.

Finally, to get a probability measure, we normalize by $\sum_x \tilde{\pi}(x) = \mathbf{E}_z(\tau_z^+)$:

$$\pi(x) = \frac{\tilde{\pi}(x)}{\mathbf{E}_z(\tau_z^+)} \quad \text{satisfies } \pi = \pi P. \qquad (1.25)$$

In particular, for any $x \in \Omega$,

$$\pi(x) = \frac{1}{\mathbf{E}_x(\tau_x^+)}. \qquad (1.26)$$

■

The computation at the heart of the proof of Proposition 1.14 can be generalized. A ***stopping time*** τ for (X_t) is a $\{0, 1, \dots, \} \cup \{\infty\}$-valued random variable such that, for each t, the event $\{\tau = t\}$ is determined by $X_0, \dots, X_t$. (Stopping times are discussed in detail in Section 6.2.1.) If a stopping time τ replaces τ_z^+ in the definition (1.19) of $\tilde{\pi}$, then the proof that $\tilde{\pi}$ satisfies $\tilde{\pi} = \tilde{\pi}P$ works, provided that τ satisfies both $\mathbf{P}_z\{\tau < \infty\} = 1$ and $\mathbf{P}_z\{X_\tau = z\} = 1$.

If τ is a stopping time, then an immediate consequence of the definition and the Markov property is

$$\begin{aligned}
&\mathbf{P}_{x_0}\{(X_{\tau+1}, X_{\tau+2}, \dots, X_\ell) \in A \mid \tau = k \text{ and } (X_1, \dots, X_k) = (x_1, \dots, x_k)\} \\
&\qquad\qquad = \mathbf{P}_{x_k}\{(X_1, \dots, X_\ell) \in A\}, \qquad (1.27)
\end{aligned}$$

for any $A \subset \Omega^\ell$. This is referred to as the ***strong Markov property***. Informally, we say that the chain "starts afresh" at a stopping time. While this is an easy fact for countable state space, discrete-time Markov chains, establishing it for processes in the continuum is more subtle.

1.5.4. Uniqueness of the stationary distribution. Earlier this chapter we pointed out the difference between multiplying a row vector by P on the right and a column vector by P on the left: the former advances a distribution by one step of the chain, while the latter gives the expectation of a function on states, one step of the chain later. We call distributions invariant under right multiplication by P ***stationary***. What about functions that are invariant under left multiplication?

Call a function $h : \Omega \to \mathbb{R}$ ***harmonic at*** x if

$$h(x) = \sum_{y \in \Omega} P(x, y)h(y). \qquad (1.28)$$

A function is ***harmonic on*** $D \subset \Omega$ if it is harmonic at every state $x \in D$. If h is regarded as a column vector, then a function which is harmonic on all of Ω satisfies the matrix equation $Ph = h$.

LEMMA 1.16. *Suppose that P is irreducible. A function h which is harmonic at every point of Ω is constant.*

PROOF. Since Ω is finite, there must be a state x_0 such that $h(x_0) = M$ is maximal. If for some state z such that $P(x_0, z) > 0$ we have $h(z) < M$, then

$$h(x_0) = P(x_0, z)h(z) + \sum_{y \neq z} P(x_0, y)h(y) < M, \tag{1.29}$$

a contradiction. It follows that $h(z) = M$ for all states z such that $P(x_0, z) > 0$.

For any $y \in \Omega$, irreducibility implies that there is a sequence $x_0, x_1, \ldots, x_n = y$ with $P(x_i, x_{i+1}) > 0$. Repeating the argument above tells us that $h(y) = h(x_{n-1}) = \cdots = h(x_0) = M$. Thus h is constant. ∎

COROLLARY 1.17. *Let P be the transition matrix of an irreducible Markov chain. There exists a unique probability distribution π satisfying $\pi = \pi P$.*

PROOF. By Proposition 1.14 there exists at least one such measure. Lemma 1.16 implies that the kernel of $P - I$ has dimension 1, so the column rank of $P - I$ is $|\Omega| - 1$. Since the row rank of any square matrix is equal to its column rank, the row-vector equation $\nu = \nu P$ also has a one-dimensional space of solutions. This space contains only one vector whose entries sum to 1. ∎

REMARK 1.18. Another proof of Corollary 1.17 follows from the Convergence Theorem (Theorem 4.9, proved below). Another simple direct proof is suggested in Exercise 1.13.

1.6. Reversibility and Time Reversals

Suppose a probability π on Ω satisfies

$$\pi(x)P(x, y) = \pi(y)P(y, x) \quad \text{for all } x, y \in \Omega. \tag{1.30}$$

The equations (1.30) are called the ***detailed balance equations***.

PROPOSITION 1.19. *Let P be the transition matrix of a Markov chain with state space Ω. Any distribution π satisfying the detailed balance equations (1.30) is stationary for P.*

PROOF. Sum both sides of (1.30) over all y:

$$\sum_{y \in \Omega} \pi(y)P(y, x) = \sum_{y \in \Omega} \pi(x)P(x, y) = \pi(x),$$

since P is stochastic. ∎

Checking detailed balance is often the simplest way to verify that a particular distribution is stationary. Furthermore, when (1.30) holds,

$$\pi(x_0)P(x_0, x_1)\cdots P(x_{n-1}, x_n) = \pi(x_n)P(x_n, x_{n-1})\cdots P(x_1, x_0). \tag{1.31}$$

We can rewrite (1.31) in the following suggestive form:

$$\mathbf{P}_\pi\{X_0 = x_0, \ldots, X_n = x_n\} = \mathbf{P}_\pi\{X_0 = x_n, X_1 = x_{n-1}, \ldots, X_n = x_0\}. \tag{1.32}$$

In other words, if a chain (X_t) satisfies (1.30) and has stationary initial distribution, then the distribution of $(X_0, X_1, \ldots, X_n)$ is the same as the distribution of $(X_n, X_{n-1}, \ldots, X_0)$. For this reason, a chain satisfying (1.30) is called ***reversible***.

EXAMPLE 1.20. Consider the simple random walk on a graph G. We saw in Example 1.12 that the distribution $\pi(x) = \deg(x)/2|E|$ is stationary.

Since

$$\pi(x)P(x,y) = \frac{\deg(x)}{2|E|}\frac{\mathbf{1}_{\{x\sim y\}}}{\deg(x)} = \frac{\mathbf{1}_{\{x\sim y\}}}{2|E|} = \pi(y)P(x,y),$$

the chain is reversible. (Note: here the notation $\mathbf{1}_A$ represents the ***indicator function*** of a set A, for which $\mathbf{1}_A(a) = 1$ if and only if $a \in A$; otherwise $\mathbf{1}_A(a) = 0$.)

EXAMPLE 1.21. Consider the ***biased random walk on the n-cycle***: a particle moves clockwise with probability p and moves counterclockwise with probability $q = 1 - p$.

The stationary distribution remains uniform: if $\pi(k) = 1/n$, then

$$\sum_{j\in\mathbb{Z}_n} \pi(j)P(j,k) = \pi(k-1)p + \pi(k+1)q = \frac{1}{n},$$

whence π is the stationary distribution. However, if $p \neq 1/2$, then

$$\pi(k)P(k,k+1) = \frac{p}{n} \neq \frac{q}{n} = \pi(k+1)P(k+1,k).$$

The ***time reversal*** of an irreducible Markov chain with transition matrix P and stationary distribution π is the chain with matrix

$$\widehat{P}(x,y) := \frac{\pi(y)P(y,x)}{\pi(x)}. \tag{1.33}$$

The stationary equation $\pi = \pi P$ implies that $\widehat{P}$ is a stochastic matrix. Proposition 1.22 shows that the terminology "time reversal" is deserved.

PROPOSITION 1.22. *Let (X_t) be an irreducible Markov chain with transition matrix P and stationary distribution π. Write $(\widehat{X}_t)$ for the time-reversed chain with transition matrix $\widehat{P}$. Then π is stationary for $\widehat{P}$, and for any $x_0, \ldots, x_t \in \Omega$ we have*

$$\mathbf{P}_\pi\{X_0 = x_0, \ldots, X_t = x_t\} = \mathbf{P}_\pi\{\widehat{X}_0 = x_t, \ldots, \widehat{X}_t = x_0\}.$$

PROOF. To check that π is stationary for $\widehat{P}$, we simply compute

$$\sum_{y\in\Omega} \pi(y)\widehat{P}(y,x) = \sum_{y\in\Omega} \pi(y)\frac{\pi(x)P(x,y)}{\pi(y)} = \pi(x).$$

To show the probabilities of the two trajectories are equal, note that

$$\begin{aligned}\mathbf{P}_\pi\{X_0 = x_0, \ldots, X_n = x_n\} &= \pi(x_0)P(x_0,x_1)P(x_1,x_2)\cdots P(x_{n-1},x_n)\\ &= \pi(x_n)\widehat{P}(x_n,x_{n-1})\cdots\widehat{P}(x_2,x_1)\widehat{P}(x_1,x_0)\\ &= \mathbf{P}_\pi\{\hat{X}_0 = x_n, \ldots, \hat{X}_n = x_0\},\end{aligned}$$

since $P(x_{i-1},x_i) = \pi(x_i)\widehat{P}(x_i,x_{i-1})/\pi(x_{i-1})$ for each i. ∎

Observe that if a chain with transition matrix P is reversible, then $\widehat{P} = P$.

$y \notin \mathcal{C}$, whence π is supported on $\mathcal{C}$. Consequently, for $x \in \mathcal{C}$,

$$\pi(x) = \sum_{y \in \Omega} \pi(y)P(y,x) = \sum_{y \in \mathcal{C}} \pi(y)P(y,x) = \sum_{y \in \mathcal{C}} \pi(y)P_{|\mathcal{C}}(y,x),$$

and π restricted to $\mathcal{C}$ is stationary for $P_{|\mathcal{C}}$. By uniqueness of the stationary distribution for $P_{|\mathcal{C}}$, it follows that $\pi(x) = \pi^{\mathcal{C}}(x)$ for all $x \in \mathcal{C}$. Therefore,

$$\pi(x) = \begin{cases} \pi^{\mathcal{C}}(x) & \text{if } x \in \mathcal{C}, \\ 0 & \text{if } x \notin \mathcal{C}, \end{cases}$$

and the solution to $\pi = \pi P$ is unique.

Suppose there are distinct essential communicating classes for P, say $\mathcal{C}_1$ and $\mathcal{C}_2$. The restriction of P to each of these classes is irreducible. Thus for $i = 1, 2$, there exists a measure π supported on $\mathcal{C}_i$ which is stationary for $P_{|\mathcal{C}_i}$. Moreover, it is easily verified that each π_i is stationary for P, and so P has more than one stationary distribution. ■

Exercises

EXERCISE 1.1. Let P be the transition matrix of random walk on the n-cycle, where n is odd. Find the smallest value of t such that $P^t(x,y) > 0$ for all states x and y.

EXERCISE 1.2. A graph G is ***connected*** when, for two vertices x and y of G, there exists a sequence of vertices $x_0, x_1, \ldots, x_k$ such that $x_0 = x$, $x_k = y$, and $x_i \sim x_{i+1}$ for $0 \le i \le k-1$. Show that random walk on G is irreducible if and only if G is connected.

EXERCISE 1.3. We define a graph to be a ***tree*** if it is connected but contains no cycles. Prove that the following statements about a graph T with n vertices and m edges are equivalent:

(a) T is a tree.
(b) T is connected and $m = n - 1$.
(c) T has no cycles and $m = n - 1$.

EXERCISE 1.4. Let T be a tree. A ***leaf*** is a vertex of degree 1.

(a) Prove that T contains a leaf.
(b) Prove that between any two vertices in T there is a unique simple path.
(c) Prove that T has at least 2 leaves.

EXERCISE 1.5. Let T be a tree. Show that the graph whose vertices are proper 3-colorings of T and whose edges are pairs of colorings which differ at only a single vertex is connected.

EXERCISE 1.6. Let P be an irreducible transition matrix of period b. Show that Ω can be partitioned into b sets $\mathcal{C}_1, \mathcal{C}_2, \ldots, \mathcal{C}_b$ in such a way that $P(x,y) > 0$ only if $x \in \mathcal{C}_i$ and $y \in \mathcal{C}_{i+1}$. (The addition $i+1$ is modulo b.)

EXERCISE 1.7. A transition matrix P is ***symmetric*** if $P(x,y) = P(y,x)$ for all $x, y \in \Omega$. Show that if P is symmetric, then the uniform distribution on Ω is stationary for P.

EXERCISE 1.8. Let P be a transition matrix which is reversible with respect to the probability distribution π on Ω. Show that the transition matrix P^2 corresponding to two steps of the chain is also reversible with respect to π.

EXERCISE 1.9. Let π be a stationary distribution for an irreducible transition matrix P. Prove that $\pi(x) > 0$ for all $x \in \Omega$, without using the explicit formula (1.25).

EXERCISE 1.10. Check carefully that equation (1.19) is true.

EXERCISE 1.11. Here we outline another proof, more analytic, of the existence of stationary distributions. Let P be the transition matrix of a Markov chain on a finite state space Ω. For an arbitrary initial distribution μ on Ω and $n > 0$, define the distribution ν_n by

$$\nu_n = \frac{1}{n}\left(\mu + \mu P + \cdots + \mu P^{n-1}\right).$$

(a) Show that for any $x \in \Omega$ and $n > 0$,

$$|\nu_n P(x) - \nu_n(x)| \leq \frac{2}{n}.$$

(b) Show that there exists a subsequence $(\nu_{n_k})_{k\geq 0}$ such that $\lim_{k\to\infty} \nu_{n_k}(x)$ exists for every $x \in \Omega$.

(c) For $x \in \Omega$, define $\nu(x) = \lim_{k\to\infty} \nu_{n_k}(x)$. Show that ν is a stationary distribution for P.

EXERCISE 1.12. Let P be the transition matrix of an irreducible Markov chain with state space Ω. Let $B \subset \Omega$ be a non-empty subset of the state space, and assume $h : \Omega \to \mathbb{R}$ is a function harmonic at all states $x \notin B$.

Prove that if h is non-constant and $h(y) = \max_{x\in\Omega} h(x)$, then $y \in B$.

(This is a discrete version of the ***maximum principle***.)

EXERCISE 1.13. Give a direct proof that the stationary distribution for an irreducible chain is unique.

Hint: Given stationary distributions π_1 and π_2, consider the state x that minimizes $\pi_1(x)/\pi_2(x)$ and show that all y with $P(x, y) > 0$ have $\pi_1(y)/\pi_2(y) = \pi_1(x)/\pi_2(x)$.

EXERCISE 1.14. Show that any stationary measure π of an irreducible chain must be strictly positive.

Hint: Show that if $\pi(x) = 0$, then $\pi(y) = 0$ whenever $P(x, y) > 0$.

EXERCISE 1.15. For a subset $A \subset \Omega$, define $f(x) = \mathbf{E}_x(\tau_A)$. Show that

(a)
$$f(x) = 0 \quad \text{for } x \in A. \tag{1.35}$$

(b)
$$f(x) = 1 + \sum_{y\in\Omega} P(x, y)f(y) \quad \text{for } x \notin A. \tag{1.36}$$

(c) f is uniquely determined by (1.35) and (1.36).

The following exercises concern the material in Section 1.7.

EXERCISE 1.16. Show that $\leftrightarrow$ is an equivalence relation on Ω.

EXERCISE 1.17. Show that the set of stationary measures for a transition matrix forms a polyhedron with one vertex for each essential communicating class.

CHAPTER 2

Classical (and Useful) Markov Chains

Here we present several basic and important examples of Markov chains. The results we prove in this chapter will be used in many places throughout the book.

This is also the only chapter in the book where the central chains are not always irreducible. Indeed, two of our examples, gambler's ruin and coupon collecting, both have absorbing states. For each we examine closely how long it takes to be absorbed.

2.1. Gambler's Ruin

Consider a gambler betting on the outcome of a sequence of independent fair coin tosses. If the coin comes up heads, she adds one dollar to her purse; if the coin lands tails up, she loses one dollar. If she ever reaches a fortune of n dollars, she will stop playing. If her purse is ever empty, then she must stop betting.

The gambler's situation can be modeled by a random walk on a path with vertices $\{0, 1, \ldots, n\}$. At all interior vertices, the walk is equally likely to go up by 1 or down by 1. That states 0 and n are absorbing, meaning that once the walk arrives at either 0 or n, it stays forever (cf. Section 1.7).

There are two questions that immediately come to mind: how long will it take for the gambler to arrive at one of the two possible fates? What are the probabilities of the two possibilities?

PROPOSITION 2.1. *Assume that a gambler making fair unit bets on coin flips will abandon the game when her fortune falls to* 0 *or rises to* n. *Let* X_t *be gambler's fortune at time* t *and let* τ *be the time required to be absorbed at one of* 0 *or* n. *Assume that* $X_0 = k$, *where* $0 \le k \le n$. *Then*

$$\mathbf{P}_k\{X_\tau = n\} = k/n \tag{2.1}$$

and

$$\mathbf{E}_k(\tau) = k(n-k). \tag{2.2}$$

PROOF. Let p_k be the probability that the gambler reaches a fortune of n before ruin, given that she starts with k dollars. We solve simultaneously for $p_0, p_1, \ldots, p_n$. Clearly $p_0 = 0$ and $p_n = 1$, while

$$p_k = \frac{1}{2}p_{k-1} + \frac{1}{2}p_{k+1} \quad \text{for } 1 \le k \le n-1. \tag{2.3}$$

Why? With probability 1/2, the walk moves to $k+1$. The conditional probability of reaching n before 0, starting from $k+1$, is exactly p_{k+1}. Similarly, with probability 1/2 the walk moves to $k-1$, and the conditional probability of reaching n before 0 from state $k-1$ is p_{k-1}.

Solving the system (2.3) of linear equations yields $p_k = k/n$ for $0 \le k \le n$.

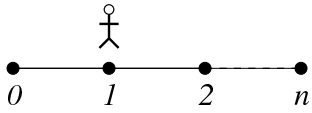


FIGURE 2.1. How long until the walk reaches either 0 or n? What is the probability of each?

For (2.2), again we try to solve for all the values at once. To this end, write f_k for the expected time $\mathbf{E}_k(\tau)$ to be absorbed, starting at position k. Clearly, $f_0 = f_n = 0$; the walk is started at one of the absorbing states. For $1 \leq k \leq n-1$, it is true that

$$f_k = \frac{1}{2}(1 + f_{k+1}) + \frac{1}{2}(1 + f_{k-1}). \tag{2.4}$$

Why? When the first step of the walk increases the gambler's fortune, then the conditional expectation of τ is 1 (for the initial step) plus the expected additional time needed. The expected additional time needed is f_{k+1}, because the walk is now at position $k+1$. Parallel reasoning applies when the gambler's fortune first decreases.

Exercise 2.1 asks the reader to solve this system of equations, completing the proof of (2.2). ■

REMARK 2.2. See Chapter 9 for powerful generalizations of the simple methods we have just applied.

2.2. Coupon Collecting

A company issues n different types of coupons. A collector desires a complete set. We suppose each coupon he acquires is equally likely to be each of the n types. How many coupons must he obtain so that his collection contains all n types?

It may not be obvious why this is a Markov chain. Let X_t denote the number of different types represented among the collector's first t coupons. Clearly $X_0 = 0$. When the collector has coupons of k different types, there are $n-k$ types missing. Of the n possibilities for his next coupon, only $n-k$ will expand his collection. Hence

$$\mathbf{P}\{X_{t+1} = k+1 \mid X_t = k\} = \frac{n-k}{n}$$

and

$$\mathbf{P}\{X_{t+1} = k \mid X_t = k\} = \frac{k}{n}.$$

Every trajectory of this chain is non-decreasing. Once the chain arrives at state n (corresponding to a complete collection), it is absorbed there. We are interested in the number of steps required to reach the absorbing state.

PROPOSITION 2.3. *Consider a collector attempting to collect a complete set of coupons. Assume that each new coupon is chosen uniformly and independently from the set of n possible types, and let τ be the (random) number of coupons collected when the set first contains every type. Then*

$$\mathbf{E}(\tau) = n \sum_{k=1}^{n} \frac{1}{k}.$$

PROOF. The expectation $\mathbf{E}(\tau)$ can be computed by writing τ as a sum of geometric random variables. Let τ_k be the total number of coupons accumulated when the collection first contains k distinct coupons. Then

$$\tau = \tau_n = \tau_1 + (\tau_2 - \tau_1) + \cdots + (\tau_n - \tau_{n-1}). \tag{2.5}$$

Furthermore, $\tau_k - \tau_{k-1}$ is a geometric random variable with success probability $(n-k+1)/n$: after collecting τ_{k-1} coupons, there are $n-k+1$ types missing from the collection. Each subsequent coupon drawn has the same probability $(n-k+1)/n$ of being a type not already collected, until a new type is finally drawn. Thus $\mathbf{E}(\tau_k - \tau_{k-1}) = n/(n-k+1)$ and

$$\mathbf{E}(\tau) = \sum_{k=1}^{n} \mathbf{E}(\tau_k - \tau_{k-1}) = n \sum_{k=1}^{n} \frac{1}{n-k+1} = n \sum_{k=1}^{n} \frac{1}{k}. \tag{2.6}$$

■

While the argument for Proposition 2.3 is simple and vivid, we will often need to know more about the distribution of τ in future applications. Recall that $|\sum_{k=1}^{n} 1/k - \log n| \le 1$, whence $|\mathbf{E}(\tau) - n\log n| \le n$ (see Exercise 2.4 for a better estimate). Proposition 2.4 says that τ is unlikely to be much larger than its expected value.

PROPOSITION 2.4. *Let τ be a coupon collector random variable, as in Proposition 2.3. For any $c > 0$,*

$$\mathbf{P}\{\tau > \lceil n\log n + cn\rceil\} \le e^{-c}. \tag{2.7}$$

PROOF. Let A_i be the event that the i-th type does not appear among the first $\lceil n\log n + cn\rceil$ coupons drawn. Observe first that

$$\mathbf{P}\{\tau > \lceil n\log n + cn\rceil\} = \mathbf{P}\left(\bigcup_{i=1}^{n} A_i\right) \le \sum_{i=1}^{n} \mathbf{P}(A_i).$$

Since each trial has probability $1 - n^{-1}$ of *not* drawing coupon i and the trials are independent, the right-hand side above is bounded above by

$$\sum_{i=1}^{n} \left(1 - \frac{1}{n}\right)^{\lceil n\log n + cn\rceil} \le n \exp\left(-\frac{n\log n + cn}{n}\right) = e^{-c},$$

proving (2.7). ■

2.3. The Hypercube and the Ehrenfest Urn Model

The ***n*-dimensional hypercube** is a graph whose vertices are the binary n-tuples $\{0,1\}^n$. Two vertices are connected by an edge when they differ in exactly one coordinate. See Figure 2.2 for an illustration of the three-dimensional hypercube.

The simple random walk on the hypercube moves from a vertex $(x^1, x^2, \ldots, x^n)$ by choosing a coordinate $j \in \{1, 2, \ldots, n\}$ uniformly at random and setting the new state equal to $(x^1, \ldots, x^{j-1}, 1 - x^j, x^{j+1}, \ldots, x^n)$. That is, the bit at the walk's chosen coordinate is flipped. (This is a special case of the walk defined in Section 1.4.)

Unfortunately, the simple random walk on the hypercube is periodic, since every move flips the parity of the number of 1's. The ***lazy random walk***, which does not have this problem, remains at its current position with probability 1/2 and moves

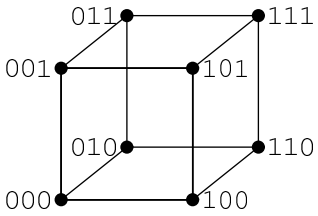


FIGURE 2.2. The three-dimensional hypercube.

as above with probability 1/2. This chain can be realized by choosing a coordinate uniformly at random and *refreshing* the bit at this coordinate by replacing it with an unbiased random bit independent of time, current state, and coordinate chosen.

Since the hypercube is an n-regular graph, Example 1.12 implies that the stationary distribution of both the simple and lazy random walks is uniform on $\{0,1\}^n$.

We now consider a process, the ***Ehrenfest urn***, which at first glance appears quite different. Suppose n balls are distributed among two urns, I and II. At each move, a ball is selected uniformly at random and transferred from its current urn to the other urn. If X_t is the number of balls in urn I at time t, then the transition matrix for (X_t) is

$$P(j,k) = \begin{cases} \frac{n-j}{n} & \text{if } k = j+1, \\ \frac{j}{n} & \text{if } k = j-1, \\ 0 & \text{otherwise.} \end{cases} \tag{2.8}$$

Thus (X_t) is a Markov chain with state space $\Omega = \{0,1,2,\ldots,n\}$ that moves by ± 1 on each move and is biased towards the middle of the interval. The stationary distribution for this chain is binomial with parameters n and 1/2 (see Exercise 2.5).

The Ehrenfest urn is a projection (in a sense that will be defined precisely in Section 2.3.1) of the random walk on the n-dimensional hypercube. This is unsurprising given the standard bijection between $\{0,1\}^n$ and subsets of $\{1,\ldots,n\}$, under which a set corresponds to the vector with 1's in the positions of its elements. We can view the position of the random walk on the hypercube as specifying the set of balls in Ehrenfest urn I; then changing a bit corresponds to moving a ball into or out of the urn.

Define the ***Hamming weight*** $W(\boldsymbol{x})$ of a vector $\boldsymbol{x} := (x^1,\ldots,x^n) \in \{0,1\}^n$ to be its number of coordinates with value 1:

$$W(\boldsymbol{x}) = \sum_{j=1}^{n} x^j. \tag{2.9}$$

Let $(\boldsymbol{X}_t)$ be the simple random walk on the n-dimensional hypercube, and let $W_t = W(\boldsymbol{X}_t)$ be the Hamming weight of the walk's position at time t.

When $W_t = j$, the weight increments by a unit amount when one of the $n-j$ coordinates with value 0 is selected. Likewise, when one of the j coordinates with value 1 is selected, the weight decrements by one unit. From this description, it is clear that (W_t) is a Markov chain with transition probabilities given by (2.8).

2.3.1. Projections of chains. The Ehrenfest urn is a *projection*, which we define in this section, of the simple random walk on the hypercube.

CHAPTER 3

Markov Chain Monte Carlo: Metropolis and Glauber Chains

3.1. Introduction

Given an irreducible transition matrix P, there is a unique stationary distribution π satisfying $\pi = \pi P$, which we constructed in Section 1.5. We now consider the inverse problem: given a probability distribution π on Ω, can we find a transition matrix P for which π is its stationary distribution? The following example illustrates why this is a natural problem to consider.

A ***random sample*** from a finite set Ω will mean a random uniform selection from Ω, i.e., one such that each element has the same chance $1/|\Omega|$ of being chosen.

Fix a set $\{1, 2, \ldots, q\}$ of *colors*. A ***proper q-coloring*** of a graph $G = (V, E)$ is an assignment of colors to the vertices V, subject to the constraint that neighboring vertices do not receive the same color. There are (at least) two reasons to look for an efficient method to sample from Ω, the set of all proper q-colorings. If a random sample can be produced, then the size of Ω can be estimated (as we discuss in detail in Section 14.4.2). Also, if it is possible to sample from Ω, then average characteristics of colorings can be studied via simulation.

For some graphs, e.g. trees, there are simple recursive methods for generating a random proper coloring (see Example 14.10). However, for other graphs it can be challenging to directly construct a random sample. One approach is to use Markov chains to sample: suppose that (X_t) is a chain with state space Ω and with stationary distribution uniform on Ω (in Section 3.3, we will construct one such chain). By the Convergence Theorem (Theorem 4.9, whose proof we have not yet given but have often foreshadowed), X_t is approximately uniformly distributed when t is large.

This method of sampling from a given probability distribution is called ***Markov chain Monte Carlo***. Suppose π is a probability distribution on Ω. If a Markov chain (X_t) with stationary distribution π can be constructed, then, for t large enough, the distribution of X_t is close to π. The focus of this book is to determine how large t must be to obtain a sufficient approximation. In this chapter we will focus on the task of finding chains with a given stationary distribution.

3.2. Metropolis Chains

Given *some* chain with state space Ω and an arbitrary stationary distribution, can the chain be modified so that the new chain has the stationary distribution π? The Metropolis algorithm accomplishes this.

3.2.1. Symmetric base chain. Suppose that Ψ is a symmetric transition matrix. In this case, Ψ is reversible with respect to the uniform distribution on Ω.

We now show how to modify transitions made according to Ψ to obtain a chain with stationary distribution π, where π is any probability distribution on Ω.

The new chain evolves as follows: when at state x, a candidate move is generated from the distribution $\Psi(x,\cdot)$. If the proposed new state is y, then the move is censored with probability $1-a(x,y)$. That is, with probability $a(x,y)$, the state y is "accepted" so that the next state of the chain is y, and with the remaining probability $1-a(x,y)$, the chain remains at x. Rejecting moves slows the chain and can reduce its computational efficiency but may be necessary to achieve a specific stationary distribution. We will discuss how to choose the acceptance probability $a(x,y)$ below, but for now observe that the transition matrix P of the new chain is

$$P(x,y) = \begin{cases} \Psi(x,y)a(x,y) & \text{if } y \neq x, \\ 1 - \sum_{z\,:\,z\neq x} \Psi(x,z)a(x,z) & \text{if } y = x. \end{cases}$$

By Proposition 1.19, the transition matrix P has stationary distribution π if

$$\pi(x)\Psi(x,y)a(x,y) = \pi(y)\Psi(y,x)a(y,x) \tag{3.1}$$

for all $x \neq y$. Since we have assumed Ψ is symmetric, equation (3.1) holds if and only if

$$b(x,y) = b(y,x), \tag{3.2}$$

where $b(x,y) = \pi(x)a(x,y)$. Because $a(x,y)$ is a probability and must satisfy $a(x,y) \leq 1$, the function b must obey the constraints

$$\begin{aligned} b(x,y) &\leq \pi(x), \\ b(x,y) = b(y,x) &\leq \pi(y). \end{aligned} \tag{3.3}$$

Since rejecting the moves of the original chain Ψ is wasteful, a solution b to (3.2) and (3.3) should be chosen which is as large as possible. Clearly, all solutions are bounded above by $b(x,y) = \pi(x) \wedge \pi(y) := \min\{\pi(x), \pi(y)\}$. For this choice, the acceptance probability $a(x,y)$ is equal to $(\pi(y)/\pi(x)) \wedge 1$.

The ***Metropolis chain*** for a probability π and a symmetric transition matrix Ψ is defined as

$$P(x,y) = \begin{cases} \Psi(x,y)\left[1 \wedge \frac{\pi(y)}{\pi(x)}\right] & \text{if } y \neq x, \\ 1 - \sum_{z\,:\,z\neq x} \Psi(x,z)\left[1 \wedge \frac{\pi(z)}{\pi(x)}\right] & \text{if } y = x. \end{cases}$$

Our discussion above shows that π is indeed a stationary distribution for the Metropolis chain.

REMARK 3.1. A very important feature of the Metropolis chain is that it only depends on the ratios $\pi(x)/\pi(y)$. Frequently $\pi(x)$ has the form $h(x)/Z$, where the function $h : \Omega \to [0,\infty)$ is known and $Z = \sum_{x\in\Omega} h(x)$ is a normalizing constant. It may be difficult to explicitly compute Z, especially if Ω is large. Because the Metropolis chain only depends on $h(x)/h(y)$, it is not necessary to compute the constant Z in order to simulate the chain. The optimization chains described below (Example 3.2) are examples of this type.

EXAMPLE 3.2 (Optimization). Let f be a real-valued function defined on the vertex set Ω of a graph. In many applications it is desirable to find a vertex x where $f(x)$ is maximal. If the domain Ω is very large, then an exhaustive search may be too expensive.

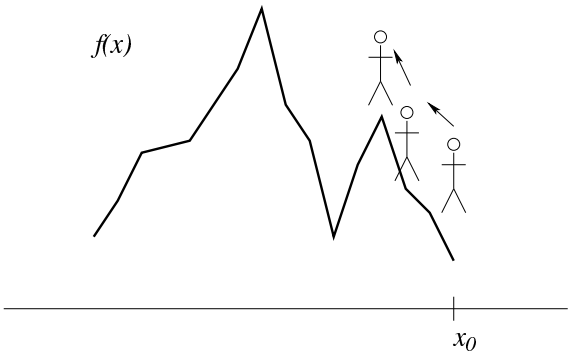


FIGURE 3.1. A hill climb algorithm may become trapped at a local maximum.

A ***hill climb*** is an algorithm which attempts to locate the maximum values of f as follows: when at x, if a neighbor y of x has $f(y) > f(x)$, move to y. When f has local maxima that are not also global maxima, then the climber may become trapped before discovering a global maximum—see Figure 3.1.

One solution is to randomize moves so that instead of always remaining at a local maximum, with some probability the climber moves to lower states.

Suppose for simplicity that Ω is a regular graph, so that simple random walk on Ω has a symmetric transition matrix. Fix $\lambda \geq 1$ and define

$$\pi_\lambda(x) = \frac{\lambda^{f(x)}}{Z(\lambda)},$$

where $Z(\lambda) := \sum_{x\in\Omega} \lambda^{f(x)}$ is the normalizing constant that makes π_λ a probability measure (as mentioned in Remark 3.1, running the Metropolis chain does not require computation of $Z(\lambda)$, which may be prohibitively expensive to compute). Since $\pi_\lambda(x)$ is increasing in $f(x)$, the measure π_λ favors vertices x for which $f(x)$ is large.

If $f(y) < f(x)$, the Metropolis chain accepts a transition $x \to y$ with probability $\lambda^{-[f(x)-f(y)]}$. As $\lambda \to \infty$, the chain more closely resembles the deterministic hill climb.

Define

$$\Omega^\star := \left\{ x \in \Omega : f(x) = f^\star := \max_{y\in\Omega} f(y) \right\}.$$

Then

$$\lim_{\lambda\to\infty} \pi_\lambda(x) = \lim_{\lambda\to\infty} \frac{\lambda^{f(x)}/\lambda^{f^\star}}{|\Omega^\star| + \sum_{x\in\Omega\setminus\Omega^\star} \lambda^{f(x)}/\lambda^{f^\star}} = \frac{\mathbf{1}_{\{x\in\Omega^\star\}}}{|\Omega^\star|}.$$

That is, as $\lambda \to \infty$, the stationary distribution π_λ of this Metropolis chain converges to the uniform distribution over the global maxima of f.

3.2.2. General base chain. The Metropolis chain can also be defined when the initial transition matrix is not symmetric. For a general (irreducible) transition matrix Ψ and an arbitrary probability distribution π on Ω, the Metropolized chain is executed as follows. When at state x, generate a state y from $\Psi(x, \cdot)$. Move to

y with probability

$$\frac{\pi(y)\Psi(y,x)}{\pi(x)\Psi(x,y)} \wedge 1, \tag{3.4}$$

and remain at x with the complementary probability. The transition matrix P for this chain is

$$P(x,y) = \begin{cases} \Psi(x,y)\left[\frac{\pi(y)\Psi(y,x)}{\pi(x)\Psi(x,y)} \wedge 1\right] & \text{if } y \neq x, \\ 1 - \sum_{z\,:\,z\neq x} \Psi(x,z)\left[\frac{\pi(z)\Psi(z,x)}{\pi(x)\Psi(x,z)} \wedge 1\right] & \text{if } y = x. \end{cases} \tag{3.5}$$

The reader should check that the transition matrix (3.5) defines a reversible Markov chain with stationary distribution π (see Exercise 3.1).

EXAMPLE 3.3. Suppose you know neither the vertex set V nor the edge set E of a graph G. However, you are able to perform a simple random walk on G. (Many computer and social networks have this form; each vertex knows who its neighbors are, but not the global structure of the graph.) If the graph is not regular, then the stationary distribution is not uniform, so the distribution of the walk will not converge to uniform. You desire a uniform sample from V. We can use the Metropolis algorithm to modify the simple random walk and ensure a uniform stationary distribution. The acceptance probability in (3.4) reduces in this case to

$$\frac{\deg(x)}{\deg(y)} \wedge 1.$$

This biases the walk against moving to higher degree vertices, giving a uniform stationary distribution. Note that it is not necessary to know the size of the vertex set to perform this modification, which can be an important consideration in applications.

3.3. Glauber Dynamics

We will study many chains whose state spaces are contained in a set of the form S^V, where V is the vertex set of a graph and S is a finite set. The elements of S^V, called ***configurations***, are the functions from V to S. We visualize a configuration as a labeling of vertices with elements of S.

Given a probability distribution π on a space of configurations, the Glauber dynamics for π, to be defined below, is a Markov chain which has stationary distribution π. This chain is often called the *Gibbs sampler*, especially in statistical contexts.

3.3.1. Two examples. As we defined in Section 3.1, a proper q-coloring of a graph $G = (V, E)$ is an element x of $\{1, 2, \ldots, q\}^V$, the set of functions from V to $\{1, 2, \ldots, q\}$, such that $x(v) \neq x(w)$ for all edges $\{v, w\}$. We construct here a Markov chain on the set of proper q-colorings of G.

For a given configuration x and a vertex v, call a color j ***allowable*** at v if j is different from all colors assigned to neighbors of v. That is, a color is allowable at v if it does *not* belong to the set $\{x(w) : w \sim v\}$. Given a proper q-coloring x, we can generate a new coloring by

- selecting a vertex $v \in V$ at random,
- selecting a color j uniformly at random from the allowable colors at v, and

- re-coloring vertex v with color j.

We claim that the resulting chain has uniform stationary distribution: why? Note that transitions are permitted only between colorings differing at a single vertex. If x and y agree everywhere except vertex v, then the chance of moving from x to y equals $|V|^{-1}|A_v(x)|^{-1}$, where $A_v(x)$ is the set of allowable colors at v in x. Since $A_v(x) = A_v(y)$, this probability equals the probability of moving from y to x. Since $P(x, y) = P(y, x)$, the detailed balance equations are satisfied by the uniform distribution.

This chain is called the ***Glauber dynamics for proper q-colorings***. Note that when a vertex v is updated in coloring x, a coloring is chosen from π conditioned on the set of colorings agreeing with x at all vertices different from v. This is the general rule for defining Glauber dynamics for any set of configurations. Before spelling out the details in the general case, we consider one other specific example.

A ***hardcore configuration*** is a placement of particles on the vertices V of a graph so that each vertex is occupied by at most one particle and no two particles are adjacent. Formally, a hardcore configuration x is an element of $\{0, 1\}^V$, the set of functions from V to $\{0, 1\}$, satisfying $x(v)x(w) = 0$ whenever v and w are neighbors. The vertices v with $x(v) = 1$ are called ***occupied***, and the vertices v with $x(v) = 0$ are called ***vacant***.

Consider the following transition rule:

- a vertex v is chosen uniformly at random, and, regardless of the current status of v,
- if any neighbor of v is occupied, v is left unoccupied, while if no adjacent vertex is occupied, a particle is placed at v with probability $1/2$.

REMARK 3.4. Note that the rule above has the same effect as the following apparently simpler rule: if no neighbor of v is occupied, then, with probability $1/2$, flip the status of v. Our original description will be much more convenient when, in the future, we attempt to couple multiple copies of this chain, since it provides a way to ensure that the status at the chosen vertex v is the same in all copies after an update. See Section 5.4.2.

The verification that this chain is reversible with respect to the uniform distribution is similar to the coloring chain just considered and is left to the reader.

3.3.2. General definition. In general, let V and S be finite sets, and suppose that Ω is a subset of S^V (both the set of proper q-colorings and the set of hardcore configurations are of this form). Let π be a probability distribution whose support is Ω. The (single-site) ***Glauber dynamics for π*** is a reversible Markov chain with state space Ω, stationary distribution π, and the transition probabilities we describe below.

In words, the Glauber chain moves from state x as follows: a vertex v is chosen uniformly at random from V, and a new state is chosen according to the measure π conditioned on the set of states equal to x at all vertices different from v. We give the details now. For $x \in \Omega$ and $v \in V$, let

$$\Omega(x, v) = \{y \in \Omega : y(w) = x(w) \text{ for all } w \neq v\} \tag{3.6}$$

be the set of states agreeing with x everywhere except possibly at v, and define

$$\pi^{x,v}(y) = \pi(y \mid \Omega(x,v)) = \begin{cases} \frac{\pi(y)}{\pi(\Omega(x,v))} & \text{if } y \in \Omega(x,v), \\ 0 & \text{if } y \notin \Omega(x,v) \end{cases}$$

to be the distribution π conditioned on the set $\Omega(x,v)$. The rule for updating a configuration x is: pick a vertex v uniformly at random, and choose a new configuration according to $\pi^{x,v}$.

The distribution π is always stationary and reversible for the Glauber dynamics (see Exercise 3.2).

3.3.3. Comparing Glauber dynamics and Metropolis chains. Suppose now that π is a probability distribution on the state space S^V, where S is a finite set and V is the vertex set of a graph. We can always define the Glauber chain as just described. Suppose on the other hand that we have a chain which picks a vertex v at random and has *some* mechanism for updating the configuration at v. (For example, the chain may pick an element of S at random to update at v.) This chain may not have stationary distribution π, but it can be modified by the Metropolis rule to obtain a chain with stationary distribution π. This chain can be very similar to the Glauber chain, but may not coincide exactly. We consider our examples.

EXAMPLE 3.5 (Chains on q-colorings). Consider the following chain on (not necessarily proper) q-colorings: a vertex v is chosen uniformly at random, a color is selected uniformly at random among *all* q colors, and the vertex v is recolored with the chosen color. We apply the Metropolis rule to this chain, where π is the probability measure which is uniform over the space of *proper* q-colorings. When at a proper coloring, if the color k is proposed to update a vertex, then the Metropolis rule accepts the proposed re-coloring with probability 1 if it yields a proper coloring and rejects otherwise.

The Glauber chain described in Section 3.3.1 is slightly different. Note in particular that the chance of remaining at the same coloring differs for the two chains. If there are a allowable colors at vertex v and this vertex v is selected for updating in the Glauber dynamics, the chance that the coloring remains the same is $1/a$. For the Metropolis chain, if vertex v is selected, the chance of remaining in the current coloring is $(1+q-a)/q$.

EXAMPLE 3.6 (Hardcore chains). Again identify elements of $\{0,1\}^V$ with a placement of particles onto the vertex set V, and consider the following chain on $\{0,1\}^V$: a vertex is chosen at random, and a particle is placed at the selected vertex with probability $1/2$. This chain does not live on the space of hardcore configurations, as there is no constraint against placing a particle on a vertex with an occupied neighbor.

We can modify this chain with the Metropolis rule to obtain a chain with stationary distribution π, where π is uniform over hardcore configurations. If x is a hardcore configuration, the move $x \to y$ is rejected if and only if y is not a hardcore configuration. The Metropolis chain and the Glauber dynamics agree in this example.

3.3.4. Hardcore model with fugacity. Let $G = (V, E)$ be a graph and let Ω be the set of hardcore configurations on G. The ***hardcore model*** with ***fugacity***

λ is the probability π on hardcore configurations defined by

$$\pi(\sigma) = \begin{cases} \frac{\lambda^{\sum_{v\in V}\sigma(v)}}{Z(\lambda)} & \text{if } \sigma(v)\sigma(w) = 0 \text{ for all } \{v,w\} \in E, \\ 0 & \text{otherwise.} \end{cases}$$

The factor $Z(\lambda) = \sum_{\sigma\in\Omega} \lambda^{\sum_{v\in V}\sigma(v)}$ normalizes π to have unit total mass.

The Glauber dynamics for the hardcore model updates a configuration $X_t = \sigma$ to a new configuration X_{t+1} as follows: a vertex w is chosen at random. Denote the set of occupied neighbors of w by $\mathcal{N}$, so that

$$\mathcal{N}(w) := \{v : v \sim w \text{ and } \sigma(v) = 1\}.$$

If $\mathcal{N}(w) \neq \varnothing$, then $X_{t+1} = \sigma$. If $\mathcal{N}(w) = \varnothing$, then set

$$X_{t+1}(w) = \begin{cases} 1 & \text{with probability } \lambda/(1+\lambda), \\ 0 & \text{with probability } 1/(1+\lambda). \end{cases}$$

Set $X_{t+1}(v) = \sigma(v)$ for all $v \neq w$.

3.3.5. The Ising model. A ***spin system*** is a probability distribution on $\Omega = \{-1, 1\}^V$, where V is the vertex set of a graph $G = (V, E)$. The value $\sigma(v)$ is called the ***spin*** at v. The physical interpretation is that magnets, each having one of the two possible orientations represented by $+1$ and -1, are placed on the vertices of the graph; a configuration specifies the orientations of these magnets.

The nearest-neighbor ***Ising model*** is the most widely studied spin system. In this system, the ***energy*** of a configuration σ is defined to be

$$H(\sigma) = -\sum_{\substack{v,w\in V \\ v\sim w}} \sigma(v)\sigma(w). \tag{3.7}$$

Clearly, the energy increases with the number of pairs of neighbors whose spins disagree (anyone who has played with magnets has observed firsthand that it is challenging to place neighboring magnets in opposite orientations and keep them there).

The ***Gibbs distribution*** corresponding to the energy H is the probability distribution μ on Ω defined by

$$\mu(\sigma) = \frac{1}{Z(\beta)} e^{-\beta H(\sigma)}. \tag{3.8}$$

Here the ***partition function*** $Z(\beta)$ is the normalizing constant required to make μ a probability distribution:

$$Z(\beta) := \sum_{\sigma\in\Omega} e^{-\beta H(\sigma)}. \tag{3.9}$$

The parameter $\beta \geq 0$ determines the importance of the energy function. In the physical interpretation, β is the reciprocal of temperature. At infinite temperature ($\beta = 0$), the energy function H plays no role and μ is the uniform distribution on Ω. In this case, there is no interaction between the spins at differing vertices and the random variables $\{\sigma(v)\}_{v\in V}$ are independent. As $\beta > 0$ increases, the bias of μ towards low-energy configurations also increases. See Figure 3.2 for an illustration of the effect of β on configurations.

The Glauber dynamics for the Gibbs distribution μ move from a starting configuration σ by picking a vertex w uniformly at random from V and then generating

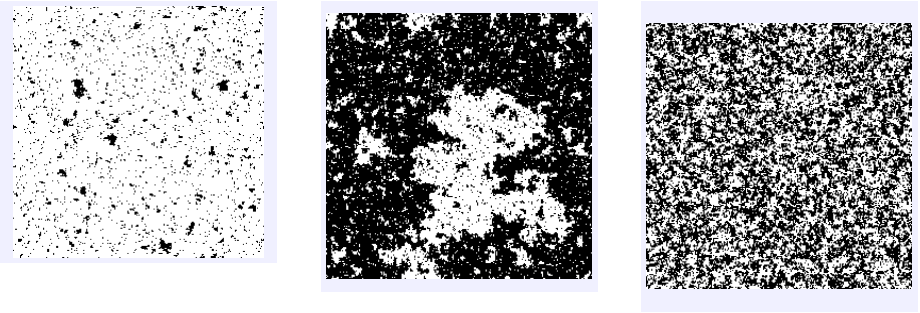


FIGURE 3.2. Glauber dynamics for the Ising model on the 250×250 torus viewed at times $t = 1{,}000$, $16{,}500$, and $1{,}000$ at low, critical, and high temperature, respectively. Simulations and graphics courtesy of Raissa D'Souza.

a new configuration according to μ conditioned on the set of configurations agreeing with σ on vertices different from w.

The reader can check that the conditional μ-probability of spin $+1$ at w is

$$p(\sigma, w) := \frac{e^{\beta S(\sigma,w)}}{e^{\beta S(\sigma,w)} + e^{-\beta S(\sigma,w)}} = \frac{1 + \tanh(\beta S(\sigma, w))}{2}, \tag{3.10}$$

where $S(\sigma, w) := \sum_{u\,:\,u\sim w} \sigma(u)$. Note that $p(\sigma, w)$ depends only on the spins at vertices adjacent to w. Therefore, the transition matrix on Ω is given by

$$P(\sigma, \sigma') = \frac{1}{|V|} \sum_{v \in V} \frac{e^{\beta\, \sigma'(w)\, S(\sigma,w)}}{e^{\beta\, \sigma'(w)\, S(\sigma,w)} + e^{-\beta\, \sigma'(w)\, S(\sigma,w)}} \cdot \mathbf{1}_{\{\sigma(v) = \sigma'(v) \text{ for } v \neq w\}}. \tag{3.11}$$

This chain has stationary distribution given by the Gibbs distribution μ.

Exercises

EXERCISE 3.1. Let Ψ be an irreducible transition matrix on Ω, and let π be a probability distribution on Ω. Show that the transition matrix

$$P(x, y) = \begin{cases} \Psi(x, y) \left[\frac{\pi(y)\Psi(y,x)}{\pi(x)\Psi(x,y)} \wedge 1 \right] & \text{if } y \neq x, \\ 1 - \sum\limits_{z\,:\,z \neq x} \Psi(x, z) \left[\frac{\pi(z)\Psi(z,x)}{\pi(x)\Psi(x,z)} \wedge 1 \right] & \text{if } y = x \end{cases}$$

defines a reversible Markov chain with stationary distribution π.

EXERCISE 3.2. Verify that the Glauber dynamics for π is a reversible Markov chain with stationary distribution π.

Notes

The Metropolis chain was introduced in Metropolis, Rosenbluth, Teller, and Teller (1953) for a specific stationary distribution. Hastings (1970) extended the

CHAPTER 4

Introduction to Markov Chain Mixing

We are now ready to discuss the long-term behavior of finite Markov chains. Since we are interested in quantifying the speed of convergence of families of Markov chains, we need to choose an appropriate metric for measuring the distance between distributions.

First we define ***total variation distance*** and give several characterizations of it, all of which will be useful in our future work. Next we prove the Convergence Theorem (Theorem 4.9), which says that for an irreducible and aperiodic chain the distribution after many steps approaches the chain's stationary distribution, in the sense that the total variation distance between them approaches 0. In the rest of the chapter we examine the effects of the initial distribution on distance from stationarity, define the *mixing time* of a chain, consider circumstances under which related chains can have identical mixing, and prove a version of the Ergodic Theorem (Theorem 4.16) for Markov chains.

4.1. Total Variation Distance

The ***total variation distance*** between two probability distributions μ and ν on Ω is defined by

$$\|\mu - \nu\|_{\text{TV}} = \max_{A \subset \Omega} |\mu(A) - \nu(A)|. \tag{4.1}$$

This definition is explicitly probabilistic: the distance between μ and ν is the maximum difference between the probabilities assigned to a single event by the two distributions.

EXAMPLE 4.1. Recall the coin-tossing frog of Example 1.1, who has probability p of jumping from east to west and probability q of jumping from west to east. His transition matrix is $\left(\begin{smallmatrix} 1-p & p \\ q & 1-q \end{smallmatrix}\right)$ and his stationary distribution is $\pi = \left(\frac{q}{p+q}, \frac{p}{p+q}\right)$. Assume the frog starts at the east pad (that is, $\mu_0 = (1, 0)$) and define

$$\Delta_t = \mu_t(e) - \pi(e).$$

Since there are only two states, there are only four possible events $A \subseteq \Omega$. Hence it is easy to check (and you should) that

$$\|\mu_t - \pi\|_{\text{TV}} = \Delta_t = P^t(e, e) - \pi(e) = \pi(w) - P^t(e, w).$$

We pointed out in Example 1.1 that $\Delta_t = (1 - p - q)^t \Delta_0$. Hence for this two-state chain, the total variation distance decreases exponentially fast as t increases. (Note that $(1 - p - q)$ is an eigenvalue of P; we will discuss connections between eigenvalues and mixing in Chapter 12.)

The definition of total variation distance (4.1) is a maximum over *all* subsets of Ω, so using this definition is not always the most convenient way to estimate

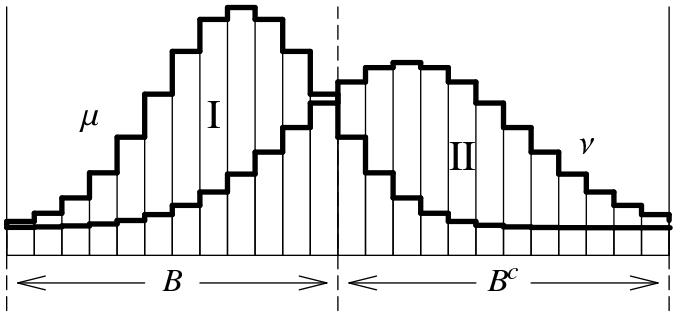


FIGURE 4.1. Recall that $B = \{x : \mu(x) > \nu(x)\}$. Region I has area $\mu(B) - \nu(B)$. Region II has area $\nu(B^c) - \mu(B^c)$. Since the total area under each of μ and ν is 1, regions I and II must have the same area—and that area is $\|\mu - \nu\|_{\text{TV}}$.

the distance. We now give three extremely useful alternative characterizations. Proposition 4.2 reduces total variation distance to a simple sum over the state space. Proposition 4.7 uses *coupling* to give another probabilistic interpretation: $\|\mu - \nu\|_{\text{TV}}$ measures how close to identical we can force two random variables realizing μ and ν to be.

PROPOSITION 4.2. *Let μ and ν be two probability distributions on Ω. Then*

$$\|\mu - \nu\|_{\text{TV}} = \frac{1}{2} \sum_{x \in \Omega} |\mu(x) - \nu(x)| . \tag{4.2}$$

PROOF. Let $B = \{x : \mu(x) \geq \nu(x)\}$ and let $A \subset \Omega$ be any event. Then

$$\mu(A) - \nu(A) \leq \mu(A \cap B) - \nu(A \cap B) \leq \mu(B) - \nu(B). \tag{4.3}$$

The first inequality is true because any $x \in A \cap B^c$ satisfies $\mu(x) - \nu(x) < 0$, so the difference in probability cannot decrease when such elements are eliminated. For the second inequality, note that including more elements of B cannot decrease the difference in probability.

By exactly parallel reasoning,

$$\nu(A) - \mu(A) \leq \nu(B^c) - \mu(B^c). \tag{4.4}$$

Fortunately, the upper bounds on the right-hand sides of (4.3) and (4.4) are actually the same (as can be seen by subtracting them; see Figure 4.1). Furthermore, when we take $A = B$ (or B^c), then $|\mu(A) - \nu(A)|$ is equal to the upper bound. Thus

$$\|\mu - \nu\|_{\text{TV}} = \frac{1}{2} [\mu(B) - \nu(B) + \nu(B^c) - \mu(B^c)] = \frac{1}{2} \sum_{x \in \Omega} |\mu(x) - \nu(x)|.$$

■

REMARK 4.3. The proof of Proposition 4.2 also shows that

$$\|\mu - \nu\|_{\text{TV}} = \sum_{\substack{x \in \Omega \\ \mu(x) \geq \nu(x)}} [\mu(x) - \nu(x)], \tag{4.5}$$

which is a useful identity.

REMARK 4.4. From Proposition 4.2 and the triangle inequality for real numbers, it is easy to see that total variation distance satisfies the triangle inequality: for probability distributions μ, ν and η,

$$\|\mu - \nu\|_{\mathrm{TV}} \leq \|\mu - \eta\|_{\mathrm{TV}} + \|\eta - \nu\|_{\mathrm{TV}} . \tag{4.6}$$

PROPOSITION 4.5. *Let μ and ν be two probability distributions on Ω. Then the total variation distance between them satisfies*

$$\begin{aligned} &\|\mu - \nu\|_{\mathrm{TV}} \\ &= \frac{1}{2} \sup \left\{ \sum_{x \in \Omega} f(x)\mu(x) - \sum_{x \in \Omega} f(x)\nu(x) \ : \ f \text{ satisfying } \max_{x \in \Omega} |f(x)| \leq 1 \right\}. \end{aligned} \tag{4.7}$$

PROOF. When f satisfies $\max_{x \in \Omega} |f(x)| \leq 1$, we have

$$\begin{aligned} \frac{1}{2} \left| \sum_{x \in \Omega} f(x)\mu(x) - \sum_{x \in \Omega} f(x)\nu(x) \right| &\leq \frac{1}{2} \sum_{x \in \Omega} |f(x)[\mu(x) - \nu(x)]| \\ &\leq \frac{1}{2} \sum_{x \in \Omega} |\mu(x) - \nu(x)| \\ &= \|\mu - \nu\|_{\mathrm{TV}} , \end{aligned}$$

which shows that the right-hand side of (4.7) is not more than $\|\mu - \nu\|_{\mathrm{TV}}$. Define

$$f^\star(x) = \begin{cases} 1 & \text{if } x \text{ satisfies } \mu(x) \geq \nu(x), \\ -1 & \text{if } x \text{ satisfies } \mu(x) < \nu(x). \end{cases}$$

Then

$$\begin{aligned} \frac{1}{2} \left[\sum_{x \in \Omega} f^\star(x)\mu(x) - \sum_{x \in \Omega} f^\star(x)\nu(x) \right] &= \frac{1}{2} \sum_{x \in \Omega} f^\star(x)[\mu(x) - \nu(x)] \\ &= \frac{1}{2} \left[\sum_{\substack{x \in \Omega \\ \mu(x) \geq \nu(x)}} [\mu(x) - \nu(x)] + \sum_{\substack{x \in \Omega \\ \nu(x) > \mu(x)}} [\nu(x) - \mu(x)] \right]. \end{aligned}$$

Using (4.5) shows that the right-hand side above equals $\|\mu - \nu\|_{\mathrm{TV}}$. Hence the right-hand side of (4.7) is at least $\|\mu - \nu\|_{\mathrm{TV}}$. ■

4.2. Coupling and Total Variation Distance

A ***coupling*** of two probability distributions μ and ν is a pair of random variables (X, Y) defined on a single probability space such that the marginal distribution of X is μ and the marginal distribution of Y is ν. That is, a coupling (X, Y) satisfies $\mathbf{P}\{X = x\} = \mu(x)$ and $\mathbf{P}\{Y = y\} = \nu(y)$.

Coupling is a general and powerful technique; it can be applied in many different ways. Indeed, Chapters 5 and 14 use couplings of entire chain trajectories to bound rates of convergence to stationarity. Here, we offer a gentle introduction by showing the close connection between couplings of two random variables and the total variation distance between those variables.

EXAMPLE 4.6. Let μ and ν both be the "fair coin" measure giving weight $1/2$ to the elements of $\{0, 1\}$.

(i) One way to couple μ and ν is to define (X, Y) to be a pair of independent coins, so that $\mathbf{P}\{X = x, Y = y\} = 1/4$ for all $x, y \in \{0, 1\}$.
(ii) Another way to couple μ and ν is to let X be a fair coin toss and define $Y = X$. In this case, $\mathbf{P}\{X = Y = 0\} = 1/2$, $\mathbf{P}\{X = Y = 1\} = 1/2$, and $\mathbf{P}\{X \neq Y\} = 0$.

Given a coupling (X, Y) of μ and ν, if q is the joint distribution of (X, Y) on $\Omega \times \Omega$, meaning that $q(x, y) = \mathbf{P}\{X = x, Y = y\}$, then q satisfies

$$\sum_{y \in \Omega} q(x, y) = \sum_{y \in \Omega} \mathbf{P}\{X = x, Y = y\} = \mathbf{P}\{X = x\} = \mu(x)$$

and

$$\sum_{x \in \Omega} q(x, y) = \sum_{x \in \Omega} \mathbf{P}\{X = x, Y = y\} = \mathbf{P}\{Y = y\} = \nu(y).$$

Conversely, given a probability distribution q on the product space $\Omega \times \Omega$ which satisfies

$$\sum_{y \in \Omega} q(x, y) = \mu(x) \quad \text{and} \quad \sum_{x \in \Omega} q(x, y) = \nu(y),$$

there is a pair of random variables (X, Y) having q as their joint distribution – and consequently this pair (X, Y) is a coupling of μ and ν. In summary, a coupling can be specified either by a pair of random variables (X, Y) defined on a common probability space or by a distribution q on $\Omega \times \Omega$.

Returning to Example 4.6, the coupling in part (i) could equivalently be specified by the probability distribution q_1 on $\{0, 1\}^2$ given by

$$q_1(x, y) = \frac{1}{4} \quad \text{for all } (x, y) \in \{0, 1\}^2.$$

Likewise, the coupling in part (ii) can be identified with the probability distribution q_2 given by

$$q_2(x, y) = \begin{cases} \frac{1}{2} & \text{if } (x, y) = (0, 0),\ (x, y) = (1, 1), \\ 0 & \text{if } (x, y) = (0, 1),\ (x, y) = (1, 0). \end{cases}$$

Any two distributions μ and ν have an independent coupling. However, when μ and ν are not identical, it will not be possible for X and Y to always have the same value. How close can a coupling get to having X and Y identical? Total variation distance gives the answer.

PROPOSITION 4.7. *Let μ and ν be two probability distributions on Ω. Then*

$$\|\mu - \nu\|_{\mathrm{TV}} = \inf \{\mathbf{P}\{X \neq Y\} : (X, Y) \text{ is a coupling of } \mu \text{ and } \nu\}. \tag{4.8}$$

REMARK 4.8. We will in fact show that there is a coupling (X, Y) which attains the infimum in (4.8). We will call such a coupling ***optimal***.

PROOF. First, we note that for any coupling (X, Y) of μ and ν and any event $A \subset \Omega$,

$$\begin{aligned} \mu(A) - \nu(A) &= \mathbf{P}\{X \in A\} - \mathbf{P}\{Y \in A\} && (4.9) \\ &\leq \mathbf{P}\{X \in A, Y \notin A\} && (4.10) \\ &\leq \mathbf{P}\{X \neq Y\}. && (4.11) \end{aligned}$$

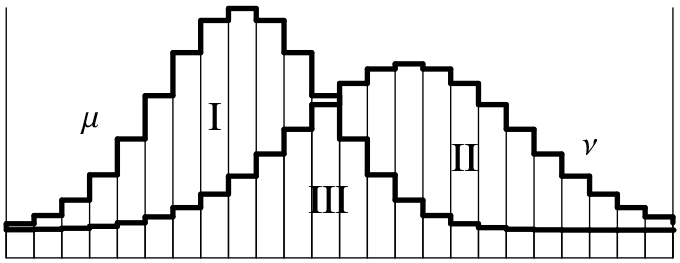


FIGURE 4.2. Since each of regions I and II has area $\|\mu-\nu\|_{\mathrm{TV}}$ and μ and ν are probability measures, region III has area $1-\|\mu-\nu\|_{\mathrm{TV}}$.

(Dropping the event $\{X \notin A, Y \in A\}$ from the second term of the difference gives the first inequality.) It immediately follows that

$$\|\mu-\nu\|_{\mathrm{TV}} \le \inf\{\mathbf{P}\{X \neq Y\} : (X,Y) \text{ is a coupling of } \mu \text{ and } \nu\}. \tag{4.12}$$

It will suffice to construct a coupling for which $\mathbf{P}\{X \neq Y\}$ is exactly equal to $\|\mu-\nu\|_{\mathrm{TV}}$. We will do so by forcing X and Y to be equal as often as they possibly can be. Consider Figure 4.2. Region III, bounded by $\mu(x)\wedge\nu(x) = \min\{\mu(x),\nu(x)\}$, can be seen as the overlap between the two distributions. Informally, our coupling proceeds by choosing a point in the union of regions I, II, and III. Whenever we "land" in region III, we take $X = Y$. Otherwise, we accept that X must be in region I and Y must be in region II; since those regions have disjoint support, X and Y cannot be equal.

More formally, we use the following procedure to generate X and Y. Let

$$p = \sum_{x\in\Omega} \mu(x)\wedge\nu(x).$$

Write

$$\sum_{x\in\Omega} \mu(x)\wedge\nu(x) = \sum_{\substack{x\in\Omega,\\ \mu(x)\le\nu(x)}} \mu(x) + \sum_{\substack{x\in\Omega,\\ \mu(x)>\nu(x)}} \nu(x).$$

Adding and subtracting $\sum_{x\,:\,\mu(x)>\nu(x)} \mu(x)$ to the right-hand side above shows that

$$\sum_{x\in\Omega} \mu(x)\wedge\nu(x) = 1 - \sum_{\substack{x\in\Omega,\\ \mu(x)>\nu(x)}} [\mu(x)-\nu(x)].$$

By equation (4.5) and the immediately preceding equation,

$$\sum_{x\in\Omega} \mu(x)\wedge\nu(x) = 1 - \|\mu-\nu\|_{\mathrm{TV}} = p. \tag{4.13}$$

Flip a coin with probability of heads equal to p.

(i) If the coin comes up heads, then choose a value Z according to the probability distribution

$$\gamma_{\mathrm{III}}(x) = \frac{\mu(x)\wedge\nu(x)}{p},$$

and set $X = Y = Z$.

(ii) If the coin comes up tails, choose X according to the probability distribution

$$\gamma_{\rm II}(x) = \begin{cases} \frac{\nu(x)-\mu(x)}{\|\mu-\nu\|_{\rm TV}} & \text{if } \nu(x) > \mu(x), \\ 0 & \text{otherwise.} \end{cases}$$

and independently choose Y according to the probability distribution

$$\gamma_{\rm III}(x) = \begin{cases} \frac{\mu(x)-\nu(x)}{\|\mu-\nu\|_{\rm TV}} & \text{if } \mu(x) > \nu(x), \\ 0 & \text{otherwise.} \end{cases}$$

Note that (4.5) ensures that $\gamma_{\rm I}$ and $\gamma_{\rm II}$ are probability distributions.

Clearly,

$$p\gamma_{\rm III} + (1-p)\gamma_{\rm I} = \mu,$$
$$p\gamma_{\rm III} + (1-p)\gamma_{\rm II} = \nu,$$

so that the distribution of X is μ and the distribution of Y is ν. Note that in the case that the coin lands tails up, $X \neq Y$ since $\gamma_{\rm I}$ and $\gamma_{\rm II}$ are positive on disjoint subsets of Ω. Thus $X = Y$ if and only if the coin toss is heads. We conclude that

$$\mathbf{P}\{X \neq Y\} = \|\mu - \nu\|_{\rm TV}\,.$$

■

4.3. The Convergence Theorem

We are now ready to prove that irreducible, aperiodic Markov chains converge to their stationary distributions—a key step, as much of the rest of the book will be devoted to estimating the rate at which this convergence occurs. The assumption of aperiodicity is indeed necessary—recall the even n-cycle of Example 1.4.

As is often true of such fundamental facts, there are many proofs of the Convergence Theorem. The one given here decomposes the chain into a mixture of repeated independent sampling from the stationary distribution and another Markov chain. See Exercise 5.1 for another proof using two coupled copies of the chain.

THEOREM 4.9 (Convergence Theorem). *Suppose that P is irreducible and aperiodic, with stationary distribution π. Then there exist constants $\alpha \in (0,1)$ and $C > 0$ such that*

$$\max_{x\in\Omega} \left\|P^t(x,\cdot) - \pi\right\|_{\rm TV} \leq C\alpha^t. \tag{4.14}$$

PROOF. Since P is irreducible and aperiodic, by Proposition 1.7 there exists an r such that P^r has strictly positive entries. Let Π be the matrix with $|\Omega|$ rows, each of which is the row vector π. For sufficiently small $\delta > 0$, we have

$$P^r(x,y) \geq \delta\pi(y)$$

for all $x, y \in \Omega$. Let $\theta = 1 - \delta$. The equation

$$P^r = (1-\theta)\Pi + \theta Q \tag{4.15}$$

defines a stochastic matrix Q.

It is a straightforward computation to check that $\boldsymbol{M}\Pi = \Pi$ for any stochastic matrix $\boldsymbol{M}$ and that $\Pi\boldsymbol{M} = \Pi$ for any matrix $\boldsymbol{M}$ such that $\pi\boldsymbol{M} = \pi$.

Next, we use induction to demonstrate that

$$P^{rk} = \left(1-\theta^k\right)\Pi + \theta^k Q^k \tag{4.16}$$

for $k \geq 1$. If $k = 1$, this holds by (4.15). Assuming that (4.16) holds for $k = n$,

$$P^{r(n+1)} = P^{rn}P^r = \left[(1-\theta^n)\,\Pi + \theta^n Q^n\right] P^r. \tag{4.17}$$

Distributing and expanding P^r in the second term (using (4.15)) gives

$$P^{r(n+1)} = \left[1-\theta^n\right]\Pi P^r + (1-\theta)\theta^n Q^n \Pi + \theta^{n+1} Q^n Q. \tag{4.18}$$

Using that $\Pi P^r = \Pi$ and $Q^n \Pi = \Pi$ shows that

$$P^{r(n+1)} = \left[1-\theta^{n+1}\right]\Pi + \theta^{n+1} Q^{n+1}. \tag{4.19}$$

This establishes (4.16) for $k = n+1$ (assuming it holds for $k = n$), and hence it holds for all k.

Multiplying by P^j and rearranging terms now yields

$$P^{rk+j} - \Pi = \theta^k \left(Q^k P^j - \Pi\right). \tag{4.20}$$

To complete the proof, sum the absolute values of the elements in row x_0 on both sides of (4.20) and divide by 2. On the right, the second factor is at most the largest possible total variation distance between distributions, which is 1. Hence for any x_0 we have

$$\left\|P^{rk+j}(x_0,\cdot) - \pi\right\|_{TV} \leq \theta^k. \tag{4.21}$$

■

REMARK 4.10. Because of Theorem 4.9, the distribution π is also called the ***equilibrium distribution.***

4.4. Standardizing Distance from Stationarity

Bounding the maximal distance (over $x_0 \in \Omega$) between $P^t(x_0,\cdot)$ and π is among our primary objectives. It is therefore convenient to define

$$d(t) := \max_{x\in\Omega} \left\|P^t(x,\cdot) - \pi\right\|_{TV}. \tag{4.22}$$

We will see in Chapter 5 that it is often possible to bound $\|P^t(x,\cdot) - P^t(y,\cdot)\|_{TV}$, uniformly over all pairs of states (x,y). We therefore make the definition

$$\bar{d}(t) := \max_{x,y\in\Omega} \left\|P^t(x,\cdot) - P^t(y,\cdot)\right\|_{TV}. \tag{4.23}$$

The relationship between d and $\bar{d}$ is given below:

LEMMA 4.11. *If $d(t)$ and $\bar{d}(t)$ are as defined in (4.22) and (4.23), respectively, then*

$$d(t) \leq \bar{d}(t) \leq 2d(t). \tag{4.24}$$

PROOF. It is immediate from the triangle inequality for the total variation distance that $\bar{d}(t) \leq 2d(t)$.

To show that $d(t) \leq \bar{d}(t)$, note first that since π is stationary, we have $\pi(A) = \sum_{y\in\Omega} \pi(y)P^t(y,A)$ for any set A. (This is the definition of stationarity if A is a singleton $\{x\}$. To get this for arbitrary A, just sum over the elements in A.) Using this shows that

$$\begin{aligned}\left\|P^t(x,\cdot) - \pi\right\|_{TV} &= \max_{A\subset\Omega} \left|P^t(x,A) - \pi(A)\right| \\ &= \max_{A\subset\Omega} \left|\sum_{y\in\Omega} \pi(y)\left[P^t(x,A) - P^t(y,A)\right]\right|.\end{aligned}$$

We can use the triangle inequality and the fact that the maximum of a sum is not larger than the sum over a maximum to bound the right-hand side above by

$$\begin{aligned}\max_{A\subset\Omega}\sum_{y\in\Omega}\pi(y)|P^t(x,A)-P^t(y,A)| &\le \sum_{y\in\Omega}\pi(y)\max_{A\subset\Omega}|P^t(x,A)-P^t(y,A)| \\ &= \sum_{y\in\Omega}\pi(y)\left\|P^t(x,\cdot)-P^t(y,\cdot)\right\|_{\mathrm{TV}}. \end{aligned} \tag{4.25}$$

Since a weighted average of a set of numbers is never larger than its maximum element, the right-hand side of (4.25) is bounded by $\max_{y\in\Omega}\|P^t(x,\cdot)-P^t(y,\cdot)\|_{\mathrm{TV}}$. ∎

Let $\mathcal{P}$ denote the collection of all probability distributions on Ω. Exercise 4.1 asks the reader to prove the following equalities:

$$\begin{aligned} d(t) &= \sup_{\mu\in\mathcal{P}}\left\|\mu P^t-\pi\right\|_{\mathrm{TV}}, \\ \bar{d}(t) &= \sup_{\mu,\nu\in\mathcal{P}}\left\|\mu P^t-\nu P^t\right\|_{\mathrm{TV}}. \end{aligned}$$

LEMMA 4.12. *The function $\bar{d}$ is submultiplicative: $\bar{d}(s+t)\le\bar{d}(s)\bar{d}(t)$.*

PROOF. Fix $x,y\in\Omega$, and let (X_s,Y_s) be the optimal coupling of $P^s(x,\cdot)$ and $P^s(y,\cdot)$ whose existence is guaranteed by Proposition 4.7. Hence

$$\|P^s(x,\cdot)-P^s(y,\cdot)\|_{\mathrm{TV}}=\mathbf{P}\{X_s\neq Y_s\}.$$

As P^{s+t} is the matrix product of P^t and P^s and the distribution of X_s is $P^s(x,\cdot)$, we have

$$P^{s+t}(x,w)=\sum_z P^s(x,z)P^t(z,w)=\sum_z\mathbf{P}\{X_s=z\}P^t(z,w)=\mathbf{E}\left(P^t(X_s,w)\right). \tag{4.26}$$

Combining this with the similar identity $P^{s+t}(y,w)=\mathbf{E}\left(P^t(Y_s,w)\right)$ allows us to write

$$\begin{aligned} P^{s+t}(x,w)-P^{s+t}(y,w) &= \mathbf{E}\left(P^t(X_s,w)\right)-\mathbf{E}\left(P^t(Y_s,w)\right) \\ &= \mathbf{E}\left(P^t(X_s,w)-P^t(Y_s,w)\right). \end{aligned} \tag{4.27}$$

Combining the expectations is possible since X_s and Y_s are defined together on the same probability space.

Summing (4.27) over $w\in\Omega$ and applying Proposition 4.2 shows that

$$\left\|P^{s+t}(x,\cdot)-P^{s+t}(y,\cdot)\right\|_{\mathrm{TV}}=\frac{1}{2}\sum_w\left|\mathbf{E}\left(P^t(X_s,w)-P^t(Y_s,w)\right)\right|. \tag{4.28}$$

The right-hand side above is less than or equal to

$$\mathbf{E}\left(\frac{1}{2}\sum_w\left|P^t(X_s,w)-P^t(Y_s,w)\right|\right). \tag{4.29}$$

Applying Proposition 4.2 again, we see that the quantity inside the expectation is exactly the distance $\|P^t(X_s,\cdot)-P^t(Y_s,\cdot)\|_{\mathrm{TV}}$, which is zero whenever $X_s=Y_s$. Moreover, this distance is always bounded by $\bar{d}(t)$. This shows that

$$\left\|P^{s+t}(x,\cdot)-P^{s+t}(y,\cdot)\right\|_{\mathrm{TV}}\le\bar{d}(t)\mathbf{E}\left(\mathbf{1}_{\{X_s\neq Y_s\}}\right)=\bar{d}(t)\mathbf{P}\{X_s\neq Y_s\}. \tag{4.30}$$

Finally, since (X_s,Y_s) is an optimal coupling, the probability on the right-hand side is equal to $\|P^s(x,\cdot)-P^s(y,\cdot)\|_{\mathrm{TV}}$. Maximizing over x,y completes the proof. ∎

Exercise 4.3 implies that $\bar{d}(t)$ is non-increasing in t. From this and Lemma 4.12 it follows that when c is any non-negative integer and t is any non-negative integer, we have

$$d(ct) \le \bar{d}(ct) \le \bar{d}(t)^c. \tag{4.31}$$

4.5. Mixing Time

It is useful to introduce a parameter which measures the time required by a Markov chain for the distance to stationarity to be small. The ***mixing time*** is defined by

$$t_{\text{mix}}(\varepsilon) := \min\{t \ : \ d(t) \le \varepsilon\} \tag{4.32}$$

and

$$t_{\text{mix}} := t_{\text{mix}}(1/4). \tag{4.33}$$

Lemma 4.11 and (4.31) show that when ℓ is a non-negative integer,

$$d(\, \ell t_{\text{mix}}(\varepsilon)\,) \le \bar{d}(\, \ell t_{\text{mix}}(\varepsilon)\,) \le \bar{d}(\, t_{\text{mix}}(\varepsilon)\,)^\ell \le (2\varepsilon)^\ell. \tag{4.34}$$

In particular, taking $\varepsilon = 1/4$ above yields

$$d(\, \ell t_{\text{mix}}\,) \le 2^{-\ell} \tag{4.35}$$

and

$$t_{\text{mix}}(\varepsilon) \le \lceil \log_2 \varepsilon^{-1} \rceil \, t_{\text{mix}}. \tag{4.36}$$

Thus, although the choice of $1/4$ is arbitrary in the definition (4.33) of t_{mix}, a value of ε less than $1/2$ is needed to make the inequality $d(\, \ell t_{\text{mix}}(\varepsilon)\,) \le (2\varepsilon)^\ell$ in (4.34) non-trivial and to achieve an inequality of the form (4.36).

4.6. Mixing and Time Reversal

For a distribution μ on a group G, the ***inverse distribution*** $\widehat{\mu}$ is defined by $\widehat{\mu}(g) := \mu(g^{-1})$ for all $g \in G$. Let P be the transition matrix of the random walk with increment distribution μ. Then the random walk with increment distribution $\widehat{\mu}$ is exactly the time reversal $\widehat{P}$ (defined in (1.33)) of P.

In Proposition 2.14 we noted that when $\widehat{\mu} = \mu$, the random walk on G with increment distribution μ is reversible, so that $P = \widehat{P}$. Even when μ is not a symmetric distribution, however, the forward and reversed walks must be at the same distance from stationarity, as we will find useful in analyzing card shuffling in Chapters 6 and 8.

LEMMA 4.13. *Let P be the transition matrix of a random walk on a group G with increment distribution μ and let $\widehat{P}$ be that of the walk on G with increment distribution $\widehat{\mu}$. Let π be the uniform distribution on G. Then for any $t \ge 0$*

$$\left\| P^t(\text{id}, \cdot) - \pi \right\|_{TV} = \left\| \widehat{P}^t(\text{id}, \cdot) - \pi \right\|_{TV}.$$

PROOF. Let $(X_t) = (\text{id}, X_1, \dots)$ be a Markov chain with transition matrix P and initial state id. We can write $X_k = g_1 g_2 \dots g_k$, where the random elements $g_1, g_2, \dots \in G$ are independent choices from the distribution μ. Similarly, let (Y_t)

Putting together (4.41) and (4.42) shows that

$$\mathbf{P}_x\left\{\lim_{n\to\infty}\frac{S_{\tau^+_{x,n}}}{\tau^+_{x,n}} = E_\pi(f)\right\} = 1.$$

Exercise 4.2 shows that (4.40) holds when $\mu = \delta_x$, the probability distribution with unit mass at x. Averaging over the starting state completes the proof. ■

Taking $f(y) = \delta_x(y) = \mathbf{1}_{\{y=x\}}$ in Theorem 4.16 shows that

$$\mathbf{P}_\mu\left\{\lim_{t\to\infty}\frac{1}{t}\sum_{s=0}^{t-1}\mathbf{1}_{\{X_s=x\}} = \pi(x)\right\} = 1,$$

so the asymptotic proportion of time the chain spends in state x equals $\pi(x)$.

Exercises

EXERCISE 4.1. Prove that

$$d(t) = \sup_\mu \left\|\mu P^t - \pi\right\|_{\mathrm{TV}},$$
$$\bar{d}(t) = \sup_{\mu,\nu} \left\|\mu P^t - \nu P^t\right\|_{\mathrm{TV}},$$

where μ and ν vary over probability distributions on a finite set Ω.

EXERCISE 4.2. Let (a_n) be a bounded sequence. If, for a sequence of integers (n_k) satisfying $\lim_{k\to\infty} n_k/n_{k+1} = 1$, we have

$$\lim_{k\to\infty}\frac{a_1+\cdots+a_{n_k}}{n_k} = a,$$

then

$$\lim_{n\to\infty}\frac{a_1+\cdots+a_n}{n} = a.$$

EXERCISE 4.3. Let P be the transition matrix of a Markov chain with state space Ω and let μ and ν be any two distributions on Ω. Prove that

$$\|\mu P - \nu P\|_{\mathrm{TV}} \le \|\mu - \nu\|_{\mathrm{TV}}.$$

(This in particular shows that $\left\|\mu P^{t+1} - \pi\right\|_{\mathrm{TV}} \le \|\mu P^t - \pi\|_{\mathrm{TV}}$, that is, advancing the chain can only move it closer to stationarity.)

EXERCISE 4.4. Let P be the transition matrix of a Markov chain with stationary distribution π. Prove that for any $t \ge 0$,

$$d(t+1) \le d(t),$$

where $d(t)$ is defined by (4.22).

EXERCISE 4.5. For $i = 1, \ldots, n$, let μ_i and ν_i be measures on Ω_i, and define measures μ and ν on $\prod_{i=1}^n \Omega_i$ by $\mu := \prod_{i=1}^n \mu_i$ and $\nu := \prod_{i=1}^n \nu_i$. Show that

$$\|\mu - \nu\|_{\mathrm{TV}} \le \sum_{i=1}^n \|\mu_i - \nu_i\|_{\mathrm{TV}}.$$

CHAPTER 5

Coupling

5.1. Definition

As we defined in Section 4.1, a coupling of two probability distributions μ and ν is a pair of random variables (X, Y), defined on the same probability space, such that the marginal distribution of X is μ and the marginal distribution of Y is ν.

Couplings are useful because a comparison between distributions is reduced to a comparison between random variables. Proposition 4.7 characterized $\|\mu - \nu\|_{\text{TV}}$ as the minimum, over all couplings (X, Y) of μ and ν, of the probability that X and Y are different. This provides an effective method of obtaining upper bounds on the distance.

In this chapter, we will extract more information by coupling not only pairs of distributions, but entire Markov chain trajectories. Here is a simple initial example.

EXAMPLE 5.1. A simple random walk on the segment $\{0, 1, \ldots, n\}$ is a Markov chain which moves either up or down at each move with equal probability. If the walk attempts to move outside the interval when at a boundary point, it stays put. It is intuitively clear that $P^t(x, n) \leq P^t(y, n)$ whenever $x \leq y$, as this says that the chance of being at the "top" value n after t steps does not decrease as you increase the height of the starting position.

A simple proof uses a coupling of the distributions $P^t(x, \cdot)$ and $P^t(y, \cdot)$. Let $\Delta_1, \Delta_2, \ldots$ be a sequence of i.i.d. (that is, independent and identically distributed) $\{-1, 1\}$-valued random variables with zero mean, so each Δ_i is equally likely to be $+1$ as -1. We will define together two random walks on $\{0, 1, \ldots, n\}$: the walk (X_t) starts at x, while the walk (Y_t) starts at y.

We use the same rule for moving in both chains (X_t) and (Y_t): if $\Delta_t = +1$, move the chain up if possible, and if $\Delta_t = -1$, move the chain down if possible. Hence the chains move in step, although they are started at different heights. Once the two chains meet (necessarily either at 0 or n), they stay together thereafter.

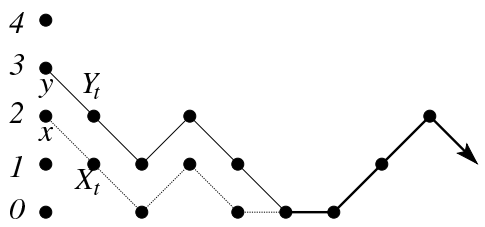


FIGURE 5.1. Coupled random walks on $\{0, 1, 2, 3, 4\}$. The walks stay together after meeting.

Clearly the distribution of X_t is $P^t(x,\cdot)$, and the distribution of Y_t is $P^t(y,\cdot)$. Importantly, X_t and Y_t are defined on the same underlying probability space, as both chains use the sequence (Δ_t) to determine their moves.

It is clear that if $x \le y$, then $X_t \le Y_t$ for all t. In particular, if $X_t = n$, the top state, then it must be that $Y_t = n$ also. From this we can conclude that

$$P^t(x,n) = \mathbf{P}\{X_t = n\} \le \mathbf{P}\{Y_t = n\} = P^t(y,n). \tag{5.1}$$

This argument shows the power of coupling. We were able to couple together the two chains in such a way that $X_t \le Y_t$ always, and from this fact about the random variables we could easily read off information about the distributions.

In the rest of this chapter, we will see how building two simultaneous copies of a Markov chain using a common source of randomness, as we did in the previous example, can be useful for getting bounds on the distance to stationarity.

We define a ***coupling of Markov chains*** with transition matrix P to be a process $(X_t, Y_t)_{t=0}^{\infty}$ with the property that both (X_t) and (Y_t) are Markov chains with transition matrix P, although the two chains may possibly have different starting distributions.

Any coupling of Markov chains with transition matrix P can be modified so that the two chains stay together at all times after their first simultaneous visit to a single state—more precisely, so that

$$\text{if } X_s = Y_s, \text{ then } X_t = Y_t \text{ for } t \ge s. \tag{5.2}$$

To construct a coupling satisfying (5.2), simply run the chains according to the original coupling until they meet; then run them together.

NOTATION. If (X_t) and (Y_t) are coupled Markov chains with $X_0 = x$ and $Y_0 = y$, then we will often write $\mathbf{P}_{x,y}$ for the probability on the space where (X_t) and (Y_t) are both defined.

5.2. Bounding Total Variation Distance

As usual, we will fix an irreducible transition matrix P on the state space Ω and write π for its stationary distribution. The following is the key tool used in this chapter.

THEOREM 5.2. *Let $\{(X_t, Y_t)\}$ be a coupling satisfying (5.2) for which $X_0 = x$ and $Y_0 = y$. Let τ_{couple} be the first time the chains meet:*

$$\tau_{\text{couple}} := \min\{t : X_t = Y_t\}. \tag{5.3}$$

Then

$$\left\|P^t(x,\cdot) - P^t(y,\cdot)\right\|_{TV} \le \mathbf{P}_{x,y}\{\tau_{\text{couple}} > t\}. \tag{5.4}$$

PROOF. Notice that $P^t(x,z) = \mathbf{P}_{x,y}\{X_t = z\}$ and $P^t(y,z) = \mathbf{P}_{x,y}\{Y_t = z\}$. Consequently, (X_t, Y_t) is a coupling of $P^t(x,\cdot)$ with $P^t(y,\cdot)$, whence Proposition 4.7 implies that

$$\left\|P^t(x,\cdot) - P^t(y,\cdot)\right\|_{\text{TV}} \le \mathbf{P}_{x,y}\{X_t \ne Y_t\}. \tag{5.5}$$

By (5.2), $\mathbf{P}_{x,y}\{X_t \ne Y_t\} = \mathbf{P}_{x,y}\{\tau_{\text{couple}} > t\}$, which with (5.5) establishes (5.4). ∎

Combining Theorem 5.2 with Lemma 4.11 proves the following:

COROLLARY 5.3. *Suppose that for each pair of states $x, y \in \Omega$ there is a coupling (X_t, Y_t) with $X_0 = x$ and $Y_0 = y$. For each such coupling, let τ_{couple} be the first time the chains meet, as defined in (5.3). Then*

$$d(t) \leq \max_{x,y\in\Omega} \mathbf{P}_{x,y}\{\tau_{\text{couple}} > t\}.$$

Given a Markov chain on Ω with transition matrix P, a ***Markovian coupling*** of P is a Markov chain with state space $\Omega \times \Omega$ whose transition matrix Q satisfies

(i) for all x, y, x' we have $\sum_{y'} Q((x,y),(x',y')) = P(x,x')$ and
(ii) for all x, y, y' we have $\sum_{x'} Q((x,y),(x',y')) = P(y,y')$.

Clearly any Markovian coupling is indeed a coupling of Markov chains, as we defined in Section 5.1.

REMARK 5.4. All couplings used in this book will be Markovian.

5.3. Examples

5.3.1. Random walk on the cycle. We defined random walk on the n-cycle in Example 1.4. The underlying graph of this walk, $\mathbb{Z}_n$, has vertex set $\{1, 2, \ldots, n\}$ and edges between j and k whenever $j \equiv k \pm 1 \mod n$. See Figure 1.3.

We consider the lazy walk, which remains in its current position with probability $1/2$, moves clockwise with probability $1/4$, and moves counterclockwise with probability $1/4$.

We construct a coupling (X_t, Y_t) of two particles performing lazy walks on $\mathbb{Z}_n$, one started from x and the other started from y. In this coupling, the two particles will never move simultaneously, ensuring that they will not jump over one another when they come to within unit distance. At each move, a fair coin is tossed, independent of all previous tosses. If heads, the chain (X_t) moves one step, the direction of which is determined by another fair coin toss, again independent of all other previous tosses. If tails, the chain (Y_t) moves one step, also determined by an independent fair coin toss. Once the two particles collide, thereafter they make identical moves. Let D_t be the clockwise distance between the two particles. Note that D_t is a simple random walk on the interior vertices of $\{0, 1, 2, \ldots, n\}$ and gets absorbed at either 0 or n. By Proposition 2.1, if $\tau = \min\{t \geq 0 : D_t \in \{0, n\}\}$, then $\mathbf{E}_{x,y}(\tau) = k(n-k)$, where k is the clockwise distance between x and y. Since $\tau = \tau_{\text{couple}}$, by Corollary 5.3,

$$d(t) \leq \max_{x,y\in\mathbb{Z}_n} \mathbf{P}_{x,y}\{\tau > t\} \leq \frac{\max_{x,y} \mathbf{E}_{x,y}(\tau)}{t} \leq \frac{n^2}{4t}.$$

The right-hand side equals $1/4$ for $t = n^2$, whence $t_{\text{mix}} \leq n^2$.

In Section 7.4.1, it is shown that $t_{\text{mix}} \geq c_1 n^2$ for a constant c_1.

5.3.2. Random walk on the torus. The *d-dimensional torus* is graph whose vertex set is the Cartesian product

$$\mathbb{Z}_n^d = \underbrace{\mathbb{Z}_n \times \cdots \times \mathbb{Z}_n}_{d \text{ times}}.$$

Vertices $\boldsymbol{x} = (x^1, \ldots, x^d)$ and $\boldsymbol{y} = (y^1, y^2, \ldots, y^d)$ are neighbors in $\mathbb{Z}_n^d$ if for some $j \in \{1, 2, \ldots, n\}$, we have $x^i = y^i$ for all $i \neq j$ and $x^j \equiv y^j \pm 1 \mod n$. See Figure 5.2 for an example.

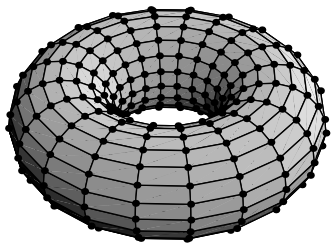


FIGURE 5.2. The 2-torus $\mathbb{Z}_{20}^2$.

When n is even, the graph $\mathbb{Z}_n^d$ is bipartite and the associated random walk is periodic. To avoid this complication, we consider the lazy random walk on $\mathbb{Z}_n^d$, defined in Section 1.3. This walk remains at its current position with probability $1/2$ at each move.

We now use coupling to bound the mixing time of the lazy random walk on $\mathbb{Z}_n^d$.

THEOREM 5.5. *For the lazy random walk on the d-dimension torus $\mathbb{Z}_n^d$,*

$$t_{\text{mix}}(\varepsilon) \leq c(d)n^2 \log_2(\varepsilon^{-1}), \tag{5.6}$$

where $c(d)$ is a constant depending on the dimension d.

PROOF. In order to apply Corollary 5.3 to prove this theorem, we construct a coupling for each pair $(\boldsymbol{x}, \boldsymbol{y})$ of starting states and bound the expected value of the coupling time $\tau_{\text{couple}} = \tau_{\boldsymbol{x},\boldsymbol{y}}$.

To couple together a random walk $(\boldsymbol{X}_t)$ started at $\boldsymbol{x}$ with a random walk $(\boldsymbol{Y}_t)$ started at $\boldsymbol{y}$, first pick one of the d coordinates at random. If the positions of the two walks agree in the chosen coordinate, we move both of the walks by $+1$, -1, or 0 in that coordinate, with probabilities $1/4$, $1/4$ and $1/2$, respectively. If the positions of the two walks differ in the chosen coordinate, we randomly choose one of the chains to move, leaving the other fixed. We then move the selected walk by $+1$ or -1 in the chosen coordinate, with the sign determined by a fair coin toss.

Let $\boldsymbol{X}_t = (X_t^1, \ldots, X_t^d)$ and $\boldsymbol{Y}_t = (Y_t^1, \ldots, Y_t^d)$, and let

$$\tau_i := \min\{t \geq 0 : X_t^i = Y_t^i\}$$

be the time required for the chains to agree in coordinate i.

The clockwise difference between X_t^i and Y_t^i, viewed at the times when coordinate i is selected, behaves just as the coupling of the lazy walk on the cycle $\mathbb{Z}_n$ discussed above. Thus, the expected number of moves in coordinate i needed to make the two chains agree on that coordinate is not more than $n^2/4$.

Since coordinate i is selected with probability $1/d$ at each move, there is a geometric waiting time between moves with expectation d. Exercise 5.3 implies that

$$\mathbf{E}_{\boldsymbol{x},\boldsymbol{y}}(\tau_i) \leq \frac{dn^2}{4}. \tag{5.7}$$

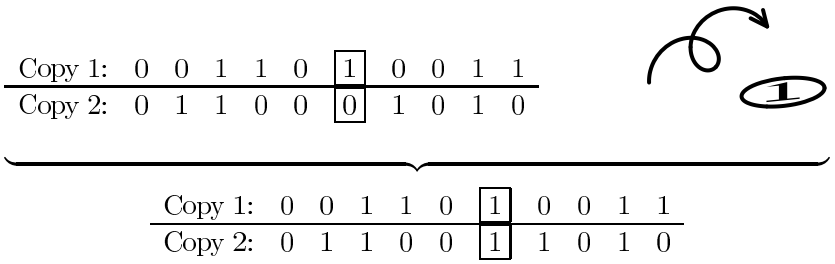


FIGURE 5.3. One step in two coupled lazy walks on the hypercube. First, choose a coordinate to update—here, the sixth. Then, flip a 0/1 coin and use the result to update the chosen coordinate to the same value in both walks.

The coupling time we are interested in is $\tau_{\text{couple}} = \max_{1\le i\le d} \tau_i$, and we can bound the maximum by a sum to get

$$\mathbf{E}_{\boldsymbol{x},\boldsymbol{y}}(\tau_{\text{couple}}) \le \frac{d^2n^2}{4}. \tag{5.8}$$

This bound is independent of the starting states, and we can use Markov's inequality to show that

$$\mathbf{P}_{\boldsymbol{x},\boldsymbol{y}}\{\tau_{\text{couple}} > t\} \le \frac{\mathbf{E}_{\boldsymbol{x},\boldsymbol{y}}(\tau_{\text{couple}})}{t} \le \frac{1}{t}\frac{d^2n^2}{4}. \tag{5.9}$$

Taking $t_0 = d^2n^2$ shows that $d(t_0) \le 1/4$, and so $t_{\text{mix}} \le d^2n^2$. By (4.36),

$$t_{\text{mix}}(\varepsilon) \le d^2n^2 \left\lceil \log(\varepsilon^{-1}) \right\rceil,$$

and we have proved the theorem. ■

Exercise 5.4 shows that the bound on $c(d)$ can be improved.

5.3.3. Random walk on the hypercube. The simple random walk on the hypercube $\{0,1\}^n$ was defined in Section 2.3: this is the simple walker on the graph having vertex set $\{0,1\}^n$, the binary words of length n, and with edges connecting words differing in exactly one letter. (Note that this graph is also the torus $\mathbb{Z}_2^n$.)

To avoid periodicity, we study the lazy chain: at each time step, the walker remains at her current position with probability 1/2 and with probability 1/2 moves to a position chosen uniformly at random among all neighboring vertices.

As remarked in Section 2.3, a convenient way to generate the lazy walk is as follows: pick one of the n coordinates uniformly at random, and *refresh* the bit at this coordinate with a random fair bit (one which equals 0 or 1 each with probability 1/2).

This algorithm for running the walk leads to the following coupling of two walks with possibly different starting positions: first, pick among the n coordinates uniformly at random; suppose that coordinate i is selected. *In both walks*, replace the bit at coordinate i *with the same* random fair bit. (See Figure 5.3.) From this time onwards, both walks will agree in the i-th coordinate. A moment's thought reveals that individually each of the walks is indeed a lazy random walk on the hypercube.

If τ is the first time when all of the coordinates have been selected at least once, then the two walkers agree with each other from time τ onwards. (If the

initial states agree in some coordinates, the first time the walkers agree could be strictly before τ.) The distribution of τ is exactly the same as the coupon collector random variable studied in Section 2.2. Using Corollary 5.3, together with the bound on the tail of τ given in Proposition 2.4, shows that

$$d(n\log n + cn) \leq \mathbf{P}\{\tau > n\log n + cn\} \leq e^{-c}.$$

It is immediate from the above that

$$t_{\text{mix}}(\varepsilon) \leq n\log n + \log(1/\varepsilon)n. \tag{5.10}$$

Simply, $t_{\text{mix}} = O(n\log n)$. The bound in (5.10) is off by a factor of two and will be sharpened in Section 18.2.2 via a more sophisticated coupling.

5.3.4. Random walk on a finite binary tree. Since trees will appear in several examples in the sequel, we collect some definitions here. A ***tree*** is a connected graph with no cycles. (See Exercise 1.3 and Exercise 1.4.) A ***rooted*** tree has a distinguished vertex, called the ***root***. The ***depth*** of a vertex v is its graph distance to the root. A ***level*** of the tree consists of all vertices at the same depth. The ***children*** of v are the neighbors of v with depth larger than v. A ***leaf*** is a vertex with degree one.

A ***rooted finite b-ary tree of depth k***, denoted by $T_{b,k}$, is a tree with a distinguished vertex v_0, the root, such that

- v_0 has degree b,
- every vertex at distance j from the root, where $1 \leq j \leq k-1$, has degree $b+1$,
- the vertices at distance k from v_0 are leaves.

There are $n = (b^{k+1}-1)/(b-1)$ vertices in $T_{b,k}$.

In this example, we consider the lazy random walk on the finite ***binary tree*** $T_{2,k}$; this walk remains at its current position with probability 1/2.

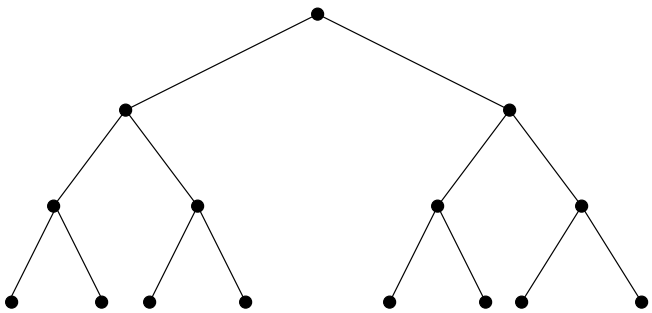


FIGURE 5.4. A binary tree of height 3.

Consider the following coupling (X_t, Y_t) of two lazy random walks, started from states x_0 and y_0 on the tree. Assume, without loss of generality, that x_0 is at least as close to the root as y_0. At each move, toss a fair coin to decide which of the two chains moves: if heads, $Y_{t+1} = Y_t$, while X_{t+1} is chosen from the neighbors of X_t uniformly at random. If the coin toss is tails, then $X_{t+1} = X_t$, and Y_{t+1} is chosen from the neighbors of Y_t uniformly at random. Run the two chains according to this rule until the first time they are at the same level of the tree. Once the two

(b) If in (a) we take $\nu = \pi$, where π is the stationary distribution, then (by definition) $\pi P^t = \pi$, and (5.18) bounds the difference between μP^t and π. The only thing left to check is that there exists a coupling guaranteed to coalesce, that is, for which $\mathbf{P}\{\tau_{\text{couple}} < \infty\} = 1$. Show that if the chains (X_t) and (Y_t) are taken to be independent of one another, then they are assured to eventually meet.

EXERCISE 5.2. Let (X_t, Y_t) be a Markovian coupling such that for some $0 < \alpha < 1$ and some $t_0 > 0$, the coupling time $\tau_{\text{couple}} = \min\{t \geq 0 : X_t = Y_t\}$ satisfies $\mathbf{P}\{\tau_{\text{couple}} \leq t_0\} \geq \alpha$ for *all* pairs of initial states (x, y). Prove that

$$\mathbf{E}(\tau_{\text{couple}}) \leq \frac{t_0}{\alpha}.$$

EXERCISE 5.3. Show that if $X_1, X_2, \ldots$ are independent and each have mean μ and if τ is a $\mathbb{Z}^+$-valued random variable independent of all the X_i's, then

$$\mathbf{E}\left(\sum_{i=1}^{\tau} X_i\right) = \mu \mathbf{E}(\tau).$$

EXERCISE 5.4. We can get a better bound on the mixing time for the lazy walker on the d-dimensional torus by sharpening the analysis of the "coordinate-by-coordinate" coupling given in the proof of Theorem 5.5.

Let $t \geq kdn^2$.

(a) Show that the probability that the first coordinates of the two walks have not yet coupled by time t is less than $(1/4)^k$.

(b) By making an appropriate choice of k and considering all the coordinates, obtain an $O((d \log d)n^2)$ bound on t_{mix}.

Notes

The use of coupling in probability is usually traced back to Doeblin (1938). Couplings of Markov chains were first studied in Pitman (1974) and Griffeath (1974/75). See also Pitman (1976). See Luby, Randall, and Sinclair (1995) and Luby, Randall, and Sinclair (2001) for interesting examples of couplings.

For Glauber dynamics on colorings, it is shown in Chapter 14 that if the number of colors q satisfies $q > 2\Delta$, then the mixing time is of order $n \log n$.

Luby and Vigoda (1999) show that for a different Markov chain with the hard-core model as its stationary distribution, for λ small enough, the mixing time is of order $n \log n$. See also Luby and Vigoda (1995) and Vigoda (2001).

Further reading. For more on coupling and its applications in probability, see Lindvall (2002) and Thorisson (2000).

CHAPTER 6

Strong Stationary Times

6.1. Top-to-Random Shuffle

We begin this chapter with an example. Consider the following (slow) method of shuffling a deck of n cards: take the top card and insert it uniformly at random in the deck. This process will eventually mix up the deck—the successive arrangements of the deck are a random walk on the group $\mathcal{S}_n$ of $n!$ possible permutations of the cards, which by Proposition 2.12 has uniform stationary distribution.

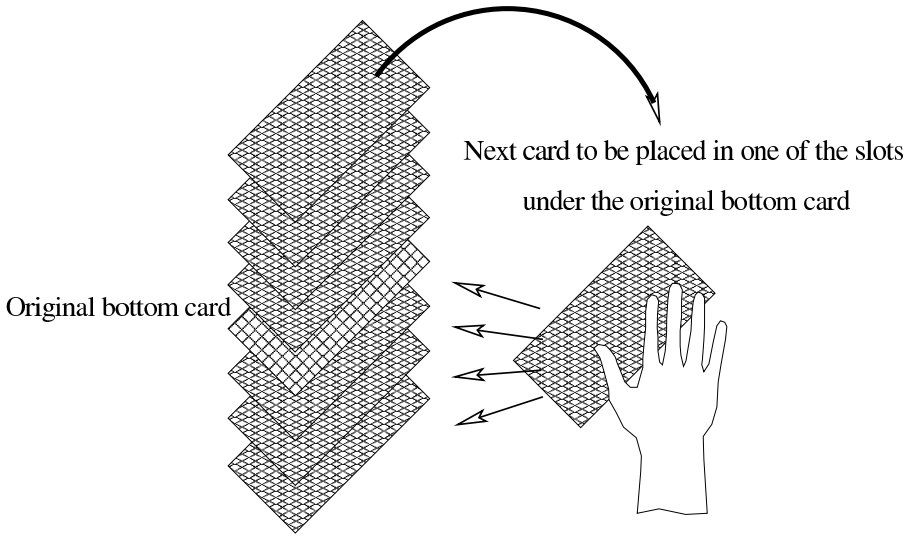


FIGURE 6.1. The top-to-random shuffle.

How long must we shuffle using this method until the arrangement of the deck is close to random?

Let τ_{top} be the time *one move after the first occasion when the original bottom card has moved to the top of the deck.* We show now that the arrangement of cards at time τ_{top} is distributed uniformly on the set $\mathcal{S}_n$ of all permutations of $\{1, \ldots, n\}$ and moreover this random element of $\mathcal{S}_n$ is independent of the time τ_{top}.

More generally, we prove the following:

PROPOSITION 6.1. *Let (X_t) be the random walk on $\mathcal{S}_n$ corresponding to the top-to-random shuffle on n cards. Given at time t that there are k cards under the original bottom card, each of the $k!$ possible orderings of these cards are equally likely. Therefore, if τ_{top} is one shuffle after the first time that the original bottom*

card moves to the top of the deck, then the distribution of $X_{\tau_{\text{top}}}$ is uniform over $\mathcal{S}_n$, and the time τ_{top} is independent of $X_{\tau_{\text{top}}}$.

PROOF. When $t = 0$, there are no cards under the original bottom card, and the claim is trivially valid. Now suppose that the claim holds at time t. There are two possibilities at time $t + 1$: either a card is placed under the original bottom card, or not. In the second case, the cards under the original bottom card remain in random order. In the first case, given that the card is placed under the original bottom card, each of the $k + 1$ possible locations for the card is equally likely, and so each of the $(k + 1)!$ orderings are equiprobable. ■

If we stop shuffling precisely one shuffle after the original bottom card rises to the top of the deck for the first time, then the order of the cards at this time is exactly uniform over all possible arrangements. That is, $X_{\tau_{\text{top}}}$ has *exactly* the stationary distribution of the chain. In this chapter, we show how we can use the distribution of the *random* time τ_{top} to bound t_{mix}, the *fixed* number of steps needed for the distribution of the chain to be *approximately* stationary.

6.2. Definitions

6.2.1. Stopping times. Suppose you give instructions to your stock broker to sell a particular security when its value next drops below 32 dollars per share. This directive can be implemented by a computer program: at each unit of time, the value of the security is checked; if the value at that time is at least 32, no action is taken, while if the value is less than 32, the asset is sold and the program quits.

You would like to tell your broker to sell a stock at the first time its value equals its maximum value over its lifetime. However, this is not a reasonable instruction, because to determine on Wednesday whether or not to sell, the broker needs to know that on Thursday the value will not rise and in fact for the entire infinite future that the value will never exceed its present value. To determine the correct decision on Wednesday, the broker must be able to see into the future!

The first instruction is an example of a *stopping time*, which we will now define, while the second rule is not.

Given a sequence $(X_t)_{t=0}^{\infty}$ of Ω-valued random variables, a $\{0, 1, 2, \ldots, \infty\}$-valued random variable τ is a ***stopping time*** for (X_t) if, for each $t \in \{0, 1, \ldots\}$, there is a set $B_t \subset \Omega^{t+1}$ such that

$$\{\tau = t\} = \{(X_0, X_1, \ldots, X_t) \in B_t\}.$$

In other words, a random time τ is a stopping time if and only if the indicator function $\mathbf{1}_{\{\tau=t\}}$ is a function of the vector $(X_0, X_1, \ldots, X_t)$.

EXAMPLE 6.2 (Hitting times). Fix $A \subseteq \Omega$. The vector $(X_0, X_1, \ldots, X_t)$ determines whether a site in A is visited for the first time at time t. That is, if

$$\tau_A = \min\{t \geq 0 : X_t \in A\}$$

is the first time that the sequence (X_t) is in A, then

$$\{\tau_A = t\} = \{X_0 \notin A, X_1 \notin A, \ldots, X_{t-1} \notin A, X_t \in A\}.$$

Therefore, τ_A is a stopping time. (We saw the special case where $A = \{x\}$ consists of a single state in Section 1.5.2.)

Consider the top-to-random shuffle, defined in Section 6.1. Let A be the set of arrangements having the original bottom card on top. Then $\tau_{\text{top}} = \tau_A + 1$. By Exercise 6.1, τ_{top} is a stopping time.

6.2.2. Randomized stopping times. The following example is instructive.

EXAMPLE 6.3 (Random walk on the hypercube). The lazy random walk $(\boldsymbol{X}_t)$ on the hypercube $\{0,1\}^n$ was introduced in Section 2.3, and we used coupling to bound the mixing time in Section 5.3.3. Recall that a move of this walk can be constructed using the following random mapping representation: an element (j, B) from $\{1, 2, \ldots, n\} \times \{0, 1\}$ is selected uniformly at random, and coordinate j of the current state is updated with the bit B.

In this construction, the chain is determined by the i.i.d. sequence (Z_t), where $Z_t = (j_t, B_t)$ is the coordinate and bit pair used to update at step t.

Define

$$\tau_{\text{refresh}} := \min\{t \geq 0 : \{j_1, \ldots, j_t\} = \{1, 2, \ldots, n\}\},$$

the first time when all the coordinates have been selected at least once for updating.

Because at time τ_{refresh} all of the coordinates have been replaced with independent fair bits, the distribution of the chain at this time is uniform on $\{0,1\}^n$. That is, $X_{\tau_{\text{refresh}}}$ is an exact sample from the stationary distribution π.

Note that τ_{refresh} is not a function of (X_t), but it is a function of (Z_t). In particular, while τ_{refresh} is not a stopping time for (X_t), it is a stopping time for (Z_t).

Recall that we showed in Section 1.2 that every transition matrix P has a random mapping representation: we can find an i.i.d. sequence $(Z_t)_{t=1}^{\infty}$ and a map f such that the sequence $(X_t)_{t=0}^{\infty}$ defined inductively by

$$X_0 = x, \qquad X_t = f(X_{t-1}, Z_t)$$

is a Markov chain with transition matrix P started from x. A random time τ is called a ***randomized stopping time*** for the Markov chain (X_t) if it is a stopping time for the sequence (Z_t).

EXAMPLE 6.4. We return to Example 6.3, the lazy random walk on the hypercube. As remarked there, the time τ_{refresh} is a stopping time for the sequence (Z_t), where Z_t is the coordinate and bit used to update at time t. Therefore, τ_{refresh} is a randomized stopping time.

6.3. Achieving Equilibrium

For the top-to-random shuffle, one shuffle after the original bottom card rises to the top, the deck is in completely random order. Likewise, for the lazy random walker on the hypercube, at the first time when all of the coordinates have been updated, the state of the chain is a random sample from $\{0,1\}^n$. These random times are examples of ***stationary times***, which we now define.

Let (X_t) be an irreducible Markov chain with stationary distribution π. A ***stationary time*** τ for (X_t) is a randomized stopping time, possibly depending on the starting position x, such that the distribution of X_τ is π:

$$\mathbf{P}_x\{X_\tau = y\} = \pi(y). \tag{6.1}$$

EXAMPLE 6.5. Let (X_t) be an irreducible Markov chain with state space Ω and stationary distribution π. Let ξ be a Ω-valued random variable with distribution π, and define

$$\tau = \min\{t \geq 0 : X_t = \xi\}.$$

The time τ is a randomized stopping time, and because $X_\tau = \xi$, it follows that τ is a stationary time.

Suppose the chain starts at x_0. If $\tau = 0$, then $X_\tau = x_0$; therefore, τ and X_τ are not independent.

EXAMPLE 6.6. Let (X_t) be the random walk on the n-cycle. Define τ by tossing a coin with probability of heads $1/n$. If "heads", let $\tau = 0$; if "tails", let τ be the first time every state has been visited at least once. Given "tails", the distribution of X_τ is uniform over all $n-1$ states different from the starting state. (See Exercise 6.9.) This shows that X_τ has the uniform distribution, whence τ is a stationary time.

However, $\tau = 0$ implies that X_τ is the starting state. Therefore, as in Example 6.5, τ and X_τ are not independent.

As mentioned at the end of Section 6.1, we want to use the time τ_{top} to bound t_{mix}. To carry out this program, we need a property of τ_{top} stronger than (6.1). We will need that τ_{top} is independent of $X_{\tau_{\text{top}}}$, a property not enjoyed by the stationary times in Example 6.5 and Example 6.6.

6.4. Strong Stationary Times and Bounding Distance

A ***strong stationary time*** for a Markov chain (X_t) with stationary distribution π is a randomized stopping time τ, possibly depending on the starting position x, such that

$$\mathbf{P}_x\{\tau = t,\ X_\tau = y\} = \mathbf{P}_x\{\tau = t\}\pi(y). \tag{6.2}$$

In words, X_τ has distribution π and is independent of τ.

EXAMPLE 6.7. For the top-to-random shuffle, the first time τ_{top} when the original bottom card gets placed into the deck by a shuffle is a strong stationary time. This is the content of Proposition 6.1.

EXAMPLE 6.8. We return to Example 6.3, the lazy random walk on the hypercube. The time τ_{refresh}, the first time each of the coordinates have been refreshed with an independent fair bit, is a strong stationary time.

We now return to the program suggested at the end of Section 6.1 and use strong stationary times to bound t_{mix}.

We first need the following technical lemma.

LEMMA 6.9. *Let (X_t) be an irreducible Markov chain with stationary distribution π. If τ is a strong stationary time for (X_t), then for all $t \geq 0$,*

$$\mathbf{P}_x\{\tau \leq t, X_t = y\} = \mathbf{P}\{\tau \leq t\}\pi(y). \tag{6.3}$$

PROOF. Let $Z_1, Z_2, \ldots$ be the i.i.d. sequence used in the random mapping representation of (X_t). For any $s \leq t$,

$$\mathbf{P}_x\{\tau = s, X_t = y\} = \sum_{z\in\Omega} \mathbf{P}_x\{X_t = y \mid \tau = s, X_s = z\}\mathbf{P}_x\{\tau = s, X_s = z\}. \tag{6.4}$$

Since τ is a stopping time for (Z_t), the event $\{\tau = s\}$ equals $\{(Z_1, \dots, Z_s) \in B\}$ for some set $B \subset \Omega^s$. Also, for integers $r, s \geq 0$, there exists a function $\tilde{f}_r : \Omega^{r+1} \to \Omega$ such that

$$X_{s+r} = \tilde{f}_r(X_s, Z_{s+1}, \dots, Z_{s+r}).$$

Since the random vectors $(Z_1, \dots, Z_s)$ and $(Z_{s+1}, \dots, Z_t)$ are independent,

$$\begin{aligned}&\mathbf{P}_x\{X_t = y \mid \tau = s,\, X_s = z\}\\ &\quad = \mathbf{P}_x\{\tilde{f}_{t-s}(z, Z_{s+1}, \dots, Z_t) = y \mid (X_1, \dots, X_s) \in B,\, X_s = z\} = P^{t-s}(z, y).\end{aligned}$$

Therefore, using the definition (6.2) along with the above equality, (6.4) can be rewritten as

$$\mathbf{P}_x\{\tau = s, X_t = y\} = \sum_{z \in \Omega} P^{t-s}(z, y)\pi(z)\mathbf{P}_x\{\tau = s\}. \tag{6.5}$$

Since π satisfies $\pi = \pi P^{t-s}$, the right-hand side of (6.5) equals $\pi(y)\mathbf{P}_x\{\tau = s\}$. Summing over $s \leq t$ establishes (6.3). ∎

The route from strong stationary times to bounding convergence time is the following proposition:

PROPOSITION 6.10. *If τ is a strong stationary time, then*

$$d(t) = \max_{x \in \Omega} \|P^t(x, \cdot) - \pi\|_{\mathrm{TV}} \leq \max_{x \in \Omega} \mathbf{P}_x\{\tau > t\}. \tag{6.6}$$

We break the proof into two lemmas. It will be convenient to introduce a parameter $s_x(t)$, called ***separation distance*** and defined by

$$s_x(t) := \max_{y \in \Omega} \left[1 - \frac{P^t(x, y)}{\pi(y)}\right]. \tag{6.7}$$

We also define

$$s(t) := \max_{x \in \Omega} s_x(t). \tag{6.8}$$

The relationship between $s_x(t)$ and strong stationary times is

LEMMA 6.11. *If τ is a strong stationary time, then*

$$s_x(t) \leq \mathbf{P}_x\{\tau > t\}. \tag{6.9}$$

PROOF. Fix $x \in \Omega$. Observe that for any $y \in \Omega$,

$$1 - \frac{P^t(x, y)}{\pi(y)} = 1 - \frac{\mathbf{P}_x\{X_t = y\}}{\pi(y)} \leq 1 - \frac{\mathbf{P}_x\{X_t = y,\, \tau \leq t\}}{\pi(y)}. \tag{6.10}$$

By Lemma 6.9, the right-hand side equals

$$1 - \frac{\pi(y)\mathbf{P}_x\{\tau \leq t\}}{\pi(y)} = \mathbf{P}_x\{\tau > t\}. \tag{6.11}$$

∎

REMARK 6.12. Given starting state x, a state y is a ***halting state*** for a stopping time τ if $X_t = y$ implies $\tau \leq t$. For example, when starting the lazy random walk on the hypercube at $(0, \dots, 0)$, the state $(1, \dots, 1)$ is a halting state for the stopping time τ_{refresh} defined in Example 6.3. Because the inequality in (6.10) is an equality if and only if y is a halting state for the starting state x, it follows that the inequality in (6.9) is an equality if and only if there exists a halting state for the starting state x.

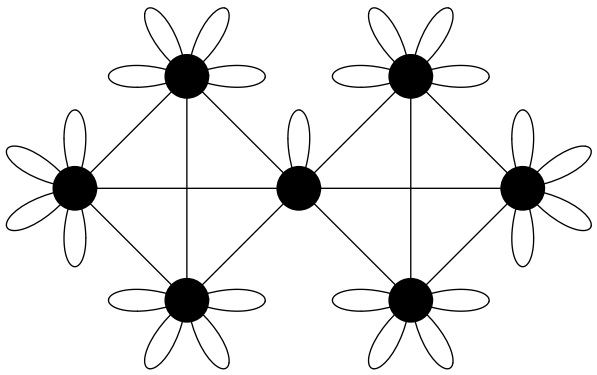


FIGURE 6.2. Two complete graphs (on 4 vertices), "glued" at a single vertex. Loops have been added so that every vertex has the same degree (count each loop as one edge).

The next lemma along with Lemma 6.11 proves Proposition 6.10.

LEMMA 6.13. *The separation distance $s_x(t)$ satisfies*

$$\left\|P^t(x,\cdot)-\pi\right\|_{\mathrm{TV}} \le s_x(t), \tag{6.12}$$

and therefore $d(t) \le s(t)$.

PROOF. We have

$$\begin{aligned}\|P^t(x,\cdot)-\pi\|_{\mathrm{TV}} &= \sum_{\substack{y\in\Omega\\ P^t(x,y)<\pi(y)}} \left[\pi(y)-P^t(x,y)\right] = \sum_{\substack{y\in\Omega\\ P^t(x,y)<\pi(y)}} \pi(y)\left[1-\frac{P^t(x,y)}{\pi(y)}\right]\\ &\le \max_y \left[1-\frac{P^t(x,y)}{\pi(y)}\right] = s_x(t).\end{aligned}$$

■

6.5. Examples

6.5.1. Two glued complete graphs. Consider the graph G obtained by taking two complete graphs on n vertices and "gluing" them together at a single vertex. We analyze here simple random walk on a slightly modified graph, G'.

Let $v^\star$ be the vertex where the two complete graphs meet. After gluing, $v^\star$ has degree $2n-2$, while every other vertex has degree $n-1$. To make the graph regular and to ensure non-zero holding probability at each vertex, in G' we add one loop at $v^\star$ and n loops at all other vertices. (See Figure 6.2 for an illustration when $n=4$.) The uniform distribution is stationary for simple random walk on G', since it is regular of degree $2n-1$.

It is clear that when at $v^\star$, the next state is equally likely to be any of the $2n-1$ vertices. For this reason, if τ is the time one step after $v^\star$ has been visited for the first time, then τ is a strong stationary time.

When the walk is *not* at $v^\star$, the probability of moving (in one step) to $v^\star$ is $1/(2n-1)$. This remains true at any subsequent move. That is, the first time $\tau_{v^\star}$

that the walk visits $v^\star$ is geometric with $\mathbf{E}(\tau_{v^\star}) = 2n - 1$. Therefore, $\mathbf{E}(\tau) = 2n$, and using Markov's inequality shows that

$$\mathbf{P}_x\{\tau \geq t\} \leq \frac{2n}{t}. \tag{6.13}$$

Taking $t = 8n$ in (6.13) and applying Proposition 6.10 shows that

$$t_{\text{mix}} \leq 8n.$$

A lower bound on t_{mix} of order n is obtained in Exercise 6.7.

6.5.2. Random walk on the hypercube. We return to Example 6.3, the lazy random walker on $\{0,1\}^n$. As noted in Example 6.8, the random variable τ_{refresh}, the time when each coordinate has been selected at least once for the first time, is a strong stationary time. The time τ_{refresh} and the coupling time τ_{couple} for the coordinate-by-coordinate coupling used in Section 5.3.3 are closely related: the coupon collector's time of Section 2.2 stochastically dominates τ_{couple} and has the same distribution as τ_{refresh}. It is therefore not surprising that we obtain here exactly the same upper bound for t_{mix} as was found using the coupling method. In particular, combining Proposition 2.4 and Lemma 6.11 shows that the separation distance satisfies, for each x,

$$s_x(n\log n + cn) \leq e^{-c}. \tag{6.14}$$

By Lemma 6.13,

$$t_{\text{mix}}(\varepsilon) \leq n\log n + \log(\varepsilon^{-1})n. \tag{6.15}$$

REMARK 6.14. The reason we explicitly give a bound on the separation distance here and appeal to Lemma 6.13, instead of applying directly Proposition 6.10, is that there is a matching lower bound on $s(t)$, which we give in Section 18.4. This contrasts with the lower bound on $d(t)$ we will find in Section 7.3.1, which implies $t_{\text{mix}}(1-\varepsilon) \geq (1/2)n\log n - c(\varepsilon)n$. In fact, the estimate on $t_{\text{mix}}(\varepsilon)$ given in (6.15) is off by a factor of two, as we will see in Section 18.2.2.

6.5.3. Top-to-random shuffle. We revisit the top-to-random shuffle introduced in Section 6.1. As noted in Example 6.7, the time τ_{top} is a strong stationary time.

Consider the motion of the original bottom card. When there are k cards beneath it, the chance that it rises one card remains k/n until a shuffle puts the top card underneath it. Thus, the distribution of τ_{top} is the same as the coupon collector's time. As above for the lazy hypercube walker, combining Proposition 6.10 and Proposition 2.4 yields

$$d(n\log n + \alpha n) \leq e^{-\alpha} \quad \text{for all } n. \tag{6.16}$$

Consequently,

$$t_{\text{mix}}(\varepsilon) \leq n\log n + \log(\varepsilon^{-1})n. \tag{6.17}$$

6.5.4. The move-to-front chain. A certain professor owns many books, arranged on his shelves. When he finishes with a book drawn from his collection, he does not waste time re-shelving it in its proper location. Instead, he puts it at the very beginning of his collection, in front of all the shelved books.

If his choice of book is random, this is an example of the ***move-to-front*** chain. It is a very natural chain which arises in many applied contexts. Any setting where

Drawing by Yelena Shvets

FIGURE 6.3. The move-to-front rule in action.

items are stored in a stack, removed at random locations, and placed on the top of the stack can be modeled by the move-to-front chain.

Let P be the transition matrix (on permutations of $\{1, 2, \ldots, n\}$) corresponding to this method of rearranging elements.

The time reversal $\widehat{P}$ of the move-to-front chain is the top-to-random shuffle, as intuition would expect. It is clear from the definition that for any permissible transition $\sigma_1 \to \sigma_2$ for move-to-front, the transition $\sigma_2 \to \sigma_1$ is permissible for top-to-random, and both have probability n^{-1}.

By Lemma 4.13, the mixing time for move-to-front will be identical to that of the top-to-random shuffle. Consequently, the mixing time for move-to-front is not more than $n \log n - \log(\varepsilon) n$.

6.5.5. Lazy random walk on cycle. Here is a recursive description of a strong stationary time τ_k for lazy random walk (X_t) on a cycle $\mathbb{Z}_n$ with $n = 2^k$ points.

For $k = 1$, waiting one step will do: $\tau_1 = 1$ with mean $m_1 = 1$. Suppose we have constructed τ_k already and are now given a cycle with 2^{k+1} points. Set $T_0 = 0$ and define $T_1 = t_1$ as the time it takes the lazy walk to make two ± 1 steps.

Exercises

EXERCISE 6.1. Show that if τ and τ' are stopping times for the sequence (X_t), then $\tau + \tau'$ is a stopping time for (X_t). In particular, if r is a non-random and non-negative integer and τ is a stopping time, then $\tau + r$ is a stopping time.

EXERCISE 6.2. Consider the top-to-random shuffle. Show that the time until the card initially one card from the bottom rises to the top, plus one more move, is a strong stationary time, and find its expectation.

EXERCISE 6.3. Show that for the Markov chain on two complete graphs in Section 6.5.1, the stationary distribution is uniform on all $2n - 1$ vertices.

EXERCISE 6.4. Let $s(t)$ be defined as in (6.8).

(a) Show that there is a stochastic matrix Q so that $P^t(x, \cdot) = [1 - s(t)]\,\pi + s(t)Q^t(x, \cdot)$ and $\pi = \pi Q$.

(b) Using the representation in (a), show that

$$P^{t+u}(x, y) = [1 - s(t)s(u)]\,\pi(y) + s(t)s(u) \sum_{z \in \Omega} Q^t(x, z)Q^u(z, y). \tag{6.19}$$

(c) Using (6.19), establish that s is submultiplicative: $s(t + u) \le s(t)s(u)$.

EXERCISE 6.5. Show that if $\max_{x \in \Omega} \mathbf{P}_x\{\tau > t_0\} \le \varepsilon$, then $d(t) \le \varepsilon^{t/t_0}$.

EXERCISE 6.6 (Wald's Identity). Let (Y_t) be a sequence of independent and identically distributed random variables such that $\mathbf{E}(|Y_t|) < \infty$.

(a) Show that if τ is a random time so that the event $\{\tau \ge t\}$ is independent of Y_t and $\mathbf{E}(\tau) < \infty$, then

$$\mathbf{E}\left(\sum_{t=1}^{\tau} Y_t\right) = \mathbf{E}(\tau)\mathbf{E}(Y_1). \tag{6.20}$$

Hint: Write $\sum_{t=1}^{\tau} Y_t = \sum_{t=1}^{\infty} Y_t \mathbf{1}_{\{\tau \ge t\}}$. First consider the case where $Y_t \ge 0$.

(b) Let τ be a stopping time for the sequence (Y_t). Show that $\{\tau \ge t\}$ is independent of Y_t, so (6.20) holds provided that $\mathbf{E}(\tau) < \infty$.

EXERCISE 6.7. Consider the Markov chain of Section 6.5.1 defined on two glued complete graphs. By considering the set $A \subset \Omega$ of all vertices in one of the two complete graphs, show that $t_{\text{mix}} \ge (n/2)\,[1 + o(1)]$.

EXERCISE 6.8. Let τ_k be the stopping time constructed in Section 6.5.5, and let $m_k = \mathbf{E}(\tau_k)$. Show that $m_{k+1} = 4m_k + 1$, so that $m_k = \sum_{i=0}^{k-1} 4^i = (4^k - 1)/3$.

EXERCISE 6.9. For a graph G, let W be the (random) vertex occupied at the first time the random walk has visited every vertex. That is, W is the last new vertex to be visited by the random walk. Prove the following remarkable fact: for random walk on an n-cycle, W is uniformly distributed over all vertices different from the starting vertex.

REMARK 6.16. Let W be the random vertex defined in Exercise 6.9. Lovász and Winkler (1993) demonstrate that cycles and complete graphs are the only graphs for which W is this close to uniformly distributed. More precisely, these families are the only ones for which W is equally likely to be any vertex other than the starting state.

CHAPTER 7

Lower Bounds on Mixing Times

To this point, we have directed our attention to finding upper bounds on t_{mix}. Rigorous upper bounds lend confidence that simulation studies or randomized algorithms perform as advertised. It is natural to ask if a given upper bound is the best possible, and so in this chapter we turn to methods of obtaining lower bounds on t_{mix}.

7.1. Counting and Diameter Bounds

7.1.1. Counting bound. If the possible locations of a chain after t steps do not form a significant fraction of the state space, then the distribution of the chain at time t cannot be close to uniform. This idea can be used to obtain lower bounds on the mixing time.

Let (X_t) be a Markov chain with irreducible and aperiodic transition matrix P on the state space Ω, and suppose that the stationary distribution π is uniform over Ω. Define $d_{\text{out}}(x) := |\{y : P(x,y) > 0\}|$ to be the number of states accessible in one step from x, and let

$$\Delta := \max_{x\in\Omega} d_{\text{out}}(x). \tag{7.1}$$

Denote by Ω_t^x the set of states accessible from x in t steps, and observe that $|\Omega_t^x| \leq \Delta^t$. If $\Delta^t < (1-\varepsilon)|\Omega|$, then from the definition of total variation distance we have that

$$\left\|P^t(x,\cdot) - \pi\right\|_{\text{TV}} \geq P_t(x,\Omega_t^x) - \pi(\Omega_t^x) \geq 1 - \frac{\Delta^t}{|\Omega|} > \varepsilon.$$

This implies that

$$t_{\text{mix}}(\varepsilon) \geq \frac{\log(|\Omega|(1-\varepsilon))}{\log\Delta}. \tag{7.2}$$

EXAMPLE 7.1 (Random walk on a d-regular graph). For random walk on a d-regular graph, the stationary distribution is uniform, so the inequality (7.2) can be applied. In this case, it yields the lower bound $t_{\text{mix}}(\varepsilon) \geq \log(|\Omega|(1-\varepsilon))/\log d$.

We use the bound (7.2) to bound below the mixing time for the riffle shuffle in Proposition 8.14.

7.1.2. Diameter bound. Given a transition matrix P on Ω, construct a graph with vertex set Ω and which includes the edge $\{x,y\}$ for all x and y with $P(x,y) + P(y,x) > 0$. Define the ***diameter*** of a Markov chain to be the diameter of this graph, that is, the maximal graph distance between distinct vertices.

Let P be an irreducible and aperiodic transition matrix on Ω with diameter L, and suppose that x_0 and y_0 are states at maximal graph distance L. Then

$P^{\lfloor (L-1)/2 \rfloor}(x_0, \cdot)$ and $P^{\lfloor (L-1)/2 \rfloor}(y_0, \cdot)$ are positive on disjoint vertex sets. Consequently, $\bar{d}(\lfloor (L-1)/2 \rfloor) = 1$ and for any $\varepsilon < 1/2$,

$$t_{\text{mix}}(\varepsilon) \geq \frac{L}{2}. \tag{7.3}$$

7.2. Bottleneck Ratio

Bottlenecks in the state space Ω of a Markov chain are geometric features that control mixing time. A bottleneck makes portions of Ω difficult to reach from some starting locations, limiting the speed of convergence. Figure 7.1 is a sketch of a graph with an obvious bottleneck.

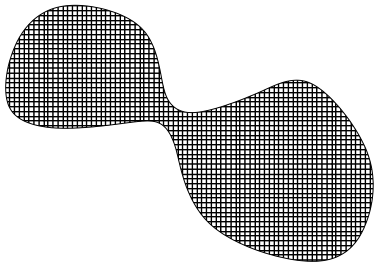


FIGURE 7.1. A graph with a bottleneck.

As usual, P is the irreducible and aperiodic transition matrix for a Markov chain on Ω with stationary distribution π.

The ***edge measure*** Q is defined by

$$Q(x,y) := \pi(x)P(x,y), \quad Q(A,B) = \sum_{x \in A, y \in B} Q(x,y). \tag{7.4}$$

Here $Q(A,B)$ is the probability of moving from A to B in one step when starting from the stationary distribution.

The ***bottleneck ratio*** of the set S is defined to be

$$\Phi(S) := \frac{Q(S, S^c)}{\pi(S)}, \tag{7.5}$$

while the bottleneck ratio of the whole chain is

$$\Phi_\star := \min_{S\,:\,\pi(S) \leq \frac{1}{2}} \Phi(S). \tag{7.6}$$

For simple random walk on a graph with vertices Ω and edge set E,

$$Q(x,y) = \begin{cases} \frac{\deg(x)}{2|E|}\frac{1}{\deg(x)} = \frac{1}{2|E|} & \text{if } \{x,y\} \text{ is an edge,} \\ 0 & \text{otherwise.} \end{cases}$$

In this case, $2|E|Q(S, S^c)$ is the size of the ***boundary*** ∂S of S, the collection of edges having one vertex in S and one vertex in S^c. The bottleneck ratio, in this case, becomes

$$\Phi(S) = \frac{|\partial S|}{\sum_{x \in S} \deg(x)}. \tag{7.7}$$

REMARK 7.2. If the walk is lazy, then $Q(x,y) = (4|E|)^{-1}\mathbf{1}_{\{\{x,y\} \in E\}}$, and the bottleneck ratio equals $\Phi(S) = 2|\partial S|/(\sum_{x \in S} \deg(x))$.

If the graph is regular with degree d, then $\Phi(S) = d^{-1}|\partial S|/|S|$, which is proportional to the ratio of the size of the boundary of S to the volume of S.

The relationship of $\Phi_\star$ to t_{mix} is the following theorem:

THEOREM 7.3. *If $\Phi_\star$ is the bottleneck ratio defined in (7.6), then*

$$t_{\text{mix}} = t_{\text{mix}}(1/4) \geq \frac{1}{4\Phi_\star}. \tag{7.8}$$

PROOF. Denote by π_S the restriction of π to S, so that $\pi_S(A) = \pi(A \cap S)$, and define μ_S to be π conditioned on S:

$$\mu_S(A) = \frac{\pi_S(A)}{\pi(S)}.$$

From Remark 4.3,

$$\pi(S)\,\|\mu_S P - \mu_S\|_{\text{TV}} = \pi(S) \sum_{\substack{y\in\Omega,\\ \mu_S P(y)\geq \mu_S(y)}} [\mu_S P(y) - \mu_S(y)]\,. \tag{7.9}$$

Because $\pi_S P(y) = \pi(S)\mu_S P(y)$ and $\pi_S(y) = \pi(S)\mu_S(y)$, the inequality $\mu_S P(y) \geq \mu_S(y)$ holds if and only if $\pi_S P(y) \geq \pi_S(y)$. Thus

$$\pi(S)\,\|\mu_S P - \mu_S\|_{\text{TV}} = \sum_{\substack{y\in\Omega,\\ \pi_S P(y)\geq \pi_S(y)}} [\pi_S P(y) - \pi_S(y)]\,. \tag{7.10}$$

Because $\pi_S(x) > 0$ only for $x \in S$ and $\pi_S(x) = \pi(x)$ for $x \in S$,

$$\pi_S P(y) = \sum_{x\in\Omega} \pi_S(x)P(x,y) = \sum_{x\in S} \pi(x)P(x,y) \leq \sum_{x\in\Omega} \pi(x)P(x,y) = \pi(y). \tag{7.11}$$

Again using that $\pi(y) = \pi_S(y)$ for $y \in S$, from (7.11) follows the inequality

$$\pi_S P(y) \leq \pi_S(y) \quad \text{for } y \in S. \tag{7.12}$$

On the other hand, because π_S vanishes on S^c,

$$\pi_S P(y) \geq 0 = \pi_S(y) \quad \text{for } y \in S^c. \tag{7.13}$$

Combining (7.12) and (7.13) shows that the sum on the right in (7.10) can be taken over S^c:

$$\pi(S)\,\|\mu_S P - \mu_S\|_{\text{TV}} = \sum_{y\in S^c} [\pi_S P(y) - \pi_S(y)]\,. \tag{7.14}$$

Again because $\pi_S(y) = 0$ for $y \in S^c$,

$$\pi(S)\,\|\mu_S P - \mu_S\|_{\text{TV}} = \sum_{y\in S^c}\sum_{x\in S} \pi(x)P(x,y) = Q(S, S^c).$$

Dividing by $\pi(S)$,

$$\|\mu_S P - \mu_S\|_{\text{TV}} = \Phi(S).$$

By Exercise 4.3, for any $u \geq 0$,

$$\left\|\mu_S P^{u+1} - \mu_S P^u\right\|_{\text{TV}} \leq \|\mu_S P - \mu_S\|_{\text{TV}} = \Phi(S).$$

Using the triangle inequality on $\mu_S P^t - \mu_S = \sum_{u=0}^{t-1}(\mu_S P^{u+1} - \mu_S P^u)$ shows that

$$\left\|\mu_S P^t - \mu_S\right\|_{\text{TV}} \leq t\Phi(S). \tag{7.15}$$

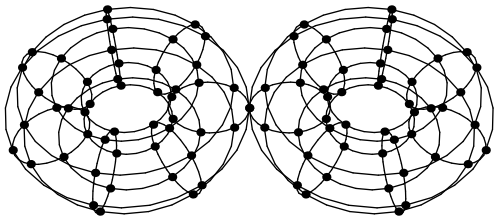


FIGURE 7.2. Two "glued" two-dimensional tori.

Assume that $\pi(S) \le \frac{1}{2}$. In this case, because $\mu_S(S^c) = 0$,

$$\|\mu_S - \pi\|_{TV} \ge \pi(S^c) - \mu_S(S^c) = 1 - \pi(S) \ge \frac{1}{2}.$$

Using the triangle inequality again shows that

$$\frac{1}{2} \le \|\mu_S - \pi\|_{TV} \le \left\|\mu_S - \mu_S P^t\right\|_{TV} + \left\|\mu_s P^t - \pi\right\|_{TV}. \tag{7.16}$$

Taking $t = t_{\text{mix}} = t_{\text{mix}}(1/4)$ in (7.16), by the definition of t_{mix} and the inequality in (7.15),

$$\frac{1}{2} \le t_{\text{mix}}\Phi(S) + \frac{1}{4}.$$

Rearranging and minimizing over S establishes (7.8). ■

EXAMPLE 7.4 (Two glued tori). Consider the graph consisting of two d-dimensional tori "glued" together at a single vertex $v^\star$; see Figure 7.2 for an example of dimension two. Denote by V_1 and V_2 the sets of vertices in the right and left tori, respectively. Note that $V_1 \cap V_2 = v^\star$.

The set ∂V_1 consists of all edges $\{v^\star, v\}$, where $v \in V_2$. The size of ∂V_1 is $2d$. Also, $\sum_{x \in V_1} \deg(x) = 2dn^2 + 2d$. Consequently, the lazy random walk on this graph has

$$\Phi_\star \le \Phi(V_1) = \frac{2(2d)}{2d(n^2+1)} \le 2n^{-2}.$$

(See Remark 7.2.) Theorem 7.3 implies that $t_{\text{mix}} \ge n^2/8$. We return to this example in Section 10.6, where it is proved that t_{mix} is of order $n^2 \log n$. Thus the lower bound here does not give the correct order.

EXAMPLE 7.5 (Coloring the star). Let Ω be the set of all proper q-colorings of a graph G, and let π be the uniform distribution on Ω. Recall from Example 3.5 that Glauber dynamics for π is the Markov chain which makes transitions as follows: at each unit of time, a vertex is chosen from V uniformly at random, and the color at this vertex is chosen uniformly at random from all ***feasible colors***. The feasible colors at vertex v are all colors *not* present among the neighbors of v.

We will prove (Theorem 14.8) that if $q > 2\Delta$, where Δ is the maximum degree of the graph, then the Glauber dynamics has mixing time of the order $|V| \log |V|$.

We show, by example, that quite different behavior may occur if the maximal degree is not bounded.

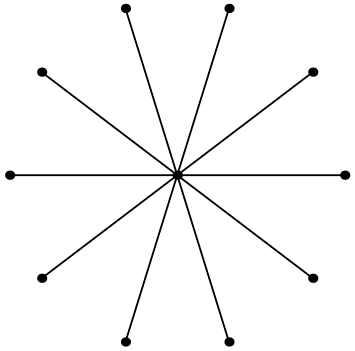


FIGURE 7.3. The star graph with 11 vertices.

The graph we study here is the ***star*** with n vertices, shown in Figure 7.3. This graph is a tree of depth 1 with $n-1$ leaves.

Let $v_\star$ denote the root vertex and let $S \subseteq \Omega$ be the set of proper colorings such that $v_\star$ has color 1:

$$S := \{x \in \Omega : x(v_\star) = 1\}.$$

For $(x, y) \in S \times S^c$, the edge measure $Q(x, y)$ is non-zero if and only if

- $x(v_\star) = 1$ and $y(v_\star) \neq 1$,
- $x(v) = y(v)$ for all leaves v, and
- $x(v) \notin \{1, y(v_\star)\}$ for all leaves v.

The number of such (x, y) pairs is therefore equal to $(q-1)(q-2)^{n-1}$, since there are $(q-1)$ possibilities for the color $y(v_\star)$ and $(q-2)$ possibilities for the color (identical in both x and y) of each of the $n-1$ leaves. Also, for such pairs, $Q(x, y) \leq (|\Omega|n)^{-1}$. It follows that

$$\sum_{x \in S, y \in S^c} Q(x, y) \leq \frac{1}{|\Omega|n}(q-1)(q-2)^{n-1}. \tag{7.17}$$

Since $x \in S$ if and only if $x(v_\star) = 1$ and $x(v) \neq 1$ for all $v \neq v_\star$, we have that $|S| = (q-1)^{n-1}$. Together with (7.17), this implies

$$\frac{Q(S, S^c)}{\pi(S)} = \frac{(q-1)(q-2)^{n-1}}{n(q-1)^{n-1}} = \frac{(q-1)^2}{n(q-2)}\left(1 - \frac{1}{q-1}\right)^n \leq \frac{(q-1)^2}{n(q-2)} e^{-n/(q-1)}.$$

Consequently, the mixing time is at least of exponential order:

$$t_{\text{mix}} \geq \frac{n(q-2)}{4(q-1)^2} e^{n/(q-1)}.$$

REMARK 7.6. In fact, this argument shows that if $n/(q \log q) \to \infty$, then t_{mix} is super-polynomial in n.

EXAMPLE 7.7 (Binary tree). Consider the lazy random walk on the rooted binary tree of depth k. (See Section 5.3.4 for the definition.) Let n be the number of vertices, so $n = 2^{k+1} - 1$. The number of edges is $n - 1$. In Section 5.3.4 we showed that $t_{\text{mix}} \leq 4n$. We now show that $t_{\text{mix}} \geq (n-2)/4$.

Let v_0 denote the root. Label the vertices adjacent to v_0 as v_r and v_ℓ. Call w a ***descendant*** of v if the shortest path from w to v_0 passes through v. Let S consist of the right-hand side of the tree, that is, v_r and all of its descendants.

We write $|v|$ for the length of the shortest path from v to v_0. By Example 1.12, the stationary distribution is

$$\pi(v) = \begin{cases} \frac{2}{2n-2} & \text{for } v = v_0, \\ \frac{3}{2n-2} & \text{for } 0 < |v| < k, \\ \frac{1}{2n-2} & \text{for } |v| = k. \end{cases}$$

Summing $\pi(v)$ over $v \in S$ shows that $\pi(S) = (n-2)/(2n-2)$. Since there is only one edge from S to S^c,

$$Q(S, S^c) = \pi(v_r)P(v_r, v_0) = \left(\frac{3}{2n-2}\right)\frac{1}{3} = \frac{1}{2n-2},$$

and so $\Phi(S) = 1/(n-2)$. Applying Theorem 7.3 establishes the lower bound

$$t_{\text{mix}} \geq \frac{n-2}{4} = \frac{2^{k+1}-3}{4},$$

which is exponentially large as a function of the depth k.

7.3. Distinguishing Statistics

One way to produce a lower bound on the mixing time t_{mix} is to find a statistic f (a real-valued function) on Ω such that the distance between the distribution of $f(X_t)$ and the distribution of f under the stationary distribution π can be bounded from below.

Let μ and ν be two probability distributions on Ω, and let f be a real-valued function defined on Ω. We write E_μ to indicate expectations of random variables (on sample space Ω) with respect to the probability distribution μ:

$$E_\mu(f) := \sum_{x \in \Omega} f(x)\mu(x).$$

(Note the distinction between E_μ with $\mathbf{E}_\mu$, the expectation operator corresponding to the Markov chain (X_t) started with initial distribution μ.) Likewise $\text{Var}_\mu(f)$ indicates variance computed with respect to the probability distribution μ.

PROPOSITION 7.8. *Let μ and ν be two probability distributions on Ω, and let f be a real-valued function on Ω. If*

$$|E_\mu(f) - E_\nu(f)| \geq r\sigma, \tag{7.18}$$

where $\sigma^2 = [\text{Var}_\mu(f) + \text{Var}_\nu(f)]/2$, then

$$\|\mu - \nu\|_{\text{TV}} \geq 1 - \frac{4}{4 + r^2}. \tag{7.19}$$

Before proving this, we provide a useful lemma. When μ is a probability distribution on Ω and $f : \Omega \to \Lambda$, write μf^{-1} for the probability distribution defined by

$$(\mu f^{-1})(A) := \mu(f^{-1}(A))$$

for $A \subseteq \Lambda$. When X is an Ω-valued random variable with distribution μ, then $f(X)$ has distribution μf^{-1} on Λ.

LEMMA 7.9. *Let μ and ν be probability distributions on Ω, and let $f : \Omega \to \Lambda$ be a function on Ω, where Λ is a finite set. Then*

$$\|\mu - \nu\|_{\text{TV}} \geq \left\|\mu f^{-1} - \nu f^{-1}\right\|_{\text{TV}}.$$

The inequality (7.26) follows because

$$\frac{1}{2}n\log n - \alpha n \le t_n = \left[\frac{1}{2}n\log n - \left(\alpha - \frac{1}{2}\right)n\right]\left[1 - \frac{1}{n+1}\right],$$

and the right-hand side of (7.27) evaluated at $t = t_n$ is equal to $1 - 8e^{-2\alpha+1}$. ■

7.4. Examples

7.4.1. Random walk on the cycle. We return to the lazy random walk on the cycle (see Example 1.8 and Example 2.10). The upper bound $t_{\text{mix}} \le n^2$ was found in Section 5.3.2.

We complement this by giving a lower bound of the same order. We can couple (X_t) to (S_t), a lazy simple random walk on all of $\mathbb{Z}$, so that $X_t = S_t$ until τ, the first time that $|X_t|$ hits $n/2$. Then

$$\mathbf{P}\left\{\sup_{t\le\alpha n^2}|X_t| > n/4\right\} = \mathbf{P}\left\{\sup_{t\le\alpha n^2}|S_t| > n/4\right\} \le \mathbf{P}\left\{|S_{\alpha n^2}| > n/4\right\} \le c_1\alpha,$$

by Chebyshev's inequality. For $\alpha < \alpha_0$, where α_0 is small enough, the right-hand side is less than $1/8$. If $A_n = \{k \in \mathbb{Z}_n : |k| \ge n/4\}$, then $\pi(A_n) \ge 1/2$, and

$$d(\alpha_0 n^2) \ge 1/2 - 1/8 > 1/4,$$

so $t_{\text{mix}} \ge \alpha_0 n^2$.

7.4.2. Top-to-random shuffle. The top-to-random shuffle was introduced in Section 6.1 and upper bounds on $d(t)$ and t_{mix} were obtained in Section 6.5.3. Here we obtain matching lower bounds.

The bound below, from Aldous and Diaconis (1986), uses only the definition of total variation distance.

PROPOSITION 7.14. *Let (X_t) be the top-to-random chain on n cards. For any $\varepsilon > 0$, there exists a constant α_0 such that $\alpha > \alpha_0$ implies that for all sufficiently large n,*

$$d_n(n\log n - \alpha n) \ge 1 - \varepsilon. \tag{7.28}$$

In particular, there is a constant α_1 such that for all sufficiently large n,

$$t_{\text{mix}} \ge n\log n - \alpha_1 n. \tag{7.29}$$

PROOF. The bound is based on the events

$$A_j = \{\text{the original bottom } j \text{ cards are in their original relative order}\}. \tag{7.30}$$

Let id be the identity permutation; we will bound $\|P^t(\text{id},\cdot) - \pi\|_{TV}$ from below.

Let τ_j be the time required for the card initially j-th from the bottom to reach the top. Then

$$\tau_j = \sum_{i=j}^{n-1}\tau_{j,i},$$

where $\tau_{j,i}$ is the time it takes the card initially j-th from the bottom to ascend from position i (from the bottom) to position $i+1$. The variables $\{\tau_{j,i}\}_{i=j}^{n-1}$ are

independent and $\tau_{j,i}$ has a geometric distribution with parameter $p = i/n$, whence $\mathbf{E}(\tau_{j,i}) = n/i$ and $\mathrm{Var}(\tau_{j,i}) < n^2/i^2$. We obtain the bounds

$$\mathbf{E}(\tau_j) = \sum_{i=j}^{n-1} \frac{n}{i} \geq n(\log n - \log j - 1) \tag{7.31}$$

and

$$\mathrm{Var}(\tau_j) \leq n^2 \sum_{i=j}^{\infty} \frac{1}{i(i-1)} \leq \frac{n^2}{j-1}. \tag{7.32}$$

Using the bounds (7.31) and (7.32), together with Chebyshev's inequality, yields

$$\begin{aligned} \mathbf{P}\{\tau_j < n\log n - \alpha n\} &\leq \mathbf{P}\{\tau_j - \mathbf{E}(\tau_j) < -n(\alpha - \log j - 1)\} \\ &\leq \frac{1}{(j-1)}, \end{aligned}$$

provided that $\alpha \geq \log j + 2$. Define $t_n(\alpha) = n\log n - \alpha n$. If $\tau_j \geq t_n(\alpha)$, then the original j bottom cards remain in their original relative order at time $t_n(\alpha)$, so

$$P^{t_n(\alpha)}(\mathrm{id}, A_j) \geq \mathbf{P}\{\tau_j \geq t_n(\alpha)\} \geq 1 - \frac{1}{(j-1)},$$

for $\alpha \geq \log j + 2$. On the other hand, for the uniform stationary distribution

$$\pi(A_j) = 1/(j!) \leq (j-1)^{-1},$$

whence, for $\alpha \geq \log j + 2$,

$$d_n(\,t_n(\alpha)\,) \geq \left\| P^{t_n(\alpha)}(\mathrm{id}, \cdot) - \pi \right\|_{\mathrm{TV}} \geq P^{t_n(\alpha)}(\mathrm{id}, A_j) - \pi(A_j) > 1 - \frac{2}{j-1}. \tag{7.33}$$

Taking $j = e^{\alpha-2}$, provided $n \geq e^{\alpha-2}$, we have

$$d_n(\,t_n(\alpha)\,) > g(\alpha) := 1 - \frac{2}{e^{\alpha-2} - 1}.$$

Therefore,

$$\liminf_{n\to\infty} d_n(\,t_n(\alpha)\,) \geq g(\alpha),$$

where $g(\alpha) \to 1$ as $\alpha \to \infty$. ■

7.4.3. East model. Let

$$\Omega := \{x \in \{0,1\}^{n+1} : x(n+1) = 1\}.$$

The ***East model*** is the Markov chain on Ω which moves from x by selecting a coordinate k from $\{1, 2, \ldots, n\}$ at random and flipping the value $x(k)$ at k if and only if $x(k+1) = 1$. The reader should check that the uniform measure on Ω is stationary for these dynamics.

THEOREM 7.15. *For the East model,* $t_{\mathrm{mix}} \geq n^2 - 2n^{3/2}$.

PROOF. If $A = \{x : x(1) = 1\}$, then $\pi(A) = 1/2$.

On the other hand, we now show that it takes order n^2 steps until $X_t(1) = 1$ with probability near $1/2$ when starting from $x_0 = (0, 0, \ldots, 0, 1)$. Consider the motion of the left-most 1: it moves to the left by one if and only if the site immediately to its left is chosen. Thus, the waiting time for the left-most 1 to move from k to $k-1$ is bounded below by a geometric random variable G_k with mean

CHAPTER 11

Cover Times

11.1. Cover Times

Let (X_t) be a finite Markov chain with state space Ω. The ***cover time*** τ_{cov} of (X_t) is the first time at which all the states have been visited. More formally, τ_{cov} is the minimal value such that, for every state $y \in \Omega$, there exists $t \le \tau_{\text{cov}}$ with $X_t = y$.

We also define a deterministic version of the cover time by taking the expected value from the worst-case initial state:

$$t_{\text{cov}} = \max_{x \in \Omega} \mathbf{E}_x \tau_{\text{cov}}. \tag{11.1}$$

The cover time of a Markov chain is a natural concept. It can be large enough for relatively small chains to arouse mathematical curiosity. Of course, there are also "practical" interpretations of the cover time. For instance, we might view the progress of a web crawler as a random walk on the graph of World Wide Web pages: at each step, the crawler chooses a linked page at random and goes there. The time taken to scan the entire collection of pages is the cover time of the underlying graph.

EXAMPLE 11.1 (Cover time of cycle). Lovász (1993) gives an elegant computation of the expected cover time t_{cov} of simple random walk on the n-cycle. This walk is simply the remainder modulo n of a simple random walk on $\mathbb{Z}$. The walk on the remainders has covered all n states exactly when the walk on $\mathbb{Z}$ has first visited n distinct states.

Let c_n be the expected value of the time when a simple random walk on $\mathbb{Z}$ has first visited n distinct states, and consider a walk which has just reached its $(n-1)$-st new state. The visited states form a subsegment of the number line and the walk must be at one end of that segment. Reaching the n-th new state is now a gambler's ruin situation: the walker's position corresponds to a fortune of 1 (or $n-1$), and we are waiting for her to reach either 0 or n. Either way, the expected time is $(1)(n-1) = n-1$, as shown in Exercise 2.1. It follows that

$$c_n = c_{n-1} + (n-1) \quad \text{for} \quad n \ge 1.$$

Since $c_1 = 0$ (the first state visited is $X_0 = 0$), we have $c_n = n(n-1)/2$.

11.2. The Matthews Method

It is not surprising that there is an essentially monotone relationship between hitting times and cover times: the longer it takes to travel between states, the longer it should take to visit all of them. In one direction, it is easy to write down a bound. Fix an irreducible chain with state space Ω. Recall the definition (10.5) of t_{hit}, and let $x, y \in \Omega$ be states for which $t_{\text{hit}} = \mathbf{E}_x \tau_y$. Since any walk started at

x must have visited y by the time all states are covered, we have

$$t_{\text{hit}} = \mathbf{E}_x \tau_y \le \mathbf{E}_x \tau_{\text{cov}} \le t_{\text{cov}}. \tag{11.2}$$

It is more interesting to give an upper bound on cover times in terms of hitting times. A walk covering all the states can visit them in many different orders, and this indeterminacy can be exploited. Randomizing the order in which we check whether states have been visited (which, following Aldous and Fill (1999), we will call the Matthews method—see Matthews (1988a) for the original version) allows us to prove both upper and lower bounds. Despite the simplicity of the arguments, these bounds are often remarkably good.

THEOREM 11.2 (Matthews (1988a)). *Let (X_t) be an irreducible finite Markov chain on n states. Then*

$$t_{\text{cov}} \le t_{\text{hit}} \left(1 + \frac{1}{2} + \cdots + \frac{1}{n}\right).$$

PROOF. Without loss of generality, we may assume that our state space is $\{1, \ldots, n\}$. Choose an arbitrary initial state $x \in \Omega$ and let $\sigma \in S_n$ be a uniform random permutation, chosen independently of the chain. We will look for states in order σ. Let T_k be the first time that the states $\sigma(1), \ldots, \sigma(k)$ have all been visited, and let $L_k = X_{T_k}$ be the last state among $\sigma(1), \ldots, \sigma(k)$ to be visited.

Of course, when $\sigma(1) = x$, we have $T_1 = 0$. We will not usually be so lucky. For any $s \in \Omega$, we have

$$\mathbf{E}_x(T_1 \mid \sigma(1) = s) = \mathbf{E}_x(\tau_s) \le t_{\text{hit}}.$$

Since the events $\{\sigma(1) = s\}$ are disjoint for distinct $s \in \Omega$, Exercise 11.1 ensures that $\mathbf{E}_x(T_1) \le t_{\text{hit}}$.

How much further along is T_2 than T_1?

- When the chain visits $\sigma(1)$ before $\sigma(2)$, then $T_2 - T_1$ is the time required to travel from $\sigma(1)$ to $\sigma(2)$, and $L_2 = \sigma(2)$.
- When the chain visits $\sigma(2)$ before $\sigma(1)$, we have $T_2 - T_1 = 0$ and $L_2 = \sigma(1)$.

Let's analyze the first case a little more closely. For any two distinct states $r, s \in \Omega$, define the event

$$A_2(r, s) = \{\sigma(1) = r \text{ and } \sigma(2) = L_2 = s\}.$$

Clearly

$$\mathbf{E}_x(T_2 - T_1 \mid A_2(r, s)) = \mathbf{E}_r(\tau_s) \le t_{\text{hit}}.$$

Conveniently,

$$A_2 = \bigcup_{r \ne s} A_2(r, s)$$

is simply the event that $\sigma(2)$ is visited after $\sigma(1)$, that is, $L_2 = \sigma(2)$. By Exercise 11.1,

$$\mathbf{E}_x(T_2 - T_1 \mid A_2) \le t_{\text{hit}}.$$

Just as conveniently, A_2^c is the event that $\sigma(2)$ is visited before $\sigma(1)$. It immediately follows that

$$\mathbf{E}_x(T_2 - T_1 \mid A_2^c) = 0.$$

Since σ was chosen uniformly and independently of the chain trajectory, it is equally likely for the chain to visit $\sigma(2)$ before $\sigma(1)$ or after $\sigma(1)$. Thus

$$\begin{aligned}\mathbf{E}_x(T_2 - T_1) &= \mathbf{P}_x(A_2)\mathbf{E}_x(T_2 - T_1 \mid A_2) + \mathbf{P}_x(A_2^c)\mathbf{E}_x(T_2 - T_1 \mid A_2^c) \\ &\leq \frac{1}{2}t_{\text{hit}}.\end{aligned}$$

We estimate $T_k - T_{k-1}$ for $3 \leq k \leq n$ in the same fashion. Now we carefully track whether $L_k = \sigma(k)$ or not. For any distinct $r, s \in \Omega$, define

$$A_k(r,s) = \{\sigma(k-1) = r \text{ and } \sigma(k) = L_k = s\}.$$

Suppose $L_{k-1} = X_{T_k}$ has distribution μ. Then by Exercise 11.1 we have

$$\mathbf{E}_x(T_k - T_{k-1} \mid A_k(r,s)) = \mathbf{E}_\mu(\tau_s) = \sum_{i=1}^{n} \mu(i)\mathbf{E}_i(\tau_s) \leq t_{\text{hit}} \tag{11.3}$$

and

$$A_k = \bigcup_{r \neq s} A_k(r,s)$$

is the event that $L_k = \sigma(k)$. Just as above, Exercise 11.1 implies that

$$\mathbf{E}_x(T_k - T_{k-1} \mid A_k) \leq t_{\text{hit}},$$

while

$$\mathbf{E}_x(T_k - T_{k-1} \mid A_k^c) = 0.$$

Since the permutation σ was chosen both uniformly and independently of the trajectory of the chain, each of $\sigma(1), \ldots, \sigma(k)$ is equally likely to be the last visited. Thus $\mathbf{P}_x(A_k) = 1/k$ and

$$\begin{aligned}\mathbf{E}_x(T_k - T_{k-1}) &= \mathbf{P}_x(A_k)\mathbf{E}_x(T_k - T_{k-1} \mid A_k) + \mathbf{P}_x(A_k^c)\mathbf{E}_x(T_k - T_{k-1} \mid A_k^c) \\ &\leq \frac{1}{k}t_{\text{hit}}.\end{aligned}$$

Finally, summing all these estimates yields

$$\begin{aligned}\mathbf{E}_x(\tau_{\text{cov}}) &= \mathbf{E}_x(T_n) \\ &= \mathbf{E}_x(T_1) + \mathbf{E}_x(T_2 - T_1) + \cdots + \mathbf{E}_x(T_n - T_{n-1}) \\ &\leq t_{\text{hit}}\left(1 + \frac{1}{2} + \cdots + \frac{1}{n}\right).\end{aligned}$$

■

EXAMPLE 11.3. The proof above strongly parallels the standard argument for the coupon collecting problem, which we discussed in Section 2.2 and have applied several times: for instance, coupon collector bounds were used to lower bound mixing times for both random walk on the hypercube (Proposition 7.13) and Glauber dynamics on the graph with no edges (Exercise 7.3). For random walk on a complete graph with self-loops, the cover time coincides with the time to "collect" all coupons. In this case $\mathbf{E}_\alpha(\tau_\beta) = n$ is constant for $\alpha \neq \beta$, so the upper bound is tight.

A slight modification of this technique can be used to prove lower bounds: instead of looking for every state along the way to the cover time, we look for the elements of some $A \subseteq \Omega$. Define τ_{cov}^A to be the first time such that every state of A has been visited by the chain. When the elements of A are far away from each other, in the sense that the hitting time between any two of them is large, the time to visit just the elements of A can give a good lower bound on the overall cover time.

PROPOSITION 11.4. *Let $A \subset X$. Set $t_{\text{min}}^A = \min_{a,b\in A, a\neq b} \mathbf{E}_a(\tau_b)$. Then*

$$t_{\text{cov}} \geq \max_{A\subseteq\Omega} t_{\text{min}}^A \left(1 + \frac{1}{2} + \cdots + \frac{1}{|A|-1}\right).$$

PROOF. Fix an initial state $x \in A$ and let σ be a uniform random permutation of the elements of A, chosen independently of the chain trajectory. Let T_k be the first time at which all of $\sigma(1), \sigma(2), \ldots, \sigma(k)$ have been visited, and let $L_k = X_{T_k}$.

With probability $1/|A|$ we have $\sigma(1) = x$ and $T_1 = 0$. Otherwise, the walk must proceed from x to $\sigma(1)$. Thus

$$\mathbf{E}_x(T_1) \geq \frac{1}{|A|}0 + \frac{|A|-1}{|A|} t_{\text{min}}^A = \left(1 - \frac{1}{|A|}\right) t_{\text{min}}^A. \tag{11.4}$$

For $2 \leq k \leq |A|$ and $r, s \in A$, define

$$B_k(r,s) = \{\sigma(k-1) = r \text{ and } \sigma(k) = L_k = s\},$$

so that, by an argument similar to that of (11.3), using (an obvious corollary to) Exercise 11.1, we have

$$\mathbf{E}_x(T_k - T_{k-1} \mid B_k(r,s)) \geq t_{\text{min}}^A.$$

Then

$$B_k = \bigcup_{r,s\in A} B_k(r,s)$$

is the event that $L_k = \sigma(k)$. Now

$$\mathbf{E}_x(T_k - T_{k-1} \mid B_k^c) = 0 \quad \text{and} \quad \mathbf{E}_x(T_k - T_{k-1} | B_k) \geq t_{\text{min}}^A.$$

By the uniformity and independence of σ we have $\mathbf{P}(B_k) = 1/k$ and thus

$$\mathbf{E}_x(T_k - T_{k-1}) \geq \frac{1}{k}\, t_{\text{min}}^A. \tag{11.5}$$

Adding up (11.4) and the bound of (11.5) for $2 \leq k \leq |A|$ gives

$$\mathbf{E}_x(\tau_{\text{cov}}^A) \geq t_{\text{min}}^A \left(1 + \frac{1}{2} + \cdots + \frac{1}{|A|-1}\right)$$

(note that the negative portion of the first term cancels with the last term).

Since $t_{\text{cov}} \geq \mathbf{E}_x(\tau_{\text{cov}}) \geq \mathbf{E}_x(\tau_{\text{cov}}^A)$ for every $x \in A$, we are done. ∎

REMARK 11.5. While any subset A yields a lower bound, some choices for A are uninformative. For example, when the underlying graph of (Y_t) contains a leaf, $t_{\text{min}}^A = 1$ for any set A containing both the leaf and its (unique) neighbor.

11.3. Applications of the Matthews Method

11.3.1. Binary trees. Consider simple random walk on the rooted binary tree with depth k and $n = 2^{k+1}-1$ vertices, which we first discussed in Section 5.3.4. The maximal hitting time will be realized by pairs of leaves a, b whose most recent common ancestor is the root (see Exercise 10.4). For such a pair, the hitting time will, by symmetry, be the same as the commute time between the root and one of the leaves. By Proposition 10.6 (the Commute Time Identity), we have

$$\mathbf{E}_a \tau_b = 2(n-1)k$$

(since the effective resistance between the root and the leaf is k, by Example 9.7, and the total conductance c_G of the network is twice the number of edges). Hence Theorem 11.2 gives

$$t_{\text{cov}} \le 2(n-1)k\left(1+\frac{1}{2}+\cdots+\frac{1}{n}\right) = (2+o(1))(\log 2)nk^2. \tag{11.6}$$

What about a lower bound? We need an appropriate set $A \subseteq X$. Fix a level h in the tree, and let A be a set of 2^h leaves chosen so that each vertex at level h has a unique descendant in A. Notice that the larger h is, the more vertices there are in A—and the closer together those vertices can be. We will choose a value of h below to optimize our bound.

Consider two distinct leaves $a, b \in A$. If their closest common ancestor is at level $h' < h$, then the hitting time from one to the other is the same as the commute time from their common ancestor to one of them, say a. Again, by the Commute Time Identity (Proposition 10.6) and Example 9.7, this is exactly

$$\mathbf{E}_a \tau_b = 2(n-1)(k-h'),$$

which is clearly minimized when $h' = h-1$. By Proposition 11.4,

$$t_{\text{cov}} \ge 2(n-1)(k-h+1)\left(1+\frac{1}{2}+\cdots+\frac{1}{2^h-1}\right) = (2+o(1))(\log 2)n(k-h)h. \tag{11.7}$$

Setting $h = \lfloor k/2 \rfloor$ in (11.7) gives

$$t_{\text{cov}} \ge \frac{1}{4}\cdot(2+o(1))(\log 2)nk^2. \tag{11.8}$$

There is still a factor of 4 gap between the upper bound of (11.6) and the lower bound of (11.8). In fact, the upper bound is sharp. See the Notes.

11.3.2. Tori. In Section 10.4 we derived fairly sharp (up to constants) estimates for the hitting times of simple random walks on finite tori of various dimensions. Let's use these bounds and the Matthews method to determine equally sharp bounds on the expected cover times of tori. We discuss the case of dimension at least 3 first, since the details are a bit simpler.

When the dimension $d \ge 3$, Proposition 10.13 tells us that there exist constants c_d and C_d such that for any distinct vertices x, y of $\mathbb{Z}_n^d$,

$$c_d n^d \le \mathbf{E}_x(\tau_y) \le C_d n^d.$$

CHAPTER 12

Eigenvalues

12.1. The Spectral Representation of a Reversible Transition Matrix

We begin by collecting some elementary facts about the eigenvalues of transition matrices, which we leave to the reader to verify (Exercise 12.1):

LEMMA 12.1. *Let P be the transition matrix of a finite Markov chain.*

(i) *If λ is an eigenvalue of P, then $|\lambda| \le 1$.*
(ii) *If P is irreducible, the vector space of eigenfunctions corresponding to the eigenvalue 1 is the one-dimensional space generated by the column vector $\mathbf{1} := (1, 1, \ldots, 1)^T$.*
(iii) *If P is irreducible and aperiodic, then -1 is not an eigenvalue of P.*

Denote by $\langle \cdot, \cdot \rangle$ the usual inner product on $\mathbb{R}^\Omega$, given by $\langle f, g \rangle = \sum_{x \in \Omega} f(x)g(x)$. We will also need another inner product, denoted by $\langle \cdot, \cdot \rangle_\pi$ and defined by

$$\langle f, g \rangle_\pi := \sum_{x \in \Omega} f(x)g(x)\pi(x). \tag{12.1}$$

We write $\ell^2(\pi)$ for the vector space $\mathbb{R}^\Omega$ equipped with the inner product (12.1). Because we regard elements of $\mathbb{R}^\Omega$ as functions from Ω to $\mathbb{R}$, we will call eigenvectors of the matrix P eigenfunctions.

Recall that the transition matrix P is reversible with respect to the stationary distribution π if $\pi(x)P(x, y) = \pi(y)P(y, x)$ for all $x, y \in \Omega$. The reason for introducing the inner product (12.1) is

LEMMA 12.2. *Let P be reversible with respect to π.*

(i) *The inner product space $(\mathbb{R}^\Omega, \langle \cdot, \cdot \rangle_\pi)$ has an orthonormal basis of real-valued eigenfunctions $\{f_j\}_{j=1}^{|\Omega|}$ corresponding to real eigenvalues $\{\lambda_j\}$.*
(ii) *The matrix P can be decomposed as*

$$\frac{P^t(x, y)}{\pi(y)} = \sum_{j=1}^{|\Omega|} f_j(x)f_j(y)\lambda_j^t.$$

(iii) *The eigenfunction f_1 corresponding to the eigenvalue 1 can be taken to be the constant vector $\mathbf{1}$, in which case*

$$\frac{P^t(x, y)}{\pi(y)} = 1 + \sum_{j=2}^{|\Omega|} f_j(x)f_j(y)\lambda_j^t. \tag{12.2}$$

PROOF. Define $A(x, y) := \pi(x)^{1/2}\pi(y)^{-1/2}P(x, y)$. Reversibility of P implies that A is symmetric. The spectral theorem for symmetric matrices (Theorem A.11) guarantees that the inner product space $(\mathbb{R}^\Omega, \langle \cdot, \cdot \rangle)$ has an orthonormal basis $\{\varphi_j\}_{j=1}^{|\Omega|}$ such that φ_j is an eigenfunction with real eigenvalue λ_j.

The reader should directly check that $\sqrt{\pi}$ is an eigenfunction of A with corresponding eigenvalue 1; we set $\varphi_1 := \sqrt{\pi}$ and $\lambda_1 := 1$.

If D_π denotes the diagonal matrix with diagonal entries $D_\pi(x,x) = \pi(x)$, then $A = D_\pi^{\frac{1}{2}} P D_\pi^{-\frac{1}{2}}$. If $f_j := D_\pi^{-\frac{1}{2}}\varphi_j$, then f_j is an eigenfunction of P with eigenvalue λ_j:

$$Pf_j = PD_\pi^{-\frac{1}{2}}\varphi_j = D_\pi^{-\frac{1}{2}}(D_\pi^{\frac{1}{2}} P D_\pi^{-\frac{1}{2}})\varphi_j = D_\pi^{-\frac{1}{2}} A\varphi_j = D_\pi^{-\frac{1}{2}}\lambda_j\varphi_j = \lambda_j f_j.$$

Although the eigenfunctions $\{f_j\}$ are not necessarily orthonormal with respect to the usual inner product, they are orthonormal with respect to the inner product $\langle\cdot,\cdot\rangle_\pi$ defined in (12.1):

$$\delta_{ij} = \langle\varphi_i,\varphi_j\rangle = \langle D_\pi^{\frac{1}{2}} f_i, D_\pi^{\frac{1}{2}} f_j\rangle = \langle f_i, f_j\rangle_\pi. \tag{12.3}$$

(The first equality follows since $\{\varphi_j\}$ is orthonormal with respect to the usual inner product.) This proves (i).

Let δ_y be the function

$$\delta_y(x) = \begin{cases} 1 & \text{if } y = x, \\ 0 & \text{if } y \neq x. \end{cases}$$

Considering $(\mathbb{R}^\Omega, \langle\cdot,\cdot\rangle_\pi)$ with its orthonormal basis of eigenfunctions $\{f_j\}_{j=1}^{|\Omega|}$, the function δ_y can be written via basis decomposition as

$$\delta_y = \sum_{j=1}^{|\Omega|} \langle\delta_y, f_j\rangle_\pi f_j = \sum_{j=1}^{|\Omega|} f_j(y)\pi(y) f_j. \tag{12.4}$$

Since $P^t f_j = \lambda_j^t f_j$ and $P^t(x,y) = (P^t\delta_y)(x)$,

$$P^t(x,y) = \sum_{j=1}^{|\Omega|} f_j(y)\pi(y)\lambda_j^t f_j(x).$$

Dividing by $\pi(y)$ completes the proof of (ii), and (iii) follows from observations above. ■

It follows from Lemma 12.2 that for a function $f : \Omega \to \mathbb{R}$,

$$P^t f = \sum_{j=1}^{|\Omega|} \langle f, f_j\rangle_\pi f_j \lambda_j^t. \tag{12.5}$$

12.2. The Relaxation Time

For a reversible transition matrix P, we label the eigenvalues of P in decreasing order:

$$1 = \lambda_1 > \lambda_2 \geq \cdots \geq \lambda_{|\Omega|} \geq -1. \tag{12.6}$$

Define

$$\lambda_\star := \max\{|\lambda| : \lambda \text{ is an eigenvalue of } P,\ \lambda \neq 1\}. \tag{12.7}$$

The difference $\gamma_\star := 1 - \lambda_\star$ is called the ***absolute spectral gap***. Lemma 12.1 implies that if P is aperiodic and irreducible, then $\gamma_\star > 0$.

The ***spectral gap*** of a reversible chain is defined by $\gamma := 1 - \lambda_2$. Exercise 12.3 shows that if the chain is lazy, then $\gamma_\star = \gamma$.

The ***relaxation time*** t_{rel} of a reversible Markov chain with absolute spectral gap $\gamma_\star$ is defined to be

$$t_{\text{rel}} := \frac{1}{\gamma_\star}.$$

One operational meaning of the relaxation time comes from the inequality

$$\operatorname{Var}_\pi(P^t f) \le (1-\gamma_\star)^{2t} \operatorname{Var}_\pi(f). \tag{12.8}$$

(Exercise 12.4 asks for a proof.) By the Convergence Theorem (Theorem 4.9), $P^t f(x) \to E_\pi(f)$ for any $x \in \Omega$, i.e., the function $P^t f$ approaches a constant function. Using (12.8), we can make a quantitative statement: if $t \ge t_{\text{rel}}$, then the standard deviation of $P^t f$ is bounded by $1/e$ times the standard deviation of f. Let $i_\star$ be the value for which $|\lambda_{i_\star}|$ is maximized. Then equality in (12.8) is achieved for $f = f_{i_\star}$, whence the inequality is sharp.

We prove both upper and lower bounds on the mixing time in terms of the relaxation time and the stationary distribution of the chain.

THEOREM 12.3. *Let P be the transition matrix of a reversible, irreducible Markov chain with state space Ω, and let $\pi_{\min} := \min_{x\in\Omega} \pi(x)$. Then*

$$t_{\text{mix}}(\varepsilon) \le \log\left(\frac{1}{\varepsilon \pi_{\min}}\right) t_{\text{rel}}. \tag{12.9}$$

PROOF. Using (12.2) and applying the Cauchy-Schwarz inequality yields

$$\left|\frac{P^t(x,y)}{\pi(y)} - 1\right| \le \sum_{j=2}^{|\Omega|} |f_j(x) f_j(y)| \lambda_\star^t \le \lambda_\star^t \left[\sum_{j=2}^{|\Omega|} f_j^2(x) \sum_{j=2}^{|\Omega|} f_j^2(y)\right]^{1/2}. \tag{12.10}$$

Using (12.4) and the orthonormality of $\{f_j\}$ shows that

$$\pi(x) = \langle \delta_x, \delta_x \rangle_\pi = \left\langle \sum_{j=1}^{|\Omega|} f_j(x)\pi(x) f_j, \sum_{j=1}^{|\Omega|} f_j(x)\pi(x) f_j \right\rangle_\pi = \pi(x)^2 \sum_{j=1}^{|\Omega|} f_j(x)^2.$$

Consequently, $\sum_{j=2}^{|\Omega|} f_j(x)^2 \le \pi(x)^{-1}$. This bound and (12.10) imply that

$$\left|\frac{P^t(x,y)}{\pi(y)} - 1\right| \le \frac{\lambda_\star^t}{\sqrt{\pi(x)\pi(y)}} \le \frac{\lambda_\star^t}{\pi_{\min}} = \frac{(1-\gamma_\star)^t}{\pi_{\min}} \le \frac{e^{-\gamma_\star t}}{\pi_{\min}}. \tag{12.11}$$

Applying Lemma 6.13 shows that $d(t) \le \pi_{\min}^{-1} \exp(-\gamma_\star t)$. The conclusion now follows from the definition of $t_{\text{mix}}(\varepsilon)$. ∎

THEOREM 12.4. *For a reversible, irreducible, and aperiodic Markov chain,*

$$t_{\text{mix}}(\varepsilon) \ge (t_{\text{rel}} - 1) \log\left(\frac{1}{2\varepsilon}\right). \tag{12.12}$$

REMARK 12.5. If the absolute spectral gap $\gamma_\star$ is small because the smallest eigenvalue $\lambda_{|\Omega|}$ is near -1, but the spectral gap γ is not small, the slow mixing suggested by this lower bound can be rectified by passing to a lazy chain to make the eigenvalues positive.

PROOF. Suppose that f is an eigenfunction of P with eigenvalue $\lambda \neq 1$, so that $Pf = \lambda f$. Since the eigenfunctions are orthogonal with respect to $\langle \cdot, \cdot \rangle_\pi$ and $\mathbf{1}$ is an eigenfunction, $\sum_{y \in \Omega} \pi(y) f(y) = \langle \mathbf{1}, f \rangle_\pi = 0$. It follows that

$$|\lambda^t f(x)| = |P^t f(x)| = \left| \sum_{y \in \Omega} \left[P^t(x,y) f(y) - \pi(y) f(y) \right] \right| \leq \|f\|_\infty 2d(t).$$

With this inequality, we can obtain a lower bound on the mixing time. Taking x with $|f(x)| = \|f\|_\infty$ yields

$$|\lambda|^t \leq 2d(t). \tag{12.13}$$

Therefore, $|\lambda|^{t_{\mathrm{mix}}(\varepsilon)} \leq 2\varepsilon$, whence

$$t_{\mathrm{mix}}(\varepsilon) \left(\frac{1}{|\lambda|} - 1 \right) \geq t_{\mathrm{mix}}(\varepsilon) \log \left(\frac{1}{|\lambda|} \right) \geq \log \left(\frac{1}{2\varepsilon} \right).$$

Minimizing the left-hand side over eigenvalues different from 1 and rearranging finishes the proof. ■

COROLLARY 12.6. *For a reversible, irreducible, and aperiodic Markov chain,*

$$\lim_{t \to \infty} d(t)^{1/t} = \lambda_\star.$$

PROOF. One direction is immediate from (12.13), and the other follows from (12.11). ■

EXAMPLE 12.7 (Relaxation time of random transpositions). By Corollary 8.10 and Proposition 8.11, we know that for the random transpositions chain on n cards,

$$t_{\mathrm{mix}} = \Theta(n \log n).$$

Hence $t_{\mathrm{rel}} = O(n \log n)$. The stationary distribution is uniform on $\mathcal{S}_n$. Since Stirling's Formula implies $\log(n!) \sim n \log n$, Theorem 12.3 gives only a constant lower bound. In fact, the relaxation time is known (through other methods) to be exactly $n/2$. See Diaconis (1988).

12.3. Eigenvalues and Eigenfunctions of Some Simple Random Walks

Simple random walk on the n-cycle was introduced in Example 1.4. In Example 2.10, we noted that it can be viewed as a random walk on an n-element cyclic group. Here we use that interpretation to find the eigenvalues and eigenfunctions of this chain and some closely related chains.

12.3.1. The cycle. Let $\omega = e^{2\pi i/n}$. In the complex plane, the set $W_n := \{\omega, \omega^2, \ldots, \omega^{n-1}, 1\}$ of the ***n-th roots of unity*** forms a regular n-gon inscribed in the unit circle. Since $\omega^n = 1$, we have

$$\omega^j \omega^k = \omega^{k+j} = \omega^{k+j \bmod n}.$$

Hence $(W_n, \cdot)$ is a cyclic group of order n, generated by ω. In this section, we view simple random walk on the n-cycle as the random walk on the (multiplicative) group W_n with increment distribution uniform on $\{\omega, \omega^{-1}\}$. Let P be the transition matrix of this walk. Every (possibly complex-valued) eigenfunction f of P satisfies

$$\lambda f(\omega^k) = Pf(\omega^k) = \frac{f(\omega^{k-1}) + f(\omega^{k+1})}{2}$$

for $0 \leq k \leq n-1$.

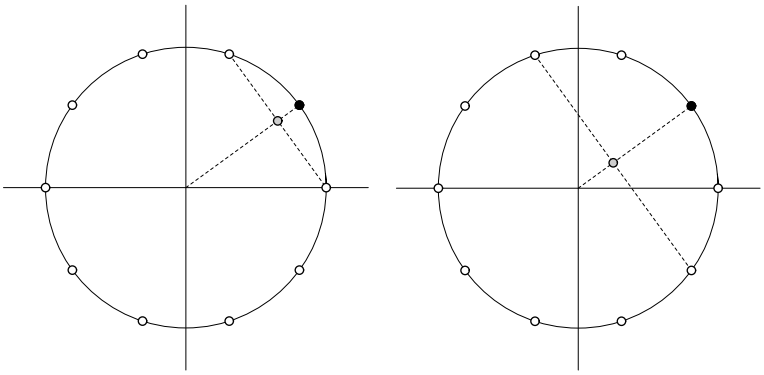


FIGURE 12.1. For simple random walk on the cycle, the eigenvalues must be the cosines. Here $n = 10$. The black vertices represent $\omega = e^{2\pi i/10}$, while the grey vertices represent $(1/2)(\omega^2 + \omega^0)$ and $(1/2)(\omega^3 + \omega^{-1})$, respectively.

For $0 \le j \le n-1$, define $\varphi_j(\omega^k) := \omega^{kj}$. Then

$$P\varphi_j(\omega^k) = \frac{\varphi_j(\omega^{k-1}) + \varphi_j(\omega^{k+1})}{2} = \frac{\omega^{jk+j} + \omega^{jk-j}}{2} = \omega^{jk}\left(\frac{\omega^j + \omega^{-j}}{2}\right). \quad (12.14)$$

Hence φ_j is an eigenfunction of P with eigenvalue $\frac{\omega^j+\omega^{-j}}{2} = \cos(2\pi j/n)$. What is the underlying geometry? As Figure 12.1 illustrates, for any ℓ and j the average of the vectors $\omega^{\ell-j}$ and $\omega^{\ell+j}$ is a scalar multiple of ω^ℓ. Since the chord connecting $\omega^{\ell+j}$ with $\omega^{\ell-j}$ is perpendicular to ω^ℓ, the projection of $\omega^{\ell+j}$ onto ω^ℓ has length $\cos(2\pi j/n)$.

Because φ_j is an eigenfunction of the real matrix P with a real eigenvalue, both its real part and its imaginary parts are eigenfunctions. In particular, the function $f_j : W_n \to \mathbb{R}$ defined by

$$f_j(\omega^k) = \mathrm{Re}(\varphi_j(\omega^k)) = \mathrm{Re}(e^{2\pi ijk/n}) = \cos\left(\frac{2\pi jk}{n}\right) \quad (12.15)$$

is an eigenfunction. We note for future reference that f_j is invariant under complex conjugation of the states of the chain.

We have $\lambda_2 = \cos(2\pi/n) = 1 - \frac{4\pi^2}{2n^2} + O(n^{-4})$, so the spectral gap γ is of order n^{-2} and the relaxation time is of order n^2.

When $n = 2m$ is even, $\cos(2\pi m/n) = -1$ is an eigenvalue, so $\gamma_\star = 0$. The walk in this case is periodic, as we pointed out in Example 1.8.

12.3.2. Lumped chains and the path. Consider the projection of simple random walk on the n-th roots of unity, as described in the preceding section, onto the real axis. The resulting process can take values on a discrete set of points. At most of them (ignoring for the moment those closest to 1 and -1), it is equally likely to move to the right or to the left—just like random walk on the path. Using this idea, we can determine the eigenvalues and eigenfunctions of the random walk on a path with either reflecting boundary conditions or an even chance of holding at the endpoints. First, we give a general lemma on the eigenvalues and eigenfunctions of projected chains (defined in Section 2.3.1).

CHAPTER 14

The Transportation Metric and Path Coupling

Let P be a transition matrix on a metric space (Ω, ρ), where the metric ρ satisfies $\rho(x,y) \geq \mathbf{1}\{x \neq y\}$. Suppose, for all states x and y, there exists a coupling (X_1, Y_1) of $P(x,\cdot)$ with $P(y,\cdot)$ that contracts ρ on average, i.e., which satisfies

$$\mathbf{E}_{x,y}\rho(X_1, Y_1) \leq e^{-\alpha}\rho(x,y) \tag{14.1}$$

for some $\alpha > 0$. The ***diameter*** of Ω is defined to be $\operatorname{diam}(\Omega) := \max_{x,y\in\Omega} \rho(x,y)$. By iterating (14.1), we have

$$\mathbf{E}_{x,y}\rho(X_t, Y_t) \leq e^{-\alpha t}\operatorname{diam}(\Omega).$$

We conclude that

$$\begin{aligned}\left\|P^t(x,\cdot) - P^t(y,\cdot)\right\|_{\mathrm{TV}} &\leq \mathbf{P}_{x,y}\{X_t \neq Y_t\} = \mathbf{P}_{x,y}\{\rho(X_t, Y_t) \geq 1\} \\ &\leq \mathbf{E}_{x,y}\rho(X_t, Y_t) \leq \operatorname{diam}(\Omega)e^{-\alpha t},\end{aligned}$$

whence

$$t_{\mathrm{mix}}(\varepsilon) \leq \left\lceil \frac{1}{\alpha}\left[\log(\operatorname{diam}(\Omega)) + \log(1/\varepsilon)\right]\right\rceil.$$

This is the method used in Theorem 5.7 to bound the mixing time of the Metropolis chain for proper colorings and also used in Theorem 5.8 for the hardcore chain.

Path coupling is a technique that simplifies the construction of couplings satisfying (14.1), when ρ is a *path metric*, defined below. While the argument just given requires verification of (14.1) *for all* pairs $x, y \in \Omega$, the path-coupling technique shows that it is enough to construct couplings satisfying (14.1) only for neighboring pairs.

14.1. The Transportation Metric

Recall that a coupling of probability distributions μ and ν is a pair (X, Y) of random variables defined on a single probability space such that X has distribution μ and Y has distribution ν.

For a given distance ρ defined on the state space Ω, the ***transportation metric*** between two distributions on Ω is defined by

$$\rho_K(\mu,\nu) := \inf\{\mathbf{E}(\rho(X,Y)) : (X,Y) \text{ is a coupling of } \mu \text{ and } \nu\}. \tag{14.2}$$

By Proposition 4.7, if $\rho(x,y) = \mathbf{1}_{\{x\neq y\}}$, then $\rho_K(\mu,\nu) = \|\mu - \nu\|_{\mathrm{TV}}$.

REMARK 14.1. It is sometimes convenient to describe couplings using probability distributions on the product space $\Omega\times\Omega$, instead of random variables. When q is a probability distribution on $\Omega\times\Omega$, its ***projection onto the first coordinate*** is the probability distribution on Ω equal to

$$q(\cdot\times\Omega) = \sum_{y\in\Omega} q(\cdot, y).$$

Likewise, its ***projection onto the second coordinate*** is the distribution $q(\Omega\times\cdot)$.

Given a coupling (X,Y) of μ and ν as defined above, the distribution of (X,Y) on $\Omega\times\Omega$ has projections μ and ν on the first and second coordinates, respectively. Conversely, given a probability distribution q on $\Omega\times\Omega$ with projections μ and ν, the identity function on the probability space $(\Omega\times\Omega, q)$ is a coupling of μ and ν.

Consequently, since $\mathbf{E}(\rho(X,Y)) = \sum_{(x,y)\in\Omega\times\Omega}\rho(x,y)q(x,y)$ when (X,Y) has distribution q, the transportation metric can also be written as

$$\rho_K(\mu,\nu) = \inf\left\{\sum_{(x,y)\in\Omega\times\Omega}\rho(x,y)q(x,y) \,:\, q(\cdot\times\Omega)=\mu,\ q(\Omega\times\cdot)=\nu\right\}. \quad (14.3)$$

REMARK 14.2. The set of probability distributions on $\Omega\times\Omega$ can be identified with the ($|\Omega|^2$-1)-dimensional simplex, which is a compact subset of $\mathbb{R}^{|\Omega|^2}$. The set of distributions on $\Omega\times\Omega$ which project on the first coordinate to μ and project on the second coordinate to ν is a closed subset of this simplex and hence is compact. The function

$$q \mapsto \sum_{(x,y)\in\Omega\times\Omega}\rho(x,y)q(x,y)$$

is continuous on this set. Hence there is a $q_\star$ such that

$$\sum_{(x,y)\in\Omega\times\Omega}\rho(x,y)q_\star(x,y) = \rho_K(\mu,\nu).$$

Such a $q_\star$ is called an ***optimal coupling*** of μ and ν. Equivalently, there is a pair of random variables $(X_\star, Y_\star)$, also called an optimal coupling, such that

$$\mathbf{E}(\rho(X_\star,Y_\star)) = \rho_K(\mu,\nu).$$

LEMMA 14.3. *The function ρ_K defined in (14.2) is a metric on the space of probability distributions on Ω.*

PROOF. We check the triangle inequality and leave the verification of the other two conditions to the reader.

Let μ,ν and η be probability distributions on Ω. Let p be a probability distribution on $\Omega\times\Omega$ which is a coupling of μ and ν, and let q be a probability distribution on $\Omega\times\Omega$ which is a coupling of ν and η. Define the probability distribution r on $\Omega\times\Omega\times\Omega$ by

$$r(x,y,z) := \frac{p(x,y)q(y,z)}{\nu(y)}. \quad (14.4)$$

(See Remark 14.4 for the motivation of this definition.) Note that the projection of r onto its first two coordinates is p, and the projection of r onto its last two coordinates is q. The projection of r onto the first and last coordinates is a coupling of μ and η.

Assume now that p is an optimal coupling of μ and ν. (See Remark 14.2.) Likewise, suppose that q is an optimal coupling of ν and η.

Let (X,Y,Z) be a random vector with probability distribution r. Since ρ is a metric,

$$\rho(X,Z) \le \rho(X,Y) + \rho(Y,Z).$$

Taking expectation, because (X,Y) is an optimal coupling of μ and ν and (Y,Z) is an optimal coupling of ν and η,

$$\mathbf{E}(\rho(X,Z)) \le \mathbf{E}(\rho(X,Y)) + \mathbf{E}(\rho(Y,Z)) = \rho_K(\mu,\nu) + \rho_K(\nu,\eta).$$

Since (X, Z) is a coupling (although not necessarily optimal) of μ and η, we conclude that

$$\rho_K(\mu, \eta) \le \rho_K(\mu, \nu) + \rho_K(\nu, \eta).$$

■

The transportation metric ρ_K extends the metric ρ on Ω to a metric on the space of probability distributions on Ω. In particular, if δ_x denotes the probability distribution which puts unit mass on x, then $\rho_K(\delta_x, \delta_y) = \rho(x, y)$.

REMARK 14.4. The probability distribution r defined in (14.4) can be thought of as three steps of a time-inhomogeneous Markov chain. The first state X is generated according to μ. Given $X = x$, the second state Y is generated according to $p(x, \cdot)/\mu(x)$, and given $Y = y$, the third state Z is generated according to $q(y, \cdot)/\nu(y)$. Thus,

$$\mathbf{P}\{X = x, Y = y, Z = z\} = \mu(x)\frac{p(x, y)}{\mu(x)}\frac{q(y, z)}{\nu(y)} = r(x, y, z).$$

14.2. Path Coupling

Suppose that the state space Ω of a Markov chain (X_t) is the vertex set of a connected graph $G = (\Omega, E_0)$ and ℓ is a length function defined on E_0. That is, ℓ assigns length $\ell(x, y)$ to each edge $\{x, y\} \in E_0$. We assume that $\ell(x, y) \ge 1$ for all edges $\{x, y\}$.

REMARK 14.5. This graph structure may be different from the structure inherited from the permissible transitions of the Markov chain (X_t).

Define a ***path*** in Ω from x to y to be a sequence of states $\xi = (x_0, x_1, \ldots, x_r)$ such that $x_0 = x$ and $x_r = y$ and such that $\{x_{i-1}, x_i\}$ is an edge for $i = 1, \ldots, r$. The ***length*** of the path is defined to be $\sum_{i=1}^{r} \ell(x_{i-1}, x_i)$. The ***path metric*** on Ω is defined by

$$\rho(x, y) = \min\{\text{length of } \xi : \xi \text{ a path from } x \text{ to } y\}. \tag{14.5}$$

Since we have assumed that $\ell(x, y) \ge 1$, it follows that $\rho(x, y) \ge \mathbf{1}\{x \ne y\}$, whence for any pair (X, Y),

$$\mathbf{P}\{X \ne Y\} = \mathbf{E}\left(\mathbf{1}_{\{X \ne Y\}}\right) \le \mathbf{E}\rho(X, Y). \tag{14.6}$$

Minimizing over all couplings (X, Y) of μ and ν shows that

$$\rho_{\mathrm{TV}}(\mu, \nu) \le \rho_K(\mu, \nu). \tag{14.7}$$

While Bubley and Dyer (1997) discovered the following theorem and applied it to mixing, the key idea is the application of the triangle inequality for the transportation metric, which goes back to Kantorovich (1942).

THEOREM 14.6 (Bubley and Dyer (1997)). *Suppose the state space Ω of a Markov chain is the vertex set of a graph with length function ℓ defined on edges. Let ρ be the corresponding path metric defined in (14.5). Suppose that for each edge $\{x, y\}$ there exists a coupling (X_1, Y_1) of the distributions $P(x, \cdot)$ and $P(y, \cdot)$ such that*

$$\mathbf{E}_{x,y}\left(\rho(X_1, Y_1)\right) \le \rho(x, y)e^{-\alpha} = \ell(x, y)e^{-\alpha}. \tag{14.8}$$

Then for any two probability measures μ and ν on Ω,

$$\rho_K(\mu P, \nu P) \le e^{-\alpha}\rho_K(\mu, \nu). \tag{14.9}$$

Recall that $d(t) = \max_{x\in\Omega} \|P^t(x,\cdot) - \pi\|_{\mathrm{TV}}$ and $\mathrm{diam}(\Omega) = \max_{x,y\in\Omega} \rho(x,y)$.

COROLLARY 14.7. *Suppose that the hypotheses of Theorem 14.6 hold. Then*

$$d(t) \le e^{-\alpha t}\mathrm{diam}(\Omega),$$

and consequently

$$t_{\mathrm{mix}}(\varepsilon) \le \left\lceil \frac{-\log(\varepsilon) + \log(\mathrm{diam}(\Omega))}{\alpha} \right\rceil.$$

PROOF. By iterating (14.9), it follows that

$$\rho_K(\mu P^t, \nu P^t) \le e^{-\alpha t}\rho_K(\mu,\nu) \le e^{-\alpha t}\max_{x,y}\rho(x,y). \tag{14.10}$$

Applying (14.7) and setting $\mu = \delta_x$ and $\nu = \pi$ shows that

$$\left\|P^t(x,\cdot) - \pi\right\|_{\mathrm{TV}} \le e^{-\alpha t}\mathrm{diam}(\Omega). \tag{14.11}$$

■

PROOF OF THEOREM 14.6. We begin by showing that for arbitrary (not necessarily neighboring) $x, y \in \Omega$,

$$\rho_K(P(x,\cdot), P(y,\cdot)) \le e^{-\alpha}\rho(x,y). \tag{14.12}$$

Fix $x, y \in \Omega$, and let $(x_0, x_1, \ldots, x_r)$ be a path achieving the minimum in (14.5). By the triangle inequality for ρ_K,

$$\rho_K(P(x,\cdot), P(y,\cdot)) \le \sum_{k=1}^{r} \rho_K(P(x_{k-1},\cdot), P(x_k,\cdot)). \tag{14.13}$$

Since ρ_K is a minimum over all couplings, the hypotheses of the theorem imply that, for any edge $\{a,b\}$,

$$\rho_K(P(a,\cdot), P(b,\cdot)) \le e^{-\alpha}\ell(a,b). \tag{14.14}$$

Using the bound (14.14) on each of the terms in the sum appearing on the right-hand side of (14.13) shows that

$$\rho_K(P(x,\cdot), P(y,\cdot)) \le e^{-\alpha}\sum_{k=1}^{r} \ell(x_{k-1}, x_k).$$

Since the path $(x_0, \ldots, x_k)$ was chosen to be of shortest length, the sum on the right-hand side above equals $\rho(x,y)$. This establishes (14.12).

Let η by an optimal coupling of μ and ν, so that

$$\rho_K(\mu,\nu) = \sum_{x,y\in\Omega} \rho(x,y)\eta(x,y). \tag{14.15}$$

By (14.12), we know that for all x, y there exists a coupling $\theta_{x,y}$ of $P(x,\cdot)$ and $P(y,\cdot)$ such that

$$\sum_{u,w\in\Omega} \rho(u,w)\theta_{x,y}(u,w) \le e^{-\alpha}\rho(x,y). \tag{14.16}$$

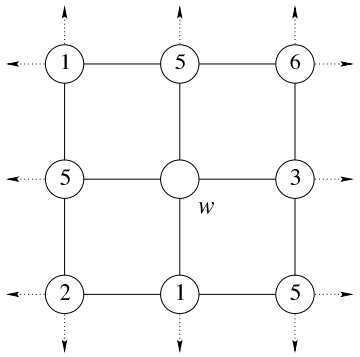


FIGURE 14.1. Updating at vertex w. The colors of the neighbors are not available, as indicated.

Consider the probability distribution $\theta := \sum_{x,y\in\Omega} \eta(x,y)\theta_{x,y}$ on $\Omega\times\Omega$. (This is a coupling of μP with νP.) We have by (14.16) and (14.15) that

$$\begin{aligned}\sum_{u,w\in\Omega} \rho(u,w)\theta(u,w) &= \sum_{x,y\in\Omega}\sum_{u,w\in\Omega} \rho(u,w)\theta_{x,y}(u,w)\eta(x,y)\\ &\le e^{-\alpha}\sum_{x,y\in\Omega}\rho(x,y)\eta(x,y)\\ &= e^{-\alpha}\rho_K(\mu,\nu).\end{aligned}$$

Therefore, the theorem is proved, because $\rho_K(\mu P,\nu P)\le \sum_{u,w\in\Omega}\rho(u,w)\theta(u,w)$. ■

14.3. Fast Mixing for Colorings

Recall from Section 3.1 that proper q-colorings of a graph $G=(V,E)$ are elements of $x\in\Omega=\{1,2,\dots,q\}^V$ such that $x(v)\neq x(w)$ for $\{v,w\}\in E$.

In Section 5.4.1, the mixing time of the Metropolis chain for proper q-colorings was analyzed for sufficiently large q. Here we analyze the mixing time for the Glauber dynamics.

As defined in Section 3.3, Glauber dynamics for proper q-colorings of a graph G with n vertices operate as follows: at each move, a vertex is chosen uniformly at random and the color of this vertex is updated. To update, a color is chosen uniformly at random from the allowable colors, which are those colors not seen among the neighbors of the chosen vertex.

We will use path coupling to bound the mixing time of this chain.

THEOREM 14.8. *Consider the Glauber dynamics chain for random proper q-colorings of a graph with n vertices and maximum degree Δ. If $q>2\Delta$, then the mixing time satisfies*

$$t_{\text{mix}}(\varepsilon)\le\left\lceil\left(\frac{q-\Delta}{q-2\Delta}\right)n\left(\log n-\log\varepsilon\right)\right\rceil. \tag{14.17}$$

CHAPTER 15

The Ising Model

The Ising model on a graph with vertex set V at inverse temperature β was introduced in Section 3.3.5. It is the probability distribution on $\Omega = \{-1, 1\}^V$ defined by

$$\pi(\sigma) = Z(\beta)^{-1} \exp\left(\beta \sum_{\substack{v,w \in V \\ v \sim w}} \sigma(v)\sigma(w)\right).$$

Here we study in detail the Glauber dynamics for this distribution. As discussed in Section 3.3.5, the transition matrix for this chain is given by

$$P(\sigma, \sigma') = \frac{1}{n} \sum_{v \in V} \frac{e^{\beta\, \sigma'(v)\, S(\sigma, v)}}{e^{\beta\, \sigma'(v)\, S(\sigma, v)} + e^{-\beta\, \sigma'(v)\, S(\sigma, v)}} \cdot \mathbf{1}_{\{\sigma'(w) = \sigma(w) \text{ for all } w \neq v\}},$$

where $S(\sigma, v) = \sum_{w\,:\, w \sim v} \sigma(w)$.

This chain evolves by selecting a vertex v at random and updating the spin at v according to the distribution π conditioned to agree with the spins at all vertices not equal to v. If the current configuration is σ and vertex v is selected, then the chance the spin at v is updated to $+1$ is equal to

$$p(\sigma, v) := \frac{e^{\beta S(\sigma, v)}}{e^{\beta S(\sigma, v)} + e^{-\beta S(\sigma, v)}} = \frac{1 + \tanh(\beta S(\sigma, v))}{2}.$$

We will be particularly interested in how the mixing time varies with β. Generically, for small values of β, the chain will mix in a short amount of time, while for large values of β, the chain will converge slowly. Understanding this phase transition between slow and fast mixing has been a topic of great interest and activity over the past twenty years; here we only scratch the surface.

15.1. Fast Mixing at High Temperature

In this section we use the path coupling technique of Chapter 14 to show that on any graph of bounded degree, for small values of β, the Glauber dynamics for the Ising model is fast mixing.

THEOREM 15.1. *Consider the Glauber dynamics for the Ising model on a graph with n vertices and maximal degree Δ.*

(i) *Let $c(\beta) := 1 - \Delta \tanh(\beta)$. If $\Delta \cdot \tanh(\beta) < 1$, then*

$$t_{\text{mix}}(\varepsilon) \leq \left\lceil \frac{n(\log n + \log(1/\varepsilon))}{c(\beta)} \right\rceil. \tag{15.1}$$

In particular, (15.1) holds whenever $\beta < \Delta^{-1}$.

(ii) *Suppose every vertex of the graph has even degree. Let*

$$c_e(\beta) := 1 - (\Delta/2)\tanh(2\beta).$$

If $(\Delta/2)\cdot\tanh(2\beta) < 1$, *then*

$$t_{\text{mix}}(\varepsilon) \le \left\lceil \frac{n(\log n + \log(1/\varepsilon))}{c_e(\beta)} \right\rceil. \tag{15.2}$$

LEMMA 15.2. *The function* $\varphi(x) := \tanh(\beta(x+1)) - \tanh(\beta(x-1))$ *is even and decreasing on* $[0,\infty)$, *whence*

$$\sup_{x\in\mathbb{R}} \varphi(x) = \varphi(0) = 2\tanh(\beta) \tag{15.3}$$

and

$$\sup_{k \text{ odd integer}} \varphi(k) = \varphi(1) = \tanh(2\beta). \tag{15.4}$$

PROOF. Let $\psi(x) := \tanh(\beta x)$; observe that $\psi'(x) = \beta/\cosh^2(\beta x)$. The function ψ' is strictly positive and decreasing on $[0,\infty)$ and is even. Therefore, for $x > 0$,

$$\varphi'(x) = \psi'(x+1) - \psi'(x-1) < 0,$$

as is seen by considering separately the case where $x - 1 > 0$ and the case where $x - 1 \le 0$. Because tanh is an odd function,

$$\varphi(-x) = \psi(-x+1) - \psi(-x-1) = -\psi(x-1) + \psi(x+1) = \varphi(x),$$

so φ is even. ∎

PROOF OF THEOREM 15.1. Define the distance ρ on Ω by

$$\rho(\sigma,\tau) = \frac{1}{2}\sum_{u\in V} |\sigma(u) - \tau(u)|.$$

The distance ρ is a path metric as defined in Section 14.2.

Let σ and τ be two configurations with $\rho(\sigma,\tau) = 1$. The spins of σ and τ agree everywhere except at a single vertex v. Assume that $\sigma(v) = -1$ and $\tau(v) = +1$.

Define $\mathcal{N}(v) := \{u : u \sim v\}$ to be the set of neighboring vertices to v.

We describe now a coupling (X, Y) of one step of the chain started in configuration σ with one step of the chain started in configuration τ.

Pick a vertex w uniformly at random from V. If $w \notin \mathcal{N}(v)$, then the neighbors of w agree in both σ and τ. As the probability of updating the spin at w to $+1$, given in (3.10), depends only on the spins at the neighbors of w, it is the same for the chain started in σ as for the chain started in τ. Thus we can update both chains together.

If $w \in \mathcal{N}(v)$, the probabilities of updating to $+1$ at w are no longer the same for the two chains, so we cannot *always* update together. We do, however, use a single random variable as the common source of noise to update both chains, so the two chains agree as often as is possible. In particular, let U be a uniform random variable on $[0,1]$ and set

$$X(w) = \begin{cases} +1 & \text{if } U \le p(\sigma,w), \\ -1 & \text{if } U > p(\sigma,w) \end{cases} \quad \text{and} \quad Y(w) = \begin{cases} +1 & \text{if } U \le p(\tau,w), \\ -1 & \text{if } U > p(\tau,w). \end{cases}$$

Set $X(u) = \sigma(u)$ and $Y(u) = \tau(u)$ for $u \ne w$.

If $w = v$, then $\rho(X,Y) = 0$. If $w \notin \mathcal{N}(v) \cup \{v\}$, then $\rho(X,Y) = 1$. If $w \in \mathcal{N}(v)$ and $p(\sigma,w) < U \le p(\tau,w)$, then $\rho(X,Y) = 2$. Thus,

$$\mathbf{E}_{\sigma,\tau}(\rho(X,Y)) \le 1 - \frac{1}{n} + \frac{1}{n} \sum_{w \in \mathcal{N}(v)} [p(\tau,w) - p(\sigma,w)]. \tag{15.5}$$

Noting that $S(w,\tau) = S(w,\sigma) + 2 = S + 2$, we obtain

$$\begin{aligned} p(\tau,w) - p(\sigma,w) &= \frac{e^{\beta(S+2)}}{e^{\beta(S+2)} + e^{-\beta(S+2)}} - \frac{e^{\beta S}}{e^{\beta S} + e^{-\beta S}} \\ &= \frac{1}{2}\left[\tanh(\beta(S+2)) - \tanh(\beta S)\right]. \end{aligned} \tag{15.6}$$

Letting $\tilde{S} = S + 1$ in (15.6) and then applying (15.3) shows that

$$p(\tau,w) - p(\sigma,w) = \frac{1}{2}\left[\tanh(\beta(\tilde{S}+1)) - \tanh(\beta(\tilde{S}-1))\right] \le \tanh(\beta). \tag{15.7}$$

Using the above bound in inequality (15.5) shows that

$$\mathbf{E}_{\sigma,\tau}(\rho(X,Y)) \le 1 - \frac{[1 - \Delta\tanh(\beta)]}{n} \le \exp\left(-\frac{1 - \Delta\tanh(\beta)}{n}\right) = e^{-c(\beta)/n}.$$

If $\Delta\tanh(\beta) < 1$, then $c(\beta) > 0$. Observe that $\operatorname{diam}(\Omega) = n$. Applying Corollary 14.7 with $\alpha = c(\beta)/n$ establishes (15.1).

Since $\tanh(x) \le x$, if $\beta < \Delta^{-1}$, then $\Delta\tanh(\beta) < 1$.

Proof of (ii). Note that if every vertex in the graph has even degree, then $\tilde{S} = S + 1$ takes on only odd values. Applying (15.4) shows that

$$p(\tau,w) - p(\sigma,w) = \frac{1}{2}\left[\tanh(\beta(\tilde{S}+1)) - \tanh(\beta(\tilde{S}-1))\right] \le \frac{\tanh(2\beta)}{2}.$$

Using the above bound in inequality (15.5) shows that

$$\mathbf{E}_{\sigma,\tau}(\rho(X,Y)) \le 1 - \frac{1 - (\Delta/2)\tanh(2\beta)}{n} \le e^{-c_e(\beta)/n}.$$

If $(\Delta/2)\tanh(2\beta) < 1$, then we can apply Corollary 14.7 to obtain (15.2). ■

15.2. The Complete Graph

Let G be the complete graph on n vertices, the graph which includes all $\binom{n}{2}$ possible edges. Since the interaction term $\sigma(v)\sum_{w\,:\,w\sim v}\sigma(w)$ is of order n, we take $\beta = \alpha/n$ so that the total contribution of a single site to $\beta\sum\sigma(v)\sigma(w)$ is $O(1)$.

THEOREM 15.3. *Let G be the complete graph on n vertices, and consider Glauber dynamics for the Ising model on G with $\beta = \alpha/n$.*

(i) *If $\alpha < 1$, then*

$$t_{\text{mix}}(\varepsilon) \le \frac{n(\log n + \log(1/\varepsilon))}{1 - \alpha}. \tag{15.8}$$

(ii) *If $\alpha > 1$, then there is a positive function $r(\alpha)$ so that $t_{\text{mix}} \ge O\left(\exp\left[r(\alpha)n\right]\right)$.*

PROOF. *Proof of* (i). Note that $\Delta\tanh(\beta) = (n-1)\tanh(\alpha/n) \le \alpha$. Thus if $\alpha < 1$, then Theorem 15.1(i) establishes (15.8).

Proof of (ii). Define $A_k := \{\sigma : |\{v : \sigma(v) = 1\}| = k\}$. By counting, $\pi(A_k) = a_k/Z(\alpha)$, where

$$a_k := \binom{n}{k} \exp\left\{\frac{\alpha}{n}\left[\binom{k}{2} + \binom{n-k}{2} - k(n-k)\right]\right\}.$$

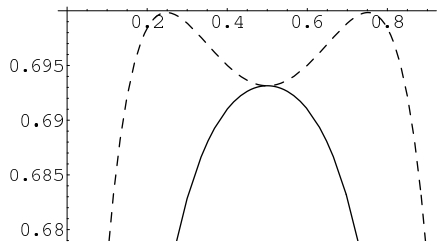


FIGURE 15.1. The function φ_α defined in (15.9). The dashed graph corresponds to $\alpha = 1.1$, the solid line to $\alpha = 0.9$.

Taking logarithms and applying Stirling's formula shows that

$$\log(a_{\lfloor cn \rfloor}) = n\varphi_\alpha(c)[1 + o(1)],$$

where

$$\varphi_\alpha(c) := -c\log(c) - (1-c)\log(1-c) + \alpha\left[\frac{(1-2c)^2}{2}\right]. \tag{15.9}$$

Taking derivatives shows that

$$\varphi'_\alpha(1/2) = 0,$$
$$\varphi''_\alpha(1/2) = -4(1-\alpha).$$

Hence $c = 1/2$ is a critical point of φ_α, and in particular it is a local maximum or minimum depending on the value of α. See Figure 15.1 for the graph of φ_α for $\alpha = 0.9$ and $\alpha = 1.1$. Take $\alpha > 1$, in which case φ_α has a local minimum at $1/2$. Define

$$S = \left\{\sigma : \sum_{u\in V}\sigma(u) < 0\right\}.$$

By symmetry, $\pi(S) \le 1/2$. Observe that the only way to get from S to S^c is through $A_{\lfloor n/2 \rfloor}$, since we are only allowed to change one spin at a time. Thus

$$Q(S, S^c) \le \frac{\lceil (n/2) \rceil}{n}\pi(A_{\lceil n/2 \rceil}) \quad \text{and} \quad \pi(S) = \sum_{j < \lceil n/2 \rceil}\pi(A_j).$$

Let c_1 be the value of c maximizing φ_α over $[0, 1/2]$. Since $1/2$ is a strict local minimum, $c_1 < 1/2$. Therefore,

$$\Phi(S) \le \frac{\exp\{\varphi_\alpha(1/2)n[1+o(1)]\}}{Z(\alpha)\pi(A_{\lfloor c_1 n \rfloor})} = \frac{\exp\{\varphi_\alpha(1/2)n[1+o(1)]\}}{\exp\{\varphi_\alpha(c_1)n[1+o(1)]\}}.$$

Since $\varphi_\alpha(c_1) > \varphi_\alpha(1/2)$, there is an $r(\alpha) > 0$ and constant $b > 0$ so that $\Phi_\star \le be^{-nr(\alpha)}$. The conclusion follows from Theorem 7.3. ∎

15.3. The Cycle

THEOREM 15.4. *Let $c_{\mathrm{O}}(\beta) := 1 - \tanh(2\beta)$. The Glauber dynamics for the Ising model on the n-cycle satisfies, for any $\beta > 0$ and fixed $\varepsilon > 0$,*

$$\frac{1+o(1)}{2c_{\mathrm{O}}(\beta)} \le \frac{t_{\mathrm{mix}}(\varepsilon)}{n\log n} \le \frac{1+o(1)}{c_{\mathrm{O}}(\beta)}. \tag{15.10}$$

CHAPTER 18

The Cutoff Phenomenon

18.1. Definition

For the top-to-random shuffle on n cards, we obtained in Section 6.5.3 the bound

$$d_n(n\log n + \alpha n) \le e^{-\alpha}, \tag{18.1}$$

while in Section 7.4.2 we showed that

$$\liminf_{n\to\infty} d_n(n\log n - \alpha n) \ge 1 - 2e^{2-\alpha}. \tag{18.2}$$

In particular, the upper bound in (18.1) tends to 0 as $\alpha \to \infty$, and the lower bound in (18.2) tends to 1 as $\alpha \to \infty$. It follows that $t_{\text{mix}}(\varepsilon) = n\log n\,[1 + h(n,\varepsilon)]$, where $\lim_{n\to\infty} h(n,\varepsilon) = 0$ for all ε. This is a much more precise statement than the fact that the mixing time is of the order $n\log n$.

The previous example motivates the following definition. Suppose, for a sequence of Markov chains indexed by $n = 1, 2, \ldots$, the mixing time for the n-th chain is denoted by $t_{\text{mix}}^{(n)}(\varepsilon)$. This sequence of chains has a ***cutoff*** if, for all $\varepsilon > 0$,

$$\lim_{n\to\infty} \frac{t_{\text{mix}}^{(n)}(\varepsilon)}{t_{\text{mix}}^{(n)}(1-\varepsilon)} = 1. \tag{18.3}$$

The bounds (18.1) and (18.2) for the top-to-random chain show that the total variation distance d_n for the n-card chain "falls off a cliff" at $t_{\text{mix}}^{(n)}$. More precisely, when time is rescaled by $n\log n$, as $n \to \infty$ the function d_n approaches a step function:

$$\lim_{n\to\infty} d_n(cn\log n) = \begin{cases} 1 & \text{if } c < 1, \\ 0 & \text{if } c > 1. \end{cases} \tag{18.4}$$

In fact, this property characterizes when a sequence of chains has a cutoff.

LEMMA 18.1. *Let $t_{\text{mix}}^{(n)}$ and d_n be the mixing time and distance to stationarity, respectively, for the n-th chain in a sequence of Markov chains. The sequence has a cutoff if and only if*

$$\lim_{n\to\infty} d_n(ct_{\text{mix}}^{(n)}) = \begin{cases} 1 & \text{if } c < 1, \\ 0 & \text{if } c > 1. \end{cases}$$

The proof is left to the reader as Exercise 18.1.

Returning again to the example of the top-to-random shuffle on n cards, the bounds (18.1) and (18.2) show that in an interval of length αn centered at $n\log n$, the total variation distance decreased from near 1 to near 0. The next definition formalizes this property.

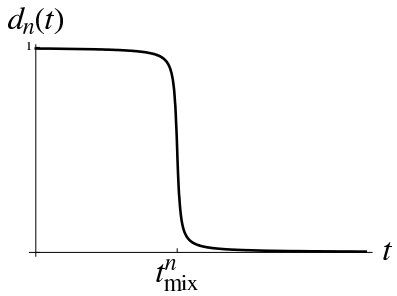


FIGURE 18.1. For a chain with a cutoff, the graph of $d_n(t)$ against t, when viewed on the time-scale of $t^{(n)}_{\text{mix}}$, approaches a step function as $n \to \infty$.

A sequence of Markov chains has a cutoff with a ***window*** of size $\{w_n\}$ if $w_n = o\left(t^{(n)}_{\text{mix}}\right)$ and

$$\lim_{\alpha\to-\infty} \liminf_{n\to\infty} d_n\left(t^{(n)}_{\text{mix}} + \alpha w_n\right) = 1,$$
$$\lim_{\alpha\to\infty} \limsup_{n\to\infty} d_n\left(t^{(n)}_{\text{mix}} + \alpha w_n\right) = 0.$$

We say a family of chains has a ***pre-cutoff*** if it satisfies the weaker condition

$$\sup_{0<\varepsilon<1/2} \limsup_{n\to\infty} \frac{t^{(n)}_{\text{mix}}(\varepsilon)}{t^{(n)}_{\text{mix}}(1-\varepsilon)} < \infty.$$

Theorem 15.4 proved that the Glauber dynamics for the Ising model on the n-cycle has a pre-cutoff; it is an open problem to show that in fact this family of chains has a cutoff.

18.2. Examples of Cutoff

18.2.1. Biased random walk on a line segment. Let $p \in (1/2, 1)$ and $q = 1 - p$, so $\beta := (p - q)/2 = p - 1/2 > 0$. Consider the lazy nearest-neighbor random walk with bias β on the interval $\Omega = \{0, 1, \ldots, n\}$, which is the Markov chain with transition probabilities

$$P(k, k+1) = \begin{cases} \frac{p}{2} & \text{if } k \notin \{0, n\}, \\ \frac{1}{2} & \text{if } k = 0, \\ 0 & \text{if } k = n, \end{cases}$$
$$P(k, k) = \frac{1}{2},$$
$$P(k, k-1) = \begin{cases} \frac{q}{2} & \text{if } k \notin \{0, n\}, \\ 0 & \text{if } k = 0, \\ \frac{1}{2} & \text{if } k = n. \end{cases}$$

That is, when at an interior vertex, the walk remains in its current position with probability $1/2$, moves to the right with probability $p/2$, and moves to the left with probability $q/2$. When at an end-vertex, the walk remains in place with probability $1/2$ and moves to the adjacent interior vertex with probability $1/2$.

THEOREM 18.2. *The lazy random walk with bias $\beta = p - 1/2$ on $\{0, 1, 2, \ldots, n\}$ has a cutoff at $\beta^{-1}n$ with a window of order $\sqrt{n}$.*

PROOF. We write $t_n(\alpha) := \beta^{-1}n + \alpha\sqrt{n}$.

Upper bound, Step 1. We first prove that if $\tau_n := \min\{t \geq 0 : X_t = n\}$, then

$$\limsup_{n\to\infty} \mathbf{P}_0\{\tau_n > t_n(\alpha)\} \leq \Phi(-c(\beta)\alpha), \tag{18.5}$$

where $c(\beta)$ depends on β only and Φ is the standard normal distribution function.

Let (S_t) be a lazy β-biased nearest-neighbor random walk on all of $\mathbb{Z}$, so $\mathbf{E}_k S_t = k + \beta t$. We couple (X_t) to (S_t) until time $\tau_n := \min\{t \geq 0 : X_t = n\}$, as follows: let $X_0 = S_0$, and set

$$X_{t+1} = \begin{cases} 1 & \text{if } X_t = 0 \text{ and } S_{t+1} - S_t = -1, \\ X_t + (S_{t+1} - S_t) & \text{otherwise.} \end{cases} \tag{18.6}$$

This coupling satisfies $X_t \geq S_t$ for all $t \leq \tau_n$.

We have $\mathbf{E}_0 S_{t_n(\alpha)} = t_n(\alpha)\beta = n + \alpha\beta\sqrt{n}$, and

$$\mathbf{P}_0\{S_{t_n(\alpha)} < n\} = \mathbf{P}_0\left\{\frac{S_{t_n(\alpha)} - \mathbf{E}S_{t_n(\alpha)}}{\sqrt{t_n(\alpha)v}} < \frac{-\alpha\beta\sqrt{n}}{\sqrt{t_n(\alpha)v}}\right\},$$

where $v = 1/2 - \beta^2$. By the Central Limit Theorem, the right-hand side above converges as $n \to \infty$ to $\Phi(-c(\beta)\alpha)$. Thus

$$\limsup_{n\to\infty} \mathbf{P}_0\{S_{t_n(\alpha)} < n\} = \Phi(-c(\beta)\alpha). \tag{18.7}$$

Since $X_t \geq S_t$ for $t \leq \tau_n$,

$$\mathbf{P}_0\{\tau_n > t_n(\alpha)\} \leq \mathbf{P}_0\left\{\max_{0\leq s\leq t_n(\alpha)} S_s < n\right\} \leq \mathbf{P}_0\left\{S_{t_n(\alpha)} \leq n\right\},$$

which with (18.7) implies (18.5).

Upper bound, Step 2. We now show that we can couple two biased random walks so that the meeting time of the two walks is bounded by τ_n.

We couple as follows: toss a coin to decide which particle to move. Move the chosen particle up one unit with probability p and down one unit with probability q, unless it is at an end-vertex, in which case move it with probability one to the neighboring interior vertex. The time τ_{couple} until the particles meet is bounded by the time it takes the left-most particle to hit n, whence

$$d_n(t_n(\alpha)) \leq \mathbf{P}_{x,y}\{\tau_{\text{couple}} > t_n(\alpha)\} \leq \mathbf{P}_0\{\tau_n > t_n(\alpha)\}.$$

This bound and (18.5) show that

$$\lim_{\alpha\to\infty} \limsup_{n\to\infty} d(t_n(\alpha)) \leq \lim_{\alpha\to\infty} \Phi(-c(\beta)\alpha) = 0.$$

Lower bound, Step 1. Let $\theta := (q/p)$. We first prove that

$$\limsup_{n\to\infty} \mathbf{P}_0\{X_{t_n(\alpha)} > n - h\} \leq 1 - \Phi(-c(\beta)\alpha) + \theta^{h-1}. \tag{18.8}$$

Let $(\tilde{X}_t)$ be the lazy biased random walk on $\{0, 1, \ldots\}$, with reflection at 0. By coupling with (X_t) so that $X_t \leq \tilde{X}_t$, for $x \geq 0$ we have

$$\mathbf{P}_0\{X_t > x\} \leq \mathbf{P}_0\{\tilde{X}_t > x\}. \tag{18.9}$$

Recall that (S_t) is the biased lazy walk on all of $\mathbb{Z}$. Couple $(\tilde{X}_t)$ with (S_t) so that $S_t \le \tilde{X}_t$. Observe that $\tilde{X}_t - S_t$ increases (by a unit amount) only when $\tilde{X}_t$ is at 0, which implies that, for any t,

$$\mathbf{P}_0\{\tilde{X}_t - S_t \ge h\} \le \mathbf{P}_0\{\text{at least } h-1 \text{ returns of } (\tilde{X}_t) \text{ to } 0\}.$$

By (9.21), the chance that the biased random walk on $\mathbb{Z}$, when starting from 1, hits 0 before n equals $1-(1-\theta)/(1-\theta^n)$. Letting $n\to\infty$, the chance that the biased random walk on $\mathbb{Z}$, when starting from 1, ever visits 0 equals θ. Therefore,

$$\mathbf{P}_0\{\text{at least } h-1 \text{ returns of } (\tilde{X}_t) \text{ to } 0\} = \theta^{h-1},$$

and consequently,

$$\mathbf{P}_0\{\tilde{X}_t - S_t \ge h\} \le \theta^{h-1}. \tag{18.10}$$

By (18.9) and (18.10),

$$\begin{aligned}\mathbf{P}_0\{X_{t_n(\alpha)} > n-h\} &\le \mathbf{P}_0\{S_{t_n(\alpha)} > n-2h\} + \mathbf{P}_0\{\tilde{X}_{t_n(\alpha)} - S_{t_n(\alpha)} \ge h\}\\ &\le \mathbf{P}_0\{S_{t_n(\alpha)} > n-2h\} + \theta^{h-1}. \end{aligned}\tag{18.11}$$

By the Central Limit Theorem,

$$\lim_{n\to\infty} \mathbf{P}_0\{S_{t_n(\alpha)} > n-2h\} = 1-\Phi(-c(\beta)\alpha),$$

which together with (18.11) establishes (18.8).

Lower bound, Step 2. The stationary distribution equals

$$\pi^{(n)}(k) = \left[\frac{(p/q)-1}{(p/q)^{n+1}-1}\right](p/q)^k.$$

If $A_h = \{n-h+1,\dots,n\}$, then

$$\pi^{(n)}(A_h) = \frac{1-(q/p)^{h+2}}{1-(q/p)^{n+1}}.$$

Therefore,

$$\begin{aligned}\liminf_{n\to\infty} d_n(t_n(\alpha)) &\ge \liminf_{n\to\infty}\left[\pi^{(n)}(A_h) - \mathbf{P}_0\{X_{t_n(\alpha)} > n-h\}\right]\\ &\ge 1-\theta^{h+2} - \left[1-\Phi(-c(\beta)\alpha) + \theta^{h-1}\right],\end{aligned}$$

and so

$$\lim_{\alpha\to-\infty}\liminf_{n\to\infty} d_n(t_n(\alpha)) \ge 1-\theta^{h+2}-\theta^{h-1}.$$

Letting $h\to\infty$ shows that

$$\lim_{\alpha\to-\infty}\liminf_{n\to\infty} d_n(t_n(\alpha)) = 1.$$

■

18.2.2. Random walk on the hypercube. We return to the lazy random walk on the n-dimensional hypercube. In Section 5.3.3, it was shown that

$$t_{\text{mix}}(\varepsilon) \le n\log n + c_u(\varepsilon)n,$$

while Proposition 7.13 proved that

$$t_{\text{mix}}(1-\varepsilon) \ge \frac{1}{2}n\log n - c_\ell(\varepsilon)n. \tag{18.12}$$

In fact, there is a cutoff, and the lower bound gives the correct constant:

THEOREM 18.3. *The lazy random walk on the n-dimensional hypercube has a cutoff at $(1/2)n\log n$ with a window of size n.*

PROOF. Let $\boldsymbol{X}_t = (X_t^1, \dots, X_t^n)$ be the position of the random walk at time t, and let $W_t = W(\boldsymbol{X}_t) = \sum_{i=1}^n X_t^i$ be the Hamming weight of $\boldsymbol{X}_t$. As follows from the discussion in Section 2.3, (W_t) is a lazy version of the Ehrenfest urn chain whose transition matrix is given in (2.8). We write π_W for the stationary distribution of (W_t), which is binomial with parameters n and $1/2$.

The study of $(\boldsymbol{X}_t)$ can be reduced to the study of (W_t) because of the following identity:

$$\|\mathbf{P}_{\mathbf{1}}\{\boldsymbol{X}_t \in \cdot\} - \pi\|_{\mathrm{TV}} = \|\mathbf{P}_n\{W_t \in \cdot\} - \pi_W\|_{\mathrm{TV}}\,. \tag{18.13}$$

Proof of (18.13). Let $\Omega_w := \{\boldsymbol{x} : W(\boldsymbol{x}) = w\}$. Note that by symmetry, the functions $\boldsymbol{x} \mapsto \mathbf{P}_{\mathbf{1}}\{\boldsymbol{X}_t = \boldsymbol{x}\}$ and π are constant over Ω_w. Therefore,

$$\begin{aligned}\sum_{\boldsymbol{x}\,:\,W(\boldsymbol{x})=w} |\mathbf{P}_1\{\boldsymbol{X}_t = \boldsymbol{x}\} - \pi(\boldsymbol{x})| &= \left|\sum_{\boldsymbol{x}\,:\,W(\boldsymbol{x})=w} \mathbf{P}_1\{\boldsymbol{X}_t = \boldsymbol{x}\} - \pi(\boldsymbol{x})\right| \\ &= |\mathbf{P}_1\{W_t = w\} - \pi_W(w)|\,.\end{aligned}$$

(The absolute values can be moved outside the sum in the first equality because all of the terms in the sum are equal.) Summing over $w \in \{0, 1, \dots, n\}$ and dividing by 2 yields (18.13).

Since $(\boldsymbol{X}_t)$ is a transitive chain,

$$d(t) = \|\mathbf{P}_{\mathbf{1}}\{\boldsymbol{X}_t \in \cdot\} - \pi\|_{\mathrm{TV}}\,,$$

and it is enough to bound the right-hand side of (18.13).

We construct now a coupling (W_t, Z_t) of the lazy Ehrenfest chain started from w with the lazy Ehrenfest chain started from z. Provided that the two chains have not yet collided, at each move, a fair coin is tossed to determine which of the two chains moves; the chosen chain makes a transition according to the matrix (2.8), while the other chain remains in its current position. The chains move together once they have met for the first time.

Suppose, without loss of generality, that $z \geq w$. Since the chains never cross each other, $Z_t \geq W_t$ for all t. Consequently, if $D_t = |Z_t - W_t|$, then $D_t = Z_t - W_t \geq 0$. Let $\tau := \min\{t \geq 0 : Z_t = W_t\}$. Supposing that $(Z_t, W_t) = (z_t, w_t)$ and $\tau > t$,

$$D_{t+1} - D_t = \begin{cases} 1 & \text{with probability } (1/2)(1 - z_t/n) + (1/2)w_t/n, \\ -1 & \text{with probability } (1/2)z_t/n + (1/2)(1 - w_t/n). \end{cases} \tag{18.14}$$

From (18.14) we see that on the event $\{\tau > t\}$,

$$\mathbf{E}_{z,w}[D_{t+1} - D_t \mid Z_t = z_t, W_t = w_t] = -\frac{(z_t - w_t)}{n}. \tag{18.15}$$

Let $\boldsymbol{Z}_t = (Z_1, \dots, Z_t)$ and $\boldsymbol{W}_t = (W_1, \dots, W_t)$. By the Markov property and because $\mathbf{1}\{\tau > t\}$ is a function of $(\boldsymbol{Z}_t, \boldsymbol{W}_t)$,

$$\begin{aligned}&\mathbf{1}\{\tau > t\}\mathbf{E}_{z,w}[D_{t+1} - D_t \mid Z_t, W_t] \\ &\qquad = \mathbf{1}\{\tau > t\}\mathbf{E}_{z,w}[D_{t+1} - D_t \mid \boldsymbol{Z}_t, \boldsymbol{W}_t] \\ &\qquad = \mathbf{E}_{z,w}[\mathbf{1}\{\tau > t\}(D_{t+1} - D_t) \mid \boldsymbol{Z}_t, \boldsymbol{W}_t].\end{aligned} \tag{18.16}$$

Combining (18.15) and (18.16) shows that

$$\mathbf{E}_{z,w}[\mathbf{1}\{\tau > t\}D_{t+1} \mid \boldsymbol{Z}_t, \boldsymbol{W}_t] \le \left(1 - \frac{1}{n}\right) D_t \mathbf{1}\{\tau > t\}.$$

Taking expectation, we have

$$\mathbf{E}_{z,w}[D_{t+1}\mathbf{1}\{\tau > t\}] = \left(1 - \frac{1}{n}\right) \mathbf{E}_{z,w}[D_t \mathbf{1}\{\tau > t\}].$$

Since $\mathbf{1}\{\tau > t+1\} \le \mathbf{1}\{\tau > t\}$, we have

$$\mathbf{E}_{z,w}[D_{t+1}\mathbf{1}\{\tau > t+1\}] \le \left(1 - \frac{1}{n}\right) \mathbf{E}_{z,w}[D_t \mathbf{1}\{\tau > t\}].$$

By induction,

$$\mathbf{E}_{z,w}[D_t \mathbf{1}\{\tau > t\}] \le \left(1 - \frac{1}{n}\right)^t (z - w) \le n e^{-t/n}. \tag{18.17}$$

Also, from (18.14), provided $\tau > t$, the process (D_t) is at least as likely to move downwards as it is to move upwards. Thus, until time τ, the process (D_t) can be coupled with a simple random walk (S_t) so that $S_0 = D_0$ and $D_t \le S_t$.

If $\tilde{\tau} := \min\{t \ge 0 : S_t = 0\}$, then $\tau \le \tilde{\tau}$. By Theorem 2.26, there is a constant c_1 such that for $k \ge 0$,

$$\mathbf{P}_k\{\tau > u\} \le \mathbf{P}_k\{\tilde{\tau} > u\} \le \frac{c_1 k}{\sqrt{u}}. \tag{18.18}$$

By (18.18),

$$\mathbf{P}_{z,w}\{\tau > s + u \mid D_0, D_1, \dots D_s\} = \mathbf{1}\{\tau > s\}\mathbf{P}_{D_s}\{\tau > u\} \le \frac{c_1 D_s \mathbf{1}\{\tau > s\}}{\sqrt{u}}.$$

Taking expectation above and applying (18.17) shows that

$$\mathbf{P}_{z,w}\{\tau > s + u\} \le \frac{c_1 n e^{-s/n}}{\sqrt{u}}. \tag{18.19}$$

Letting $u = \alpha n$ and $s = (1/2)n \log n$ above, by Corollary 5.3 we have

$$d((1/2)n \log n + \alpha n) \le \frac{c_1}{\sqrt{\alpha}}.$$

We conclude that

$$\lim_{\alpha \to \infty} \limsup_{n \to \infty} d((1/2)n \log n + \alpha n) = 0.$$

The lower bound (7.26) completes the proof. ∎

18.3. A Necessary Condition for Cutoff

When does a family of chains have a cutoff? The following proposition gives a necessary condition.

PROPOSITION 18.4. *For a sequence of irreducible aperiodic Markov chains with relaxation times $\{t_{\text{rel}}^{(n)}\}$ and mixing times $\{t_{\text{mix}}^{(n)}\}$, if $t_{\text{mix}}^{(n)}/t_{\text{rel}}^{(n)}$ is bounded above, then there is no pre-cutoff.*

PROOF. The proof follows from Theorem 12.4: dividing both sides of (12.12) by $t_{\text{mix}}^{(n)}$, we have

$$\frac{t_{\text{mix}}^{(n)}(\varepsilon)}{t_{\text{mix}}^{(n)}} \geq \frac{t_{\text{rel}}^{(n)} - 1}{t_{\text{mix}}^{(n)}} \log\left(\frac{1}{2\varepsilon}\right) \geq c_1 \log\left(\frac{1}{2\varepsilon}\right).$$

As $\varepsilon \to 0$, the right-hand side increases to infinity. ■

Recall that we write $a_n \asymp b_n$ to mean that there exist positive and finite constants c_1 and c_2, not depending on n, such that $c_1 \leq a_n/b_n \leq c_2$ for all n.

EXAMPLE 18.5. Consider the lazy random walk on the cycle $\mathbb{Z}_n$. In Section 5.3.1 we showed that $t_{\text{mix}}^{(n)} \leq n^2$. In fact, this is the correct order, as shown in Section 7.4.1. In Section 12.3.1, we computed the eigenvalues of the transition matrix, finding that $t_{\text{rel}}^{(n)} \asymp n^2$ also. By Proposition 18.4, there is no pre-cutoff.

EXAMPLE 18.6. Let T_n be the rooted binary tree with n vertices. In Example 7.7, we showed that the lazy simple random walk has $t_{\text{mix}} \asymp n$. Together with Theorem 12.4, this implies that there exists a constant c_1 such that $t_{\text{rel}} \leq c_1 n$. In Example 7.7, we actually showed that $\Phi_\star \leq 1/(n-2)$. Thus, by Theorem 13.14, we have $\gamma \leq 2/(n-2)$, whence $t_{\text{rel}} \geq c_2 n$ for some constant c_2. An application of Proposition 18.4 shows that there is no pre-cutoff for this family of chains.

The question remains if there are conditions which ensure that the converse of Proposition 18.4 holds. Below we give a variant of an example due to Igor Pak (personal communication) which shows the converse is not true in general.

EXAMPLE 18.7. Let $\{P_n\}$ be a family of transition matrices with $t_{\text{rel}}^{(n)} = o(t_{\text{mix}}^{(n)})$ and with a cutoff (e.g., the lazy random walk on the hypercube.) Let $L_n := \sqrt{t_{\text{rel}}^{(n)} t_{\text{mix}}^{(n)}}$, and define the matrix

$$\tilde{P}_n = (1 - 1/L_n)P_n + (1/L)\Pi_n,$$

where $\Pi_n(x, y) := \pi_n(y)$ for all x.

We first prove that

$$\left\|\tilde{P}_n^t(x,\cdot) - \pi\right\|_{\text{TV}} = \left(1 - \frac{1}{L_n}\right)^t \left\|P_n^t(x,\cdot) - \pi\right\|_{\text{TV}}. \tag{18.20}$$

Proof of (18.20). One step of the chain can be generated by first tossing a coin with probability $1/L_n$ of heads; if heads, a sample from π_n is produced, and if tails, a transition from P_n is used. If τ is the first time that the coin lands heads, then τ has a geometric distribution with success probability $1/L_n$. Accordingly,

$$\begin{aligned}
\mathbf{P}_x\{X_t^{(n)} = y\} - \pi(y) &= \mathbf{P}_x\{X_t^{(n)} = y, \tau \leq t\} + \mathbf{P}_x\{X_t^{(n)} = y, \tau > t\} - \pi(y) \\
&= -\pi(y)[1 - \mathbf{P}_x\{\tau \leq t\}] + P_n^t(x,y)\mathbf{P}_x\{\tau > t\} \\
&= \left[P_n^t(x,y) - \pi_n(y)\right]\mathbf{P}_x\{\tau > t\}.
\end{aligned}$$

Taking absolute value and summing over y gives (18.20). We conclude that

$$\tilde{d}_n(t) = (1 - L_n^{-1})^t d_n(t).$$

Therefore,

$$\tilde{d}_n(\beta L_n) \leq e^{-\beta} d_n(\beta L_n) \leq e^{-\beta},$$