

## 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  $\Omega$  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  $(X_0, X_1, \dots)$  is a **Markov chain with state space  $\Omega$  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

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

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  $x_0, x_1, \dots, x_{t-1}$  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

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

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  $(X_0, X_1, \dots)$  be the sequence of lily pads occupied by the frog on Sunday, Monday,  $\dots$ . 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

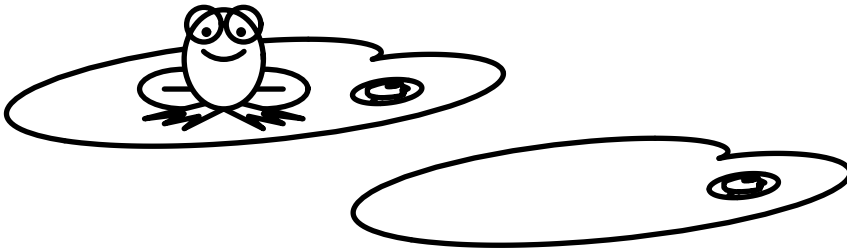


FIGURE 1.1. A randomly jumping frog. Whenever he tosses heads, he jumps to the other lily pad.

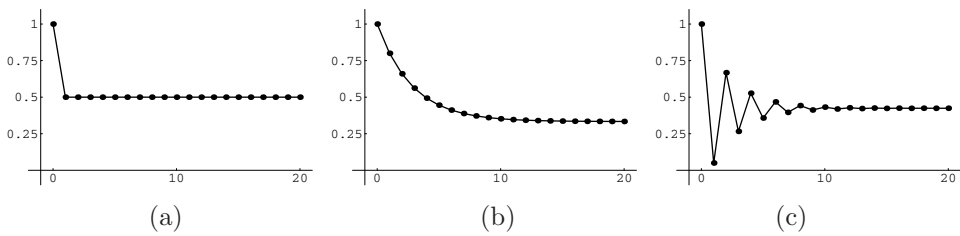


FIGURE 1.2. The probability of being on the east pad (started from the east pad) plotted versus time for (a)  $p = q = 1/2$ , (b)  $p = 0.2$  and  $q = 0.1$ , (c)  $p = 0.95$  and  $q = 0.7$ . The long-term limiting probabilities are  $1/2$ ,  $1/3$ , and  $14/33 \approx 0.42$ , respectively.

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

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

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,

$$\mathbf{P}\{X_1 = e \mid X_0 = e\} = 1-p, \quad \mathbf{P}\{X_1 = w \mid X_0 = e\} = p. \quad (1.3)$$

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

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

and

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

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

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

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:

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

Indeed, for any initial distribution  $\mu_0$ ,

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

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

$$\pi = \pi P,$$

which implies (after a little algebra) that

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

If we define

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

then by the definition of  $\mu_{t+1}$  the sequence  $(\Delta_t)$  satisfies

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

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

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

for any initial distribution  $\mu_0$ . As we suspected,  $\mu_t$  approaches  $\pi$  as  $t \rightarrow \infty$ .

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  $t \rightarrow \infty$  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  $\Omega$  and transition matrix  $P$ , and let the row vector  $\mu_t$  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

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

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  $\delta_x$ :

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

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  $\mu$  on  $\Omega$  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  $\mu$  at time  $t$ , then it has distribution  $\mu P$  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  $\Omega$  (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  $\Omega = \mathbb{Z}_n = \{0, 1, \dots, n - 1\}$ , the set of remainders modulo  $n$ . Consider the transition matrix

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

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  $\Omega$  is a function  $f : \Omega \times \Lambda \rightarrow \Omega$ , along with a  $\Lambda$ -valued random variable  $Z$ , satisfying

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

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

For the example of the simple random walk on the cycle, setting  $\Lambda = \{1, -1\}$ , each  $Z_i$  uniform on  $\Lambda$ , 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  $\Omega = \{x_1, \dots, x_n\}$ . Take  $\Lambda = [0, 1]$ ; our auxiliary random variables  $Z, Z_1, Z_2, \dots$  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  $\ell$  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  $\Omega$  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  $t \rightarrow \infty$ .

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

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

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

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

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

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

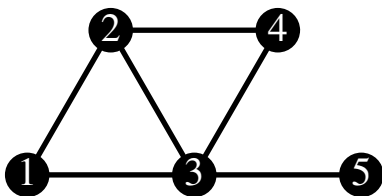


FIGURE 1.4. An example of a graph with vertex set  $\{1, 2, 3, 4, 5\}$  and 6 edges.

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  $\pi$  on  $\Omega$  satisfying

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

can have another interesting property: in that case,  $\pi$  was the long-term limiting distribution of the chain. We call a probability  $\pi$  satisfying (1.14) a *stationary distribution* of the Markov chain. Clearly, if  $\pi$  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,

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

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  $\Omega$  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  $\tau_x^+$  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  $\varepsilon$ . Hence for  $k > 0$  we have

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

Repeated application of (1.17) yields

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

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  $\pi$  on  $\Omega$  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  $\pi$  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

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

For any state  $y$ , we have  $\tilde{\pi}(y) \leq \mathbf{E}_z \tau_z^+$ . 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:

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

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

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

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

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

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

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

$$= \tilde{\pi}(y). \tag{1.24}$$

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  $\sum_x \tilde{\pi}(x) = \mathbf{E}_z(\tau_z^+)$ :

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

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

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

■

The computation at the heart of the proof of Proposition 1.14 can be generalized. A **stopping time**  $\tau$  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  $\tau$  replaces  $\tau_z^+$  in the definition (1.19) of  $\tilde{\pi}$ , then the proof that  $\tilde{\pi}$  satisfies  $\tilde{\pi} = \tilde{\pi}P$  works, provided that  $\tau$  satisfies both  $\mathbf{P}_z\{\tau < \infty\} = 1$  and  $\mathbf{P}_z\{X_\tau = z\} = 1$ .

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

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

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  $h : \Omega \rightarrow \mathbb{R}$  **harmonic at  $x$**  if

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

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  $\Omega$  satisfies the matrix equation  $Ph = h$ .

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

PROOF. Since  $\Omega$  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

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

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  $x_0, x_1, \dots, x_n = y$  with  $P(x_i, x_{i+1}) > 0$ . Repeating the argument above tells us that  $h(y) = h(x_{n-1}) = \dots = h(x_0) = M$ . Thus  $h$  is constant. ■

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

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

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  $\Omega$ . Any distribution  $\pi$  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,

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

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

$$\mathbf{P}_\pi\{X_0 = x_0, \dots, X_n = x_n\} = \mathbf{P}_\pi\{X_0 = x_n, X_1 = x_{n-1}, \dots, X_n = x_0\}. \quad (1.32)$$

In other words, if a chain  $(X_t)$  satisfies (1.30) and has stationary initial distribution, then the distribution of  $(X_0, X_1, \dots, X_n)$  is the same as the distribution of  $(X_n, X_{n-1}, \dots, X_0)$ . 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  $\pi$  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  $\pi$  is the chain with matrix

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

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  $\pi$ . Write  $(\widehat{X}_t)$  for the time-reversed chain with transition matrix  $\widehat{P}$ . Then  $\pi$  is stationary for  $\widehat{P}$ , and for any  $x_0, \dots, x_t \in \Omega$  we have*

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

PROOF. To check that  $\pi$  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

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

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  $\widehat{P} = P$ .

### 1.7. Classifying the States of a Markov Chain\*

We will occasionally need to study chains which are *not* irreducible—see, for instance, Sections 2.1, 2.2 and 2.4. In this section we describe a way to classify the states of a Markov chain. This classification clarifies what can occur when irreducibility fails.

Let  $P$  be the transition matrix of a Markov chain on a finite state space  $\Omega$ . Given  $x, y \in \Omega$ , we say that  $y$  is **accessible from**  $x$  and write  $x \rightarrow y$  if there exists an  $r > 0$  such that  $P^r(x, y) > 0$ . That is,  $x \rightarrow y$  if it is possible for the chain to move from  $x$  to  $y$  in a finite number of steps. Note that if  $x \rightarrow y$  and  $y \rightarrow z$ , then  $x \rightarrow z$ .

A state  $x \in \Omega$  is called **essential** if for all  $y$  such that  $x \rightarrow y$  it is also true that  $y \rightarrow x$ . A state  $x \in \Omega$  is **inessential** if it is not essential.

We say that  $x$  **communicates with**  $y$  and write  $x \leftrightarrow y$  if and only if  $x \rightarrow y$  and  $y \rightarrow x$ . The equivalence classes under  $\leftrightarrow$  are called **communicating classes**. For  $x \in \Omega$ , the communicating class of  $x$  is denoted by  $[x]$ .

Observe that when  $P$  is irreducible, all the states of the chain lie in a single communicating class.

LEMMA 1.23. *If  $x$  is an essential state and  $x \rightarrow y$ , then  $y$  is essential.*

PROOF. If  $y \rightarrow z$ , then  $x \rightarrow z$ . Therefore, because  $x$  is essential,  $z \rightarrow x$ , whence  $z \rightarrow y$ . ■

It follows directly from the above lemma that the states in a single communicating class are either all essential or all inessential. We can therefore classify the communicating classes as either essential or inessential.

If  $[x] = \{x\}$  and  $x$  is inessential, then once the chain leaves  $x$ , it never returns. If  $[x] = \{x\}$  and  $x$  is essential, then the chain never leaves  $x$  once it first visits  $x$ ; such states are called **absorbing**.

LEMMA 1.24. *Every finite chain has at least one essential class.*

PROOF. Define inductively a sequence  $(y_0, y_1, \dots)$  as follows: Fix an arbitrary initial state  $y_0$ . For  $k \geq 1$ , given  $(y_0, \dots, y_{k-1})$ , if  $y_{k-1}$  is essential, stop. Otherwise, find  $y_k$  such that  $y_{k-1} \rightarrow y_k$  but  $y_k \not\rightarrow y_{k-1}$ .

There can be no repeated states in this sequence, because if  $j < k$  and  $y_k \rightarrow y_j$ , then  $y_k \rightarrow y_{k-1}$ , a contradiction.

Since the state space is finite and the sequence cannot repeat elements, it must eventually terminate in an essential state. ■

Note that a transition matrix  $P$  restricted to an essential class  $[x]$  is stochastic. That is,  $\sum_{y \in [x]} P(x, y) = 1$ , since  $P(x, z) = 0$  for  $z \notin [x]$ .

PROPOSITION 1.25. *If  $\pi$  is stationary for the finite transition matrix  $P$ , then  $\pi(y_0) = 0$  for all inessential states  $y_0$ .*

PROOF. Let  $\mathcal{C}$  be an essential communicating class. Then

$$\pi P(\mathcal{C}) = \sum_{z \in \mathcal{C}} (\pi P)(z) = \sum_{z \in \mathcal{C}} \left[ \sum_{y \in \mathcal{C}} \pi(y) P(y, z) + \sum_{y \notin \mathcal{C}} \pi(y) P(y, z) \right].$$

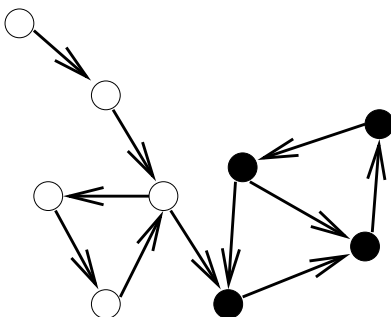


FIGURE 1.5. The directed graph associated to a Markov chain. A directed edge is placed between  $v$  and  $w$  if and only if  $P(v, w) > 0$ . Here there is one essential class, which consists of the filled vertices.

We can interchange the order of summation in the first sum, obtaining

$$\pi P(\mathcal{C}) = \sum_{y \in \mathcal{C}} \pi(y) \sum_{z \in \mathcal{C}} P(y, z) + \sum_{z \in \mathcal{C}} \sum_{y \notin \mathcal{C}} \pi(y) P(y, z).$$

For  $y \in \mathcal{C}$  we have  $\sum_{z \in \mathcal{C}} P(y, z) = 1$ , so

$$\pi P(\mathcal{C}) = \pi(\mathcal{C}) + \sum_{z \in \mathcal{C}} \sum_{y \notin \mathcal{C}} \pi(y) P(y, z). \quad (1.34)$$

Since  $\pi$  is invariant,  $\pi P(\mathcal{C}) = \pi(\mathcal{C})$ . In view of (1.34) we must have  $\pi(y)P(y, z) = 0$  for all  $y \notin \mathcal{C}$  and  $z \in \mathcal{C}$ .

Suppose that  $y_0$  is inessential. The proof of Lemma 1.24 shows that there is a sequence of states  $y_0, y_1, y_2, \dots, y_r$  satisfying  $P(y_{i-1}, y_i) > 0$ , the states  $y_0, y_1, \dots, y_{r-1}$  are inessential, and  $y_r \in \mathcal{C}$ , where  $\mathcal{C}$  is an essential communicating class. Since  $P(y_{r-1}, y_r) > 0$  and we just proved that  $\pi(y_{r-1})P(y_{r-1}, y_r) = 0$ , it follows that  $\pi(y_{r-1}) = 0$ . If  $\pi(y_k) = 0$ , then

$$0 = \pi(y_k) = \sum_{y \in \Omega} \pi(y) P(y, y_k).$$

This implies  $\pi(y)P(y, y_k) = 0$  for all  $y$ . In particular,  $\pi(y_{k-1}) = 0$ . By induction backwards along the sequence, we find that  $\pi(y_0) = 0$ . ■

Finally, we conclude with the following proposition:

**PROPOSITION 1.26.** *The stationary distribution  $\pi$  for a transition matrix  $P$  is unique if and only if there is a unique essential communicating class.*

**PROOF.** Suppose that there is a unique essential communicating class  $\mathcal{C}$ . We write  $P|_{\mathcal{C}}$  for the restriction of the matrix  $P$  to the states in  $\mathcal{C}$ . Suppose  $x \in \mathcal{C}$  and  $P(x, y) > 0$ . Then since  $x$  is essential and  $x \rightarrow y$ , it must be that  $y \rightarrow x$  also, whence  $y \in \mathcal{C}$ . This implies that  $P|_{\mathcal{C}}$  is a transition matrix, which clearly must be irreducible on  $\mathcal{C}$ . Therefore, there exists a unique stationary distribution  $\pi^{\mathcal{C}}$  for  $P|_{\mathcal{C}}$ . Let  $\pi$  be a probability on  $\Omega$  with  $\pi = \pi P$ . By Proposition 1.25,  $\pi(y) = 0$  for

$y \notin \mathcal{C}$ , whence  $\pi$  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  $\pi$  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  $\pi$  supported on  $\mathcal{C}_i$  which is stationary for  $P|_{\mathcal{C}_i}$ . Moreover, it is easily verified that each  $\pi_i$  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  $x_0, x_1, \dots, x_k$  such that  $x_0 = x$ ,  $x_k = y$ , and  $x_i \sim x_{i+1}$  for  $0 \leq i \leq k-1$ . 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  $\Omega$  can be partitioned into  $b$  sets  $\mathcal{C}_1, \mathcal{C}_2, \dots, \mathcal{C}_b$  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  $\Omega$  is stationary for  $P$ .

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

EXERCISE 1.9. Let  $\pi$  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  $\Omega$ . For an arbitrary initial distribution  $\mu$  on  $\Omega$  and  $n > 0$ , define the distribution  $\nu_n$  by

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

(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  $\lim_{k \rightarrow \infty} \nu_{n_k}(x)$  exists for every  $x \in \Omega$ .

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

EXERCISE 1.12. Let  $P$  be the transition matrix of an irreducible Markov chain with state space  $\Omega$ . Let  $B \subset \Omega$  be a non-empty subset of the state space, and assume  $h : \Omega \rightarrow \mathbb{R}$  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  $\pi_1$  and  $\pi_2$ , 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  $\pi$  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  $\Omega$ .

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.

## Notes

Markov first studied the stochastic processes that came to be named after him in Markov (1906). See Basharin, Langville, and Naumov (2004) for the early history of Markov chains.

The right-hand side of (1.1) does not depend on  $t$ . We take this as part of the definition of a Markov chain; note that other authors sometimes regard this as a special case, which they call *time homogeneous*. (This simply means that the transition matrix is the same at each step of the chain. It is possible to give a more general definition in which the transition matrix depends on  $t$ . We will not consider such chains in this book.)

Aldous and Fill (1999, Chapter 2, Proposition 4) present a version of the key computation for Proposition 1.14 which requires only that the initial distribution of the chain equals the distribution of the chain when it stops. We have essentially followed their proof.

The standard approach to demonstrating that irreducible aperiodic Markov chains have unique stationary distributions is through the Perron-Frobenius theorem. See, for instance, Karlin and Taylor (1975) or Seneta (2006).

See Feller (1968, Chapter XV) for the classification of states of Markov chains.

**Complements.** The following lemma is needed for the proof of Proposition 1.7. We include a proof here for completeness.

**LEMMA 1.27.** *If  $S \subset \mathbb{Z}^+$  has  $\gcd(S) = g_S$ , then there is some integer  $m_S$  such that for all  $m \geq m_S$  the product  $mg_S$  can be written as a linear combination of elements of  $S$  with non-negative integer coefficients.*

**PROOF.** *Step 1.* Given  $S \subset \mathbb{Z}^+$  nonempty, define  $g_S^*$  as the smallest positive integer which is an integer combination of elements of  $S$  (the smallest positive element of the additive group generated by  $S$ ). Then  $g_S^*$  divides every element of  $S$  (otherwise, consider the remainder) and  $g_S$  must divide  $g_S^*$ , so  $g_S^* = g_S$ .

*Step 2.* For any set  $S$  of positive integers, there is a finite subset  $F$  such that  $\gcd(S) = \gcd(F)$ . Indeed the non-increasing sequence  $\gcd(S \cap [1, n])$  can strictly decrease only finitely many times, so there is a last time. Thus it suffices to prove the fact for finite subsets  $F$  of  $\mathbb{Z}^+$ ; we start with sets of size 2 (size 1 is a tautology) and then prove the general case by induction on the size of  $F$ .

*Step 3.* Let  $F = \{a, b\} \subset \mathbb{Z}^+$  have  $\gcd(F) = g$ . Given  $m > 0$ , write  $mg = ca + db$  for some integers  $c, d$ . Observe that  $c, d$  are not unique since  $mg = (c + kb)a + (d - ka)b$  for any  $k$ . Thus we can write  $mg = ca + db$  where  $0 \leq c < b$ . If  $mg > (b - 1)a - b$ , then we must have  $d \geq 0$  as well. Thus for  $F = \{a, b\}$  we can take  $m_F = (ab - a - b)/g + 1$ .

*Step 4 (The induction step).* Let  $F$  be a finite subset of  $\mathbb{Z}^+$  with  $\gcd(F) = g_F$ . Then for any  $a \in \mathbb{Z}^+$  the definition of  $\gcd$  yields that  $g := \gcd(\{a\} \cup F) = \gcd(a, g_F)$ . Suppose that  $n$  satisfies  $ng \geq m_{\{a, g_F\}}g + m_F g_F$ . Then we can write  $ng - m_F g_F = ca + dg_F$  for integers  $c, d \geq 0$ . Therefore  $ng = ca + (d + m_F)g_F = ca + \sum_{f \in F} c_f f$  for some integers  $c_f \geq 0$  by the definition of  $m_F$ . Thus we can take  $m_{\{a\} \cup F} = m_{\{a, g_F\}} + m_F g_F / g$ . ■

## 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  $\{0, 1, \dots, n\}$ . 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  $\tau$  be the time required to be absorbed at one of 0 or  $n$ . Assume that  $X_0 = k$ , where  $0 \leq k \leq n$ . 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  $p_0, p_1, \dots, p_n$ . Clearly  $p_0 = 0$  and  $p_n = 1$ , while

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

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  $0 \leq k \leq n$ .

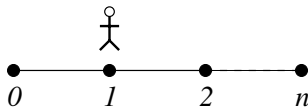


FIGURE 2.1. How long until the walk reaches either 0 or  $n$ ? What is the probability of each?

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

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

Why? When the first step of the walk increases the gambler's fortune, then the conditional expectation of  $\tau$  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  $\tau$  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  $\tau$  as a sum of geometric random variables. Let  $\tau_k$  be the total number of coupons accumulated when the collection first contains  $k$  distinct coupons. Then

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

Furthermore,  $\tau_k - \tau_{k-1}$  is a geometric random variable with success probability  $(n-k+1)/n$ : after collecting  $\tau_{k-1}$  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

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

■

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

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

$$\mathbf{P}\{\tau > \lceil n \log n + cn \rceil\} \leq e^{-c}. \quad (2.7)$$

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

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

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

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

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  $(x^1, x^2, \dots, x^n)$  by choosing a coordinate  $j \in \{1, 2, \dots, n\}$  uniformly at random and setting the new state equal to  $(x^1, \dots, x^{j-1}, 1 - x^j, x^{j+1}, \dots, x^n)$ . 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

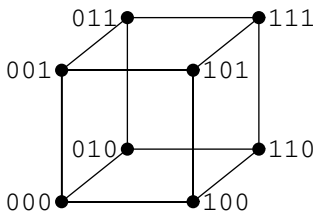


FIGURE 2.2. The three-dimensional hypercube.

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

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

Thus  $(X_t)$  is a Markov chain with state space  $\Omega = \{0, 1, 2, \dots, n\}$  that moves by  $\pm 1$  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  $\{1, \dots, n\}$ , 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*  $W(\mathbf{x})$  of a vector  $\mathbf{x} := (x^1, \dots, x^n) \in \{0, 1\}^n$  to be its number of coordinates with value 1:

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

Let  $(\mathbf{X}_t)$  be the simple random walk on the  $n$ -dimensional hypercube, and let  $W_t = W(\mathbf{X}_t)$  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.

Assume that we are given a Markov chain  $(X_0, X_1, \dots)$  with state space  $\Omega$  and transition matrix  $P$  and also some equivalence relation that partitions  $\Omega$  into equivalence classes. We denote the equivalence class of  $x \in \Omega$  by  $[x]$ . (For the Ehrenfest example, two bitstrings are equivalent when they contain the same number of 1's.)

Under what circumstances will  $([X_0], [X_1], \dots)$  also be a Markov chain? For this to happen, knowledge of what equivalence class we are in at time  $t$  must suffice to determine the distribution over equivalence classes at time  $t+1$ . If the probability  $P(x, [y])$  is always the same as  $P(x', [y])$  when  $x$  and  $x'$  are in the same equivalence class, that is clearly enough. We summarize this in the following lemma.

LEMMA 2.5. *Let  $\Omega$  be the state space of a Markov chain  $(X_t)$  with transition matrix  $P$ . Let  $\sim$  be an equivalence relation on  $\Omega$  with equivalence classes  $\Omega^\# = \{[x] : x \in \Omega\}$ , and assume that  $P$  satisfies*

$$P(x, [y]) = P(x', [y]) \quad (2.10)$$

*whenever  $x \sim x'$ . Then  $[X_t]$  is a Markov chain with state space  $\Omega^\#$  and transition matrix  $P^\#$  defined by  $P^\#([x], [y]) := P(x, [y])$ .*

The process of constructing a new chain by taking equivalence classes for an equivalence relation compatible with the transition matrix (in the sense of (2.10)) is called **projection**, or sometimes **lumping**.

## 2.4. The Pólya Urn Model

Consider the following process, known as *Pólya's urn*. Start with an urn containing two balls, one black and one white. From this point on, proceed by choosing a ball at random from those already in the urn; return the chosen ball to the urn and add another ball of the same color. If there are  $j$  black balls in the urn after  $k$  balls have been added (so that there are  $k+2$  balls total in the urn), then the probability that another black ball is added is  $j/(k+2)$ . The sequence of ordered pairs listing the numbers of black and white balls is a Markov chain with state space  $\{1, 2, \dots\}^2$ .

LEMMA 2.6. *Let  $B_k$  be the number of black balls in Pólya's urn after the addition of  $k$  balls. The distribution of  $B_k$  is uniform on  $\{1, 2, \dots, k+1\}$ .*

PROOF. Let  $U_0, U_1, \dots, U_n$  be independent and identically distributed random variables, each uniformly distributed on the interval  $[0, 1]$ . Let

$$L_k := |\{j \in \{0, 1, \dots, k\} : U_j \leq U_0\}|$$

be the number of  $U_0, U_1, \dots, U_k$  which are less than or equal to  $U_0$ .

The event  $\{L_k = j, L_{k+1} = j+1\}$  occurs if and only if  $U_0$  is the  $(j+1)$ -st smallest and  $U_{k+1}$  is one of the  $j+1$  smallest among  $\{U_0, U_1, \dots, U_{k+1}\}$ . There are  $j(k!)$  orderings of  $\{U_0, U_1, \dots, U_{k+1}\}$  making up this event; since all  $(k+2)!$  orderings are equally likely,

$$\mathbf{P}\{L_k = j, L_{k+1} = j+1\} = \frac{j(k!)}{(k+2)!} = \frac{j}{(k+2)(k+1)}. \quad (2.11)$$

Since each relative ordering of  $U_0, \dots, U_k$  is equally likely, we have  $\mathbf{P}\{L_k = j\} = 1/(k+1)$ . Together with (2.11) this implies that

$$\mathbf{P}\{L_{k+1} = j+1 \mid L_k = j\} = \frac{j}{k+2}. \quad (2.12)$$

Since  $L_{k+1} \in \{j, j+1\}$  given  $L_k = j$ ,

$$\mathbf{P}\{L_{k+1} = j \mid L_k = j\} = \frac{k+2-j}{k+2}. \quad (2.13)$$

Note that  $L_1$  and  $B_1$  have the same distribution. By (2.12) and (2.13), the sequences  $(L_k)_{k=1}^n$  and  $(B_k)_{k=1}^n$  have the same transition probabilities. Hence the sequences  $(L_k)_{k=1}^n$  and  $(B_k)_{k=1}^n$  have the same distribution. In particular,  $L_k$  and  $B_k$  have the same distribution.

Since the position of  $U_0$  among  $\{U_0, \dots, U_k\}$  is uniform among the  $k+1$  possible positions, it follows that  $L_k$  is uniform on  $\{1, \dots, k+1\}$ . Thus,  $B_k$  is uniform on  $\{1, \dots, k+1\}$ . ■

REMARK 2.7. Lemma 2.6 can also be proved by showing that  $\mathbf{P}\{B_k = j\} = 1/(k+1)$  for all  $j = 1, \dots, k+1$  using induction on  $k$ .

## 2.5. Birth-and-Death Chains

A *birth-and-death chain* has state space  $\Omega = \{0, 1, 2, \dots, n\}$ . In one step the state can increase or decrease by at most 1. The current state can be thought of as the size of some population; in a single step of the chain there can be at most one birth or death. The transition probabilities can be specified by  $\{(p_k, r_k, q_k)\}_{k=0}^n$ , where  $p_k + r_k + q_k = 1$  for each  $k$  and

- $p_k$  is the probability of moving from  $k$  to  $k+1$  when  $0 \leq k < n$ ,
- $q_k$  is the probability of moving from  $k$  to  $k-1$  when  $0 < k \leq n$ ,
- $r_k$  is the probability of remaining at  $k$  when  $0 \leq k \leq n$ ,
- $q_0 = p_n = 0$ .

PROPOSITION 2.8. *Every birth-and-death chain is reversible.*

PROOF. A function  $w$  on  $\Omega$  satisfies the detailed balance equations (1.30) if and only if

$$p_{k-1}w_{k-1} = q_k w_k$$

for  $1 \leq k \leq n$ . For our birth-and-death chain, a solution is given by  $w_0 = 1$  and

$$w_k = \prod_{i=1}^k \frac{p_{i-1}}{q_i}$$

for  $1 \leq k \leq n$ . Normalizing so that the sum is unity yields

$$\pi_k = \frac{w_k}{\sum_{j=0}^n w_j}$$

for  $0 \leq k \leq n$ . (By Proposition 1.19,  $\pi$  is also a stationary distribution.) ■

Now, fix  $\ell \in \{0, 1, \dots, n\}$ . Consider restricting the original chain to  $\{0, 1, \dots, \ell\}$ :

- For any  $k \in \{0, 1, \dots, \ell-1\}$ , the chain makes transitions from  $k$  as before, moving down with probability  $q_k$ , remaining in place with probability  $r_k$ , and moving up with probability  $p_k$ .
- At  $\ell$ , the chain either moves down or remains in place, with probabilities  $q_\ell$  and  $r_\ell + p_\ell$ , respectively.

We write  $\tilde{\mathbf{E}}$  for expectations for this new chain. By the proof of Proposition 2.8, the stationary probability  $\tilde{\pi}$  of the truncated chain is given by

$$\tilde{\pi}_k = \frac{w_k}{\sum_{j=0}^{\ell} w_j}$$

for  $0 \leq k \leq \ell$ . Since in the truncated chain the only possible moves from  $\ell$  are to stay put or to step down to  $\ell - 1$ , the expected first return time  $\tilde{\mathbf{E}}_{\ell}(\tau_{\ell}^+)$  satisfies

$$\tilde{\mathbf{E}}_{\ell}(\tau_{\ell}^+) = (r_{\ell} + p_{\ell}) \cdot 1 + q_{\ell} \left( \tilde{\mathbf{E}}_{\ell-1}(\tau_{\ell}) + 1 \right) = 1 + q_{\ell} \tilde{\mathbf{E}}_{\ell-1}(\tau_{\ell}). \quad (2.14)$$

By Proposition 1.14(ii),

$$\tilde{\mathbf{E}}_{\ell}(\tau_{\ell}^+) = \frac{1}{\tilde{\pi}(\ell)} = \frac{1}{w_{\ell}} \sum_{j=0}^{\ell} w_j. \quad (2.15)$$

We have constructed the truncated chain so that  $\tilde{\mathbf{E}}_{\ell-1}(\tau_{\ell}) = \mathbf{E}_{\ell-1}(\tau_{\ell})$ . Rearranging (2.14) and (2.15) gives

$$\mathbf{E}_{\ell-1}(\tau_{\ell}) = \frac{1}{q_{\ell}} \left( \sum_{j=0}^{\ell} \frac{w_j}{w_{\ell}} - 1 \right) = \frac{1}{q_{\ell} w_{\ell}} \sum_{j=0}^{\ell-1} w_j. \quad (2.16)$$

To find  $\mathbf{E}_a(\tau_b)$  for  $a < b$ , just sum:

$$\mathbf{E}_a(\tau_b) = \sum_{\ell=a+1}^b \mathbf{E}_{\ell-1}(\tau_{\ell}).$$

Consider two important special cases. Suppose that

$$\begin{aligned} (p_k, r_k, q_k) &= (p, r, q) \text{ for } 1 \leq k < n, \\ (p_0, r_0, q_0) &= (p, r + q, 0), \quad (p_n, r_n, q_n) = (0, r + p, q) \end{aligned}$$

for  $p, r, q \geq 0$  with  $p + r + q = 1$ . First consider the case where  $p \neq q$ . We have  $w_k = (p/q)^k$  for  $0 \leq k \leq n$ , and from (2.16), for  $1 \leq \ell \leq n$ ,

$$\mathbf{E}_{\ell-1}(\tau_{\ell}) = \frac{1}{q(p/q)^{\ell}} \sum_{j=0}^{\ell-1} (p/q)^j = \frac{(p/q)^{\ell} - 1}{q(p/q)^{\ell}[(p/q) - 1]} = \frac{1}{p - q} \left[ 1 - \left( \frac{q}{p} \right)^{\ell} \right].$$

If  $p = q$ , then  $w_j = 1$  for all  $j$  and

$$\mathbf{E}_{\ell-1}(\tau_{\ell}) = \frac{\ell}{p}.$$

## 2.6. Random Walks on Groups

Several of the examples we have already examined and many others we will study in future chapters share important symmetry properties, which we make explicit here. Recall that a **group** is a set  $G$  endowed with an associative operation  $\cdot : G \times G \rightarrow G$  and an **identity**  $\text{id} \in G$  such that for all  $g \in G$ ,

- (i)  $\text{id} \cdot g = g$  and  $g \cdot \text{id} = g$ .
- (ii) there exists an **inverse**  $g^{-1} \in G$  for which  $g \cdot g^{-1} = g^{-1} \cdot g = \text{id}$ .

Given a probability distribution  $\mu$  on a group  $(G, \cdot)$ , we define the **random walk on  $G$  with increment distribution  $\mu$**  as follows: it is a Markov chain with state space  $G$  and which moves by multiplying the current state *on the left* by a random element of  $G$  selected according to  $\mu$ . Equivalently, the transition matrix  $P$  of this chain has entries

$$P(g, hg) = \mu(h)$$

for all  $g, h \in G$ .

REMARK 2.9. We multiply the current state by the increment *on the left* because it is generally more natural in non-commutative examples, such as the symmetric group—see Section 8.1.3. For commutative examples, such as the two described immediately below, it of course does not matter on which side we multiply.

EXAMPLE 2.10 (The  $n$ -cycle). Let  $\mu$  assign probability  $1/2$  to each of  $1$  and  $n-1 \equiv -1 \pmod{n}$  in the additive cyclic group  $\mathbb{Z}_n = \{0, 1, \dots, n-1\}$ . The **simple random walk on the  $n$ -cycle** first introduced in Example 1.4 is the random walk on  $\mathbb{Z}_n$  with increment distribution  $\mu$ . Similarly, let  $\nu$  assign weight  $1/4$  to both  $1$  and  $n-1$  and weight  $1/2$  to  $0$ . Then **lazy random walk on the  $n$ -cycle**, discussed in Example 1.8, is the random walk on  $\mathbb{Z}_n$  with increment distribution  $\nu$ .

EXAMPLE 2.11 (The hypercube). The hypercube random walks defined in Section 2.3 are random walks on the group  $\mathbb{Z}_2^n$ , which is the direct product of  $n$  copies of the two-element group  $\mathbb{Z}_2 = \{0, 1\}$ . For the simple random walk the increment distribution is uniform on the set  $\{\mathbf{e}_i : 1 \leq i \leq n\}$ , where the vector  $\mathbf{e}_i$  has a  $1$  in the  $i$ -th place and  $0$  in all other entries. For the lazy version, the increment distribution gives the vector  $\mathbf{0}$  (with all zero entries) weight  $1/2$  and each  $\mathbf{e}_i$  weight  $1/2n$ .

PROPOSITION 2.12. *Let  $P$  be the transition matrix of a random walk on a finite group  $G$  and let  $U$  be the uniform probability distribution on  $G$ . Then  $U$  is a stationary distribution for  $P$ .*

PROOF. Let  $\mu$  be the increment distribution of the random walk. For any  $g \in G$ ,

$$\sum_{h \in G} U(h)P(h, g) = \frac{1}{|G|} \sum_{k \in G} P(k^{-1}g, g) = \frac{1}{|G|} \sum_{k \in G} \mu(k) = \frac{1}{|G|} = U(g).$$

For the first equality, we re-indexed by setting  $k = gh^{-1}$ . ■

**2.6.1. Generating sets, irreducibility, Cayley graphs, and reversibility.** For a set  $H \subset G$ , let  $\langle H \rangle$  be the smallest group containing all the elements of  $H$ ; recall that every element of  $\langle H \rangle$  can be written as a product of elements in  $H$  and their inverses. A set  $H$  is said to **generate**  $G$  if  $\langle H \rangle = G$ .

PROPOSITION 2.13. *Let  $\mu$  be a probability distribution on a finite group  $G$ . The random walk on  $G$  with increment distribution  $\mu$  is irreducible if and only if  $S = \{g \in G : \mu(g) > 0\}$  generates  $G$ .*

PROOF. Let  $a$  be an arbitrary element of  $G$ . If the random walk is irreducible, then there exists an  $r > 0$  such that  $P^r(\text{id}, a) > 0$ . In order for this to occur, there must be a sequence  $s_1, \dots, s_r \in G$  such that  $a = s_r s_{r-1} \dots s_1$  and  $s_i \in S$  for  $i = 1, \dots, r$ . Thus  $a \in \langle S \rangle$ .

Now assume  $S$  generates  $G$ , and consider  $a, b \in G$ . We know that  $ba^{-1}$  can be written as a word in the elements of  $S$  and their inverses. Since every element of  $G$

has finite order, any inverse appearing in the expression for  $ba^{-1}$  can be rewritten as a positive power of the same group element. Let the resulting expression be  $ba^{-1} = s_r s_{r-1} \dots s_1$ , where  $s_i \in S$  for  $i = 1, \dots, r$ . Then

$$\begin{aligned} P^m(a, b) &\geq P(a, s_1 a) P(s_1 a, s_2 s_1 a) \cdots P(s_{r-1} s_{r-2} \dots s_1 a, (ba^{-1})a) \\ &= \mu(s_1) \mu(s_2) \dots \mu(s_r) > 0. \end{aligned}$$

■

When  $S$  is a set which generates a finite group  $G$ , the **directed Cayley graph** associated to  $G$  and  $S$  is the directed graph with vertex set  $G$  in which  $(v, w)$  is an edge if and only if  $v = sw$  for some generator  $s \in S$ .

We call a set  $S$  of generators of  $G$  **symmetric** if  $s \in S$  implies  $s^{-1} \in S$ . When  $S$  is symmetric, all edges in the directed Cayley graph are bidirectional, and it may be viewed as an ordinary graph. When  $G$  is finite and  $S$  is a symmetric set that generates  $G$ , the simple random walk (as defined in Section 1.4) on the corresponding Cayley graph is the same as the random walk on  $G$  with increment distribution  $\mu$  taken to be the uniform distribution on  $S$ .

In parallel fashion, we call a probability distribution  $\mu$  on a group  $G$  **symmetric** if  $\mu(g) = \mu(g^{-1})$  for every  $g \in G$ .

**PROPOSITION 2.14.** *The random walk on a finite group  $G$  with increment distribution  $\mu$  is reversible if  $\mu$  is symmetric.*

**PROOF.** Let  $U$  be the uniform probability distribution on  $G$ . For any  $g, h \in G$ , we have that

$$U(g)P(g, h) = \frac{\mu(hg^{-1})}{|G|} \quad \text{and} \quad U(h)P(h, g) = \frac{\mu(gh^{-1})}{|G|}$$

are equal if and only if  $\mu(hg^{-1}) = \mu((hg^{-1})^{-1})$ . ■

**REMARK 2.15.** The converse of Proposition 2.14 is also true; see Exercise 2.7.

**2.6.2. Transitive chains.** A Markov chain is called **transitive** if for each pair  $(x, y) \in \Omega \times \Omega$  there is a bijection  $\varphi = \varphi_{(x,y)} : \Omega \rightarrow \Omega$  such that

$$\varphi(x) = y \quad \text{and} \quad P(z, w) = P(\varphi(z), \varphi(w)) \quad \text{for all } z, w \in \Omega. \quad (2.17)$$

Roughly, this means the chain “looks the same” from any point in the state space  $\Omega$ . Clearly any random walk on a group is transitive; set  $\varphi_{(x,y)}(g) = gx^{-1}y$ . However, there are examples of transitive chains that are not random walks on groups; see McKay and Praeger (1996).

Many properties of random walks on groups generalize to the transitive case, including Proposition 2.12.

**PROPOSITION 2.16.** *Let  $P$  be the transition matrix of a transitive Markov chain on a finite state space  $\Omega$ . Then the uniform probability distribution on  $\Omega$  is stationary for  $P$ .*

**PROOF.** Fix  $x, y \in \Omega$  and let  $\varphi : \Omega \rightarrow \Omega$  be a transition-probability-preserving bijection for which  $\varphi(x) = y$ . Let  $U$  be the uniform probability on  $\Omega$ . Then

$$\sum_{z \in \Omega} U(z)P(z, x) = \sum_{z \in \Omega} U(\varphi(z))P(\varphi(z), y) = \sum_{w \in \Omega} U(w)P(w, y),$$

where we have re-indexed with  $w = \varphi(z)$ . We have shown that when the chain is started in the uniform distribution and run one step, the total weight arriving at each state is the same. Since  $\sum_{x,z \in \Omega} U(z)P(z,x) = 1$ , we must have

$$\sum_{z \in \Omega} U(z)P(z,x) = \frac{1}{|\Omega|} = U(x).$$

■

## 2.7. Random Walks on $\mathbb{Z}$ and Reflection Principles

A *nearest-neighbor random walk* on  $\mathbb{Z}$  moves right and left by at most one step on each move, and each move is independent of the past. More precisely, if  $(\Delta_t)$  is a sequence of independent and identically distributed  $\{-1, 0, 1\}$ -valued random variables and  $X_t = \sum_{s=1}^t \Delta_s$ , then the sequence  $(X_t)$  is a nearest-neighbor random walk with increments  $(\Delta_t)$ .

This sequence of random variables is a Markov chain with infinite state space  $\mathbb{Z}$  and transition matrix

$$P(k, k+1) = p, \quad P(k, k) = r, \quad P(k, k-1) = q,$$

where  $p + r + q = 1$ .

The special case where  $p = q = 1/2$ ,  $r = 0$  is the simple random walk on  $\mathbb{Z}$ , as defined in Section 1.4. In this case

$$\mathbf{P}_0\{X_t = k\} = \begin{cases} \binom{t}{\frac{t-k}{2}} 2^{-t} & \text{if } t-k \text{ is even,} \\ 0 & \text{otherwise,} \end{cases} \quad (2.18)$$

since there are  $\binom{t}{\frac{t-k}{2}}$  possible paths of length  $t$  from 0 to  $k$ .

When  $p = q = 1/4$  and  $r = 1/2$ , the chain is the lazy simple random walk on  $\mathbb{Z}$ . (Recall the definition of lazy chains in Section 1.3.)

**THEOREM 2.17.** *Let  $(X_t)$  be simple random walk on  $\mathbb{Z}$ , and recall that*

$$\tau_0 = \min\{t \geq 0 : X_t = 0\}$$

*is the first time the walk hits zero. Then*

$$\mathbf{P}_k\{\tau_0 > r\} \leq \frac{12k}{\sqrt{r}} \quad (2.19)$$

*for any integers  $k, r > 0$ .*

We prove this by a sequence of lemmas which are of independent interest.

**LEMMA 2.18 (Reflection Principle).** *Let  $(X_t)$  be either the simple random walk or the lazy simple random walk on  $\mathbb{Z}$ . For any positive integers  $j, k$ , and  $r$ ,*

$$\mathbf{P}_k\{\tau_0 < r, X_r = j\} = \mathbf{P}_k\{X_r = -j\} \quad (2.20)$$

*and*

$$\mathbf{P}_k\{\tau_0 < r, X_r > 0\} = \mathbf{P}_k\{X_r < 0\}. \quad (2.21)$$

**PROOF.** By the Markov property, the walk “starts afresh” from 0 when it hits 0, meaning that the walk viewed from the first time it hits zero is independent of its past and has the same distribution as a walk started from zero. Hence for any  $s < r$  and  $j > 0$  we have

$$\mathbf{P}_k\{\tau_0 = s, X_r = j\} = \mathbf{P}_k\{\tau_0 = s\} \mathbf{P}_0\{X_{r-s} = j\}.$$

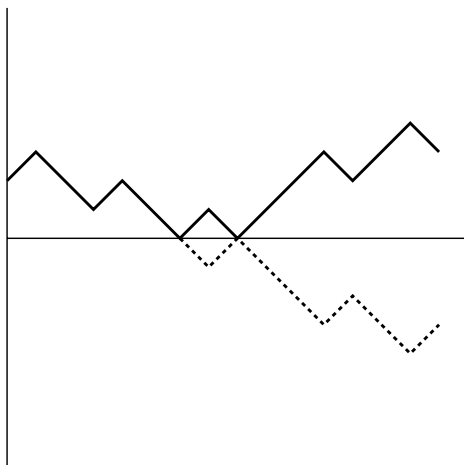


FIGURE 2.3. A path hitting zero and ending above zero can be transformed, by reflection, into a path ending below zero.

The distribution of  $X_t$  is symmetric when started at 0, so the right-hand side is equal to

$$\mathbf{P}_k\{\tau_0 = s\}\mathbf{P}_0\{X_{r-s} = -j\} = \mathbf{P}_k\{\tau_0 = s, X_r = -j\}.$$

Summing over  $s < r$ , we obtain

$$\mathbf{P}_k\{\tau_0 < r, X_r = j\} = \mathbf{P}_k\{\tau_0 < r, X_r = -j\} = \mathbf{P}_k\{X_r = -j\}.$$

To justify the last equality, note that a random walk started from  $k > 0$  must pass through 0 before reaching a negative integer.

Finally, summing (2.20) over all  $j > 0$  yields (2.21).  $\blacksquare$

REMARK 2.19. There is also a simple combinatorial interpretation of the proof of Lemma 2.18. There is a one-to-one correspondence between walk paths which hit 0 before time  $r$  and are positive at time  $r$  and walk paths which are negative at time  $r$ . This is illustrated in Figure 2.3: to obtain a bijection from the former set of paths to the latter set, reflect a path after the first time it hits 0.

EXAMPLE 2.20 (First passage time for simple random walk). A nice application of Lemma 2.18 gives the distribution of  $\tau_0$  when starting from 1 for simple random walk on  $\mathbb{Z}$ . We have

$$\begin{aligned} \mathbf{P}_1\{\tau_0 = 2m + 1\} &= \mathbf{P}_1\{\tau_0 > 2m, X_{2m} = 1, X_{2m+1} = 0\} \\ &= \mathbf{P}_1\{\tau_0 > 2m, X_{2m} = 1\} \cdot \mathbf{P}_1\{X_{2m+1} = 0 \mid X_{2m} = 1\} \\ &= \mathbf{P}_1\{\tau_0 > 2m, X_{2m} = 1\} \cdot \left(\frac{1}{2}\right). \end{aligned}$$

Rewriting and using Lemma 2.18 yields

$$\begin{aligned} \mathbf{P}_1\{\tau_0 = 2m + 1\} &= \frac{1}{2} \left[ \mathbf{P}_1\{X_{2m} = 1\} - \mathbf{P}_1\{\tau_0 \leq 2m, X_{2m} = 1\} \right] \\ &= \frac{1}{2} \left[ \mathbf{P}_1\{X_{2m} = 1\} - \mathbf{P}_1\{X_{2m} = -1\} \right]. \end{aligned}$$

Substituting using (2.18) shows that

$$\mathbf{P}_1\{\tau_0 = 2m + 1\} = \frac{1}{2} \left[ \binom{2m}{m} 2^{-2m} - \binom{2m}{m-1} 2^{-2m} \right] = \frac{1}{(m+1)2^{2m+1}} \binom{2m}{m}.$$

The right-hand side above equals  $C_m/2^{2m+1}$ , where  $C_m$  is the  $m$ -th **Catalan number**.

LEMMA 2.21. *When  $(X_t)$  is simple random walk or lazy simple random walk on  $\mathbb{Z}$ , we have*

$$\mathbf{P}_k\{\tau_0 > r\} = \mathbf{P}_0\{-k < X_r \leq k\}$$

for any  $k > 0$ .

PROOF. Observe that

$$\mathbf{P}_k\{X_r > 0\} = \mathbf{P}_k\{X_r > 0, \tau_0 \leq r\} + \mathbf{P}_k\{\tau_0 > r\}.$$

By Lemma 2.18,

$$\mathbf{P}_k\{X_r > 0\} = \mathbf{P}_k\{X_r < 0\} + \mathbf{P}_k\{\tau_0 > r\}.$$

By symmetry of the walk,  $\mathbf{P}_k\{X_r < 0\} = \mathbf{P}_k\{X_r > 2k\}$ , and so

$$\begin{aligned} \mathbf{P}_k\{\tau_0 > r\} &= \mathbf{P}_k\{X_r > 0\} - \mathbf{P}_k\{X_r > 2k\} \\ &= \mathbf{P}_k\{0 < X_r \leq 2k\} = \mathbf{P}_0\{-k < X_r \leq k\}. \end{aligned}$$

■

LEMMA 2.22. *For the simple random walk  $(X_t)$  on  $\mathbb{Z}$ ,*

$$\mathbf{P}_0\{X_t = k\} \leq \frac{3}{\sqrt{t}}. \quad (2.22)$$

REMARK 2.23. By applying Stirling's formula a bit more carefully than we do in the proof below, one can see that in fact

$$\mathbf{P}_0\{X_{2r} = 2k\} \leq \frac{1}{\sqrt{\pi r}} [1 + o(1)].$$

Hence the constant 3 is nowhere near the best possible. Our goal here is to give an explicit upper bound valid for all  $k$  without working too hard to achieve the best possible constant. Indeed, note that for simple random walk, if  $t$  and  $k$  have different parities, the probability on the left-hand side of (2.22) is 0.

PROOF. If  $X_{2r} = 2k$ , there are  $r+k$  “up” moves and  $r-k$  “down” moves. The probability of this is  $\binom{2r}{r+k} 2^{-2r}$ . The reader should check that  $\binom{2r}{r+k}$  is maximized at  $k=0$ , so for  $k=0, 1, \dots, r$ ,

$$\mathbf{P}_0\{X_{2r} = 2k\} \leq \binom{2r}{r} 2^{-2r} = \frac{(2r)!}{(r!)^2 2^{2r}}.$$

By Stirling's formula (use the bounds  $1 \leq e^{1/(12n+1)} \leq e^{1/(12n)} \leq 2$  in (A.10)), we obtain the bound

$$\mathbf{P}_0\{X_{2r} = 2k\} \leq \sqrt{\frac{8}{\pi}} \frac{1}{\sqrt{2r}}. \quad (2.23)$$

To bound  $\mathbf{P}_0\{X_{2r+1} = 2k+1\}$ , condition on the first step of the walk and use the bound above. Then use the simple bound  $[t/(t-1)]^{1/2} \leq \sqrt{2}$  to see that

$$\mathbf{P}_0\{X_{2r+1} = 2k+1\} \leq \frac{4}{\sqrt{\pi}} \frac{1}{\sqrt{2r+1}}. \quad (2.24)$$



REMARK 2.25. Figures 2.3 and 2.4 clearly illustrate versions of the same bijection. The key step in the proof of Theorem 2.24, counting the “bad” paths, is a case of (2.20): look at the paths after their first step, and set  $k = 1$ ,  $r = a + b - 1$  and  $j = b - a$ .

### Exercises

EXERCISE 2.1. Show that the system of equations for  $0 < k < n$

$$f_k = \frac{1}{2}(1 + f_{k+1}) + \frac{1}{2}(1 + f_{k-1}), \quad (2.25)$$

together with the boundary conditions  $f_0 = f_n = 0$  has a unique solution  $f_k = k(n - k)$ .

*Hint:* One approach is to define  $\Delta_k = f_k - f_{k-1}$  for  $1 \leq k \leq n$ . Check that  $\Delta_k = \Delta_{k+1} + 2$  (so the  $\Delta_k$ 's form an arithmetic progression) and that  $\sum_{k=1}^n \Delta_k = 0$ .

EXERCISE 2.2. Consider a hesitant gambler: at each time, she flips a coin with probability  $p$  of success. If it comes up heads, she places a fair one dollar bet. If tails, she does nothing that round, and her fortune stays the same. If her fortune ever reaches 0 or  $n$ , she stops playing. Assuming that her initial fortune is  $k$ , find the expected number of rounds she will play, in terms of  $n$ ,  $k$ , and  $p$ .

EXERCISE 2.3. Consider a random walk on the path  $\{0, 1, \dots, n\}$  in which the walk moves left or right with equal probability except when at  $n$  and 0. At  $n$ , it remains at  $n$  with probability  $1/2$  and moves to  $n - 1$  with probability  $1/2$ , and once the walk hits 0, it remains there forever. Compute the expected time of the walk's absorption at state 0, given that it starts at state  $n$ .

EXERCISE 2.4. By comparing the integral of  $1/x$  with its Riemann sums, show that

$$\log n \leq \sum_{k=1}^n k^{-1} \leq \log n + 1. \quad (2.26)$$

EXERCISE 2.5. Let  $P$  be the transition matrix for the Ehrenfest chain described in (2.8). Show that the binomial distribution with parameters  $n$  and  $1/2$  is the stationary distribution for this chain.

EXERCISE 2.6. Give an example of a random walk on a finite abelian group which is *not* reversible.

EXERCISE 2.7. Show that if a random walk on a group is reversible, then the increment distribution is symmetric.

EXERCISE 2.8. Show that when the transition matrix  $P$  of a Markov chain is transitive, then the transition matrix  $\widehat{P}$  of its time reversal is also transitive.

EXERCISE 2.9. Fix  $n \geq 1$ . Show that simple random walk on the  $n$ -cycle, defined in Example 1.4, is a projection (in the sense of Section 2.3.1) of the simple random walk on  $\mathbb{Z}$  defined in Section 2.7.

EXERCISE 2.10 (Reflection Principle). Let  $(S_n)$  be the simple random walk on  $\mathbb{Z}$ . Show that

$$\mathbf{P} \left\{ \max_{1 \leq j \leq n} |S_j| \geq c \right\} \leq 2\mathbf{P} \{|S_n| \geq c\}.$$

### Notes

Many of the examples in this chapter are also discussed in Feller (1968). See Chapter XIV for the gambler's ruin, Section IX.3 for coupon collecting, Section V.2 for urn models, and Chapter III for the reflection principle. Grinstead and Snell (1997, Chapter 12) discusses gambler's ruin.

See any undergraduate algebra book, for example Herstein (1975) or Artin (1991), for more information on groups. Much more can be said about random walks on groups than for general Markov chains. Diaconis (1988) is a starting place.

Pólya's urn was introduced in Eggenberger and Pólya (1923) and Pólya (1931). Urns are fundamental models for reinforced processes. See Pemantle (2007) for a wealth of information and many references on urn processes and more generally processes with reinforcement. The book Johnson and Kotz (1977) is devoted to urn models.

See Stanley (1999, pp. 219–229) and Stanley (2008) for many interpretations of the Catalan numbers.

**Complements.** Generalizations of Theorem 2.17 to walks on  $\mathbb{Z}$  other than simple random walks are very useful; we include one here.

**THEOREM 2.26.** *Let  $(\Delta_i)$  be i.i.d. integer-valued variables with mean zero and variance  $\sigma^2$ . Let  $X_t = \sum_{i=1}^t \Delta_i$ . Then*

$$\mathbf{P}\{X_t \neq 0 \text{ for } 1 \leq t \leq r\} \leq \frac{4\sigma}{\sqrt{r}}. \quad (2.27)$$

**REMARK 2.27.** The constant in this estimate is not sharp, but we will give a very elementary proof based on Chebyshev's inequality.

**PROOF.** For  $I \subseteq \mathbb{Z}$ , let

$$L_r(I) := \{t \in \{0, 1, \dots, r\} : X_t \in I\}$$

be the set of times up to and including  $r$  when the walk visits  $I$ , and write  $L_r(v) = L_r(\{v\})$ . Define also

$$A_r := \{t \in L_r(0) : X_{t+u} \neq 0 \text{ for } 1 \leq u \leq r\},$$

the set of times  $t$  in  $L_r(0)$  where the walk does not visit 0 for  $r$  steps after  $t$ . Since the future of the walk after visiting 0 is independent of the walk up until this time,

$$\mathbf{P}\{t \in A_r\} = \mathbf{P}\{t \in L_r(0)\}\alpha_r,$$

where

$$\alpha_r := \mathbf{P}_0\{X_t \neq 0, t = 1, \dots, r\}.$$

Summing this over  $t \in \{0, 1, \dots, r\}$  and noting that  $|A_r| \leq 1$  gives

$$1 \geq \mathbf{E}|A_r| = \mathbf{E}|L_r(0)|\alpha_r. \quad (2.28)$$

It remains to estimate  $\mathbf{E}|L_r(0)|$  from below, and this can be done using the local Central Limit Theorem or (in special cases) Stirling's formula.

A more direct (but less precise) approach is to first use Chebyshev's inequality to show that

$$\mathbf{P}\{|X_t| \geq \sigma\sqrt{r}\} \leq \frac{t}{r}$$

and then deduce for  $I = (-\sigma\sqrt{r}, \sigma\sqrt{r})$  that

$$\mathbf{E}|L_r(I^c)| \leq \sum_{t=1}^r \frac{t}{r} = \frac{r+1}{2},$$

whence  $\mathbf{E}|L_r(I)| \geq r/2$ . For any  $v \neq 0$ , we have

$$\mathbf{E}|L_r(v)| = \mathbf{E} \left( \sum_{t=0}^r \mathbf{1}_{\{X_t=v\}} \right) = \mathbf{E} \left( \sum_{t=\tau_v}^r \mathbf{1}_{\{X_t=v\}} \right). \quad (2.29)$$

By the Markov property, the chain after time  $\tau_v$  has the same distribution as the chain started from  $v$ . Hence the right-hand side of (2.29) is bounded above by

$$\mathbf{E}_v \left( \sum_{t=0}^r \mathbf{1}_{\{X_t=v\}} \right) = \mathbf{E}_0 \left( \sum_{t=0}^r \mathbf{1}_{\{X_t=0\}} \right).$$

We conclude that  $r/2 \leq \mathbf{E}|L_r(I)| \leq 2\sigma\sqrt{r}\mathbf{E}|L_r(0)|$ . Thus  $\mathbf{E}|L_r(0)| \geq \sqrt{r}/(4\sigma)$ . In conjunction with (2.28) this proves (2.27). ■

COROLLARY 2.28. *For the lazy simple random walk on  $\mathbb{Z}$  started at height  $k$ ,*

$$\mathbf{P}_k\{\tau_0^+ > r\} \leq \frac{8k}{\sqrt{r}}. \quad (2.30)$$

PROOF. By conditioning on the first move of the walk and then using the fact that the distribution of the walk is symmetric about 0, for  $r \geq 1$ ,

$$\mathbf{P}_0\{\tau_0^+ > r\} = \frac{1}{4}\mathbf{P}_1\{\tau_0^+ > r-1\} + \frac{1}{4}\mathbf{P}_{-1}\{\tau_0^+ > r-1\} = \frac{1}{2}\mathbf{P}_1\{\tau_0^+ > r-1\}. \quad (2.31)$$

Note that when starting from 1, the event that the walk hits height  $k$  before visiting 0 for the first time and subsequently does not hit 0 for  $r$  steps is contained in the event that the walk started from 1 does not hit 0 for  $r-1$  steps. Thus, from (2.31) and Theorem 2.26,

$$\mathbf{P}_1\{\tau_k < \tau_0\}\mathbf{P}_k\{\tau_0^+ > r\} \leq \mathbf{P}_1\{\tau_0 > r-1\} = 2\mathbf{P}_0\{\tau_0^+ > r\} \leq \frac{8}{\sqrt{r}}. \quad (2.32)$$

(The variance  $\sigma^2$  of the increments of the lazy random walk is  $1/2$ , which we bound by 1.) From the gambler's ruin formula given in (2.1), the chance that a simple random walk starting from height 1 hits  $k$  before visiting 0 is  $1/k$ . The probability is the same for a lazy random walk, so together with (2.32) this implies (2.30). ■

## From Shuffling Cards to Shuffling Genes

One reasonable restriction of the random transposition shuffle is to only allow interchanges of adjacent cards—see Figure 16.1. Restricting the moves in this

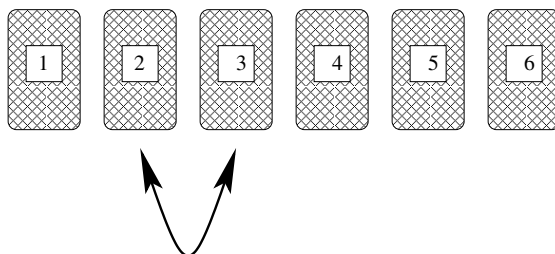


FIGURE 16.1. An adjacent transposition swaps two neighboring cards.

manner slows the shuffle down. It also breaks the symmetry of the random transpositions walk enough to require different methods of analysis.

In Section 16.1 we examine the mixing of the random adjacent transpositions walk using several different methods: upper bounds via comparison (way off) and coupling (quite sharp) and lower bounds via following a single card (off by a log factor) and Wilson’s method (sharp).

A generalization of the random adjacent transpositions model, in which entire segments of a permutation are reversed in place, can be interpreted as modeling large-scale genome changes. Varying the maximum allowed length of the reversed segments impacts the mixing time significantly. We study these reversal chains in Section 16.2.

### 16.1. Random Adjacent Transpositions

As usual we consider a lazy version of the chain to avoid periodicity problems. The resulting increment distribution assigns probability  $1/[2(n-1)]$  to each of the transpositions  $(12), \dots, (n-1n)$  and probability  $1/2$  to id.

**16.1.1. Upper bound via comparison.** We can bound the convergence of the random adjacent transposition shuffle by comparing it with the random transposition shuffle. While our analysis considers only the spectral gap and thus gives a poor upper bound on the mixing time, we illustrate the method because it can be used for many types of shuffle chains.

Note: in the course of this proof, we will introduce several constants  $C_1, C_2, \dots$ . Since we are deriving such (asymptotically) poor bounds, we will not make any effort to optimize their values. Each one does not depend on  $n$ .

First, we bound the relaxation time of the random transpositions shuffle by its mixing time. By Theorem 12.4 and Corollary 8.10,

$$t_{\text{rel}} = O(n \log n). \quad (16.1)$$

(We are already off by a factor of  $\log n$ , but we will lose so much more along the way that it scarcely matters.)

Next we compare. In order to apply Corollary 13.27, we must express an arbitrary transposition  $(ab)$ , where  $1 \leq a < b \leq n$ , in terms of adjacent transpositions. Note that

$$(ab) = (aa+1) \dots (b-1b-2)(b-1b)(b-1b-2) \dots (a+1a+2)(aa+1). \quad (16.2)$$

This path has length at most  $2n - 3$  and uses any single adjacent transposition at most twice.

We must estimate the congestion ratio

$$B = \max_{s \in \tilde{S}} \frac{1}{\mu(s)} \sum_{a \in \tilde{S}} \tilde{\mu}(a) N(s, a) |a| \leq \max_{s \in \tilde{S}} \frac{4(n-1)}{n^2} \sum_{a \in \tilde{S}} N(s, a) |a|. \quad (16.3)$$

Here  $S$  is the support of the random adjacent transposition walk,  $\mu$  is its increment distribution,  $\tilde{S}$  and  $\tilde{\mu}$  are the corresponding features of the random transpositions walk,  $N(s, a)$  is the number of times  $s$  is used in the expansion of  $a$ , and  $|a|$  is the total length of the expansion of  $a$ . Since an adjacent transposition  $s = (i \ i+1)$  lies on the generator path of  $(ab)$  exactly when  $a \leq i < i+1 \leq b$ , no generator path uses any adjacent transposition more than twice, and the length of the generator paths is bounded by  $(2n - 3)$ , the summation on the right-hand-side of (16.3) is bounded by  $2i(n-i)(2n-3) \leq n^3$ . Hence

$$B \leq 4n^2,$$

and Corollary 13.27 tells us that the relaxation time of the random adjacent transpositions chain is at most  $C_2 n^3 \log n$ .

Finally, we use Theorem 12.3 to bound the mixing time by the relaxation time. Here the stationary distribution is uniform,  $\pi(\sigma) = 1/n!$  for all  $\sigma \in \mathcal{S}_n$ . The mixing time of the random adjacent transpositions chain thus satisfies

$$t_{\text{mix}} \leq \log(4n!) C_2 n^3 \log n \leq C_3 n^4 \log^2 n.$$

**16.1.2. Upper bound via coupling.** The coupling we present here is described in Aldous (1983b) and also discussed in Wilson (2004a).

In order to couple two copies  $(\sigma_t)$  and  $(\sigma'_t)$  (the “left” and “right” decks) of the lazy version or the random adjacent transpositions chain, proceed as follows. First, choose a pair  $(i, i+1)$  of adjacent locations uniformly from the possibilities. Flip a coin to decide whether to perform the transposition on the left deck. Now, examine the cards  $\sigma_t(i)$ ,  $\sigma'_t(i)$ ,  $\sigma_t(i+1)$  and  $\sigma'_t(i+1)$  in locations  $i$  and  $i+1$  in the two decks.

- If  $\sigma_t(i) = \sigma'_t(i+1)$  or if  $\sigma_t(i+1) = \sigma'_t(i)$ , then do the opposite to the right deck: transpose if the left deck stayed still, and vice versa.
- Otherwise, perform the same action on the right deck as on the left deck.

We consider first  $\tau_a$ , the time required for a particular card  $a$  to reach the same position in the two decks. Let  $X_t$  be the (unsigned) distance between the positions of  $a$  in the two decks at time  $t$ . Our coupling ensures that  $|X_{t+1} - X_t| \leq 1$  and that if  $t \geq \tau_a$ , then  $X_t = 0$ .

Let  $M$  be the transition matrix of a random walk on the path with vertices  $\{0, \dots, n-1\}$  that moves up or down, each with probability  $1/(n-1)$ , at all interior vertices; from  $n-1$  it moves down with probability  $1/(n-1)$ , and, under all other circumstances, it stays where it is. In particular, it absorbs at state 0.

Note that for  $1 \leq i \leq n-1$ ,

$$\mathbf{P}\{X_{t+1} = i-1 \mid X_t = i, \sigma_t, \sigma'_t\} = M(i, i-1).$$

However, since one or both of the cards might be at the top or bottom of a deck and thus block the distance from increasing, we can only say

$$\mathbf{P}\{X_{t+1} = i+1 \mid X_t = i, \sigma_t, \sigma'_t\} \leq M(i, i+1).$$

Even though the sequence  $(X_t)$  is not a Markov chain, the above inequalities imply that we can couple it to a random walk  $(Y_t)$  with transition matrix  $M$  in such a way that  $Y_0 = X_0$  and  $X_t \leq Y_t$  for all  $t \geq 0$ . Under this coupling  $\tau_a$  is bounded by the time  $\tau_0^Y$  it takes  $(Y_t)$  to absorb at 0.

The chain  $(Y_t)$  is best viewed as a delayed version of a simple random walk on the path  $\{0, \dots, n-1\}$ , with a hold probability of  $1/2$  at  $n-1$  and absorption at 0. At interior nodes, with probability  $1 - 2/(n-1)$ , the chain  $(Y_t)$  does nothing, and with probability  $2/(n-1)$ , it takes a step in that walk. Exercises 2.3 and 2.2 imply that  $\mathbf{E}(\tau_0^Y)$  is bounded by  $(n-1)n^2/2$ , regardless of initial state. Hence

$$\mathbf{E}(\tau_a) < \frac{(n-1)n^2}{2}.$$

By Markov's inequality,

$$\mathbf{P}\{\tau_a > n^3\} < 1/2$$

for sufficiently large  $n$ . If we run  $2 \log_2 n$  blocks, each consisting of  $n^3$  shuffles, we can see that

$$\mathbf{P}\{\tau_a > 2n^3 \log_2 n\} < \frac{1}{n^2}.$$

Now let's look at all the cards. After  $2n^3 \log_2 n$  steps, the probability of the decks having not coupled is bounded by the sum of the probabilities of the individual cards having not coupled, so

$$\mathbf{P}\{\tau_{\text{couple}} > 2n^3 \log_2 n\} < \frac{1}{n}, \tag{16.4}$$

regardless of the initial states of the decks. Theorem 5.2 immediately implies that  $t_{\text{mix}}(\varepsilon) < 2n^3 \log_2 n$  for sufficiently large  $n$ .

**16.1.3. Lower bound via following a single card.** Consider the set of permutations

$$A = \{\sigma : \sigma(1) \geq \lfloor n/2 \rfloor\}.$$

Under the uniform distribution we have  $U(A) = (n - (\lfloor n/2 \rfloor - 1))/n \geq 1/2$ , because card 1 is equally likely to be in any of the  $n$  possible positions. However, since card 1 can change its location by at most one place in a single shuffle and since card 1 does not get to move very often, it is plausible that a large number of shuffles must be applied to a sorted deck before the event  $A$  has reasonably large probability. Below we formalize this argument.

How does card 1 move under the action of the random adjacent transposition shuffle? Let us first make the general observation that when  $(\sigma_t)$  is a random walk on  $\mathcal{S}_n$  with increment distribution  $Q$  and  $k \in [n]$ , Lemma 2.5 implies that

the sequence  $(\sigma_t(k))$  is itself a Markov chain, which we will call the *single-card chain*. Its transition matrix  $P'$  does not depend on  $k$ .

Returning to the case of (lazy) random adjacent transpositions: each interior card (neither top nor bottom of the deck) moves with probability  $1/(n-1)$ , and at each of the moves it is equally likely to jump one position to the right or one position to the left. If the card is at an endpoint, it is selected with probability  $1/2(n-1)$  and always moves in the one permitted direction. If  $(\tilde{S}_t)$  is a random walk on  $\mathbb{Z}$  which remains in place with probability  $1-1/(n-1)$  and increments by  $\pm 1$  with equal probability when it moves, then

$$\mathbf{P}\{\sigma_t(1) - 1 \geq z\} \geq \mathbf{P}\{|\tilde{S}_t| \geq z\}.$$

Thus,

$$\mathbf{P}\{\sigma_t(1) \geq n/2 + 1\} \leq \frac{4\mathbf{E}\tilde{S}_t^2}{n^2} \leq \frac{4t}{n^2(n-1)}.$$

Therefore,

$$\|P^t(\text{id}, \cdot) - U\|_{TV} \geq U(A) - P^t(\text{id}, A) \geq \frac{1}{2} - \frac{4t}{n^2(n-1)}.$$

Thus if  $t \leq n^2(n-1)/16$ , then  $d(t) \geq 1/4$ . We conclude that  $t_{\text{mix}} \geq n^2(n-1)/16$ .

**16.1.4. Lower bound via Wilson's method.** In order to apply Wilson's method (Theorem 13.5) to the random adjacent transpositions shuffle, we must specify an eigenfunction and initial state.

First, some generalities on the relationship between the eigenvalues and eigenfunctions of a shuffle chain and its single-card chain. Lemma 12.8 tells us that when  $\Phi : [n] \rightarrow \mathbb{R}$  is an eigenfunction of the single-card chain with eigenvalue  $\lambda$ , then  $\Phi^b : \mathcal{S}_n \rightarrow \mathbb{R}$  defined by  $\Phi^b(\sigma) = \Phi(\sigma(k))$  is an eigenfunction of the shuffle chain with eigenvalue  $\lambda$ .

For the random adjacent transpositions chain, the single-card chain is an extremely lazy version of a random walk on the path whose eigenfunctions and eigenvalues were determined in Section 12.3.2. Let  $M$  be the transition matrix of simple random walk on the  $n$ -path with holding probability  $1/2$  at the endpoints. Then we have

$$P' = \frac{1}{n-1}M + \frac{n-2}{n-1}I.$$

It follows from (12.18) that

$$\varphi(k) = \cos\left(\frac{(2k-1)\pi}{2n}\right)$$

is an eigenfunction of  $P'$  with eigenvalue

$$\lambda = \frac{1}{n-1} \cos\left(\frac{\pi}{n}\right) + \frac{n-2}{n-1} = 1 - \frac{\pi^2}{2n^3} + O\left(\frac{1}{n^3}\right).$$

Hence, for any  $k \in [n]$  the function  $\sigma \mapsto \varphi(\sigma(k))$  is an eigenfunction of the random transposition walk with eigenvalue  $\lambda$ . Since these eigenfunctions all lie in the same eigenspace, so will any linear combination of them. We set

$$\Phi(\sigma) = \sum_{k \in [n]} \varphi(k)\varphi(\sigma(k)). \tag{16.5}$$

REMARK 16.1. See Exercise 8.9 for some motivation of our choice of  $\Phi$ . By making sure that  $\Phi(\text{id})$  is as large as possible, we ensure that when  $\Phi(\sigma_t)$  is small, then  $\sigma_t$  is in some sense likely to be far away from the identity.

Now consider the effect of a single adjacent transposition  $(k-1\ k)$  on  $\Phi$ . Only two terms in (16.5) change, and we compute

$$\begin{aligned} |\Phi(\sigma(k-1\ k)) - \Phi(\sigma)| &= |\varphi(k)\varphi(\sigma(k-1)) + \varphi(k-1)\varphi(\sigma(k)) \\ &\quad - \varphi(k-1)\varphi(\sigma(k-1)) - \varphi(k)\varphi(\sigma(k))| \\ &= |(\varphi(k) - \varphi(k-1))(\varphi(\sigma(k)) - \varphi(\sigma(k-1)))|. \end{aligned}$$

Since  $d\varphi(x)/dx$  is bounded in absolute value by  $\pi/n$  and  $\varphi(x)$  itself is bounded in absolute value by 1, we may conclude that

$$|\Phi(\sigma(k-1\ k)) - \Phi(\sigma)| \leq \frac{\pi}{n}(2) = \frac{2\pi}{n}. \quad (16.6)$$

Combining (16.6) with Theorem 13.5 and the fact that  $\Phi(\text{id}) = n/2$  (see Exercise 8.10) tells us that when the random adjacent transposition shuffle is started with a sorted deck, after

$$t = \frac{n^3 \log n}{\pi^2} + C_\varepsilon n^3 \quad (16.7)$$

steps the variation distance from stationarity is still at least  $\varepsilon$ . (Here  $C_\varepsilon$  can be taken to be  $\log(\frac{1-\varepsilon}{64\varepsilon})$ .)

## 16.2. Shuffling Genes

Although it is amusing to view permutations as arrangements of a deck of cards, they occur in many other contexts. For example, there are (rare) mutation events involving large-scale rearrangements of segments of DNA. Biologists can use the relative order of homologous genes to estimate the evolutionary distance between two organisms. Durrett (2003) has studied the mixing behavior of the random walk on  $\mathcal{S}_n$  corresponding to one of these large-scale rearrangement mechanisms, *reversals*.

Fix  $n > 0$ . For  $1 \leq i \leq j \leq n$ , define the **reversal**  $\rho_{i,j} \in \mathcal{S}_n$  to be the permutation that reverses the order of all elements in places  $i$  through  $j$ . (The reversal  $\rho_{i,i}$  is simply the identity.)

Since not all possible reversals are equally likely in the chromosomal context, we would like to be able to limit what reversals are allowed as steps in our random walks. One (simplistic) restrictive assumption is to require that the endpoints of the reversal are at distance at most  $L$  from each other.

Applying  $\rho_{4,7}$ :

$$\boxed{9} \boxed{4} \boxed{2} \boxed{5} \boxed{1} \boxed{8} \boxed{6} \boxed{3} \boxed{7} \Rightarrow \boxed{9} \boxed{4} \boxed{2} \boxed{6} \boxed{8} \boxed{1} \boxed{5} \boxed{3} \boxed{7}$$

Applying  $\rho_{9,3}$ :

$$\boxed{9} \boxed{4} \boxed{2} \boxed{5} \boxed{1} \boxed{8} \boxed{6} \boxed{3} \boxed{7} \Rightarrow \boxed{4} \boxed{9} \boxed{7} \boxed{5} \boxed{1} \boxed{8} \boxed{6} \boxed{3} \boxed{2}$$

FIGURE 16.2. Applying reversals to permutations of length 9. Note that the second reversal wraps around the ends of the permutation.

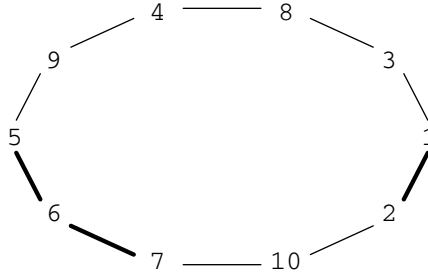


FIGURE 16.3. The permutation 1, 3, 8, 4, 9, 5, 6, 7, 10, 2 has three conserved edges.

To avoid complications at the ends of segments, we will treat our sequences as circular arrangements. Reversals will be allowed to span the “join” in the circle, and all positions will be treated mod  $n$ . See Figure 16.2. With these assumptions, we are now ready to define the  $L$ -reversal walk.

Let  $L = L(n)$  be a function of  $n$  satisfying  $1 \leq L(n) \leq n$ . The  **$L$ -reversal chain** on  $\mathcal{S}_n$  is the random walk on  $\mathcal{S}_n$  whose increment distribution is uniform on the set of all reversals of (circular) segments of length at most  $L$ . (Note that this includes the  $n$  segments of length 1; reversing a segment of length 1 results in the identity permutation.)

Equivalently, to perform a step in the  $L$ -reversal chain: choose  $i \in [n]$  uniformly, and then choose  $k \in [0, L - 1]$  uniformly. Perform the reversal  $\rho_{i, i+k}$  (which will wrap around the ends of the sequence when  $i + k > n$ ). Note that the total probability assigned to id is  $n/nL = 1/L$ .

Since each reversal is its own inverse, Proposition 2.14 ensures that the  $L$ -reversal chain is reversible.

In Section 16.2.1 we give a lower bound on the mixing time of the  $L$ -reversal chain that is sharp in some cases. In Section 16.2.2, we will present an upper bound for their mixing.

**16.2.1. Lower bound.** Although a single reversal can move many elements, it can break at most two adjacencies. We use the number of preserved adjacencies to lower bound the mixing time.

**PROPOSITION 16.2.** *Consider the family of  $L$ -reversal chains, where  $L = L(n)$  satisfies  $1 \leq L(n) < n/2$ . Fix  $0 < \varepsilon < 1$  and let  $t = t(n) = (1 - \varepsilon)\frac{n}{2} \log n$ . Then*

$$\lim_{n \rightarrow \infty} d(t) = 1.$$

**PROOF.** Superimpose the edges of a cycle onto our permutation, and say an edge is **conserved** if its endpoints are consecutive—in either order (see Figure 16.3).

Under the uniform distribution on  $\mathcal{S}_n$ , each cycle edge has probability  $2/n$  of being conserved. Hence the expected number of conserved edges is 2.

Now consider running the  $L$ -reversal chain. Each reversal breaks the cycle at 2 edges and reverses the segment in between them. Call an edge **undisturbed** if it has not been cut by any reversal. There are two reasons that a disturbed edge might end up conserved: a reversal of a segment of length 1 is simply the identity permutation and does not change adjacencies, and vertices cut apart by one reversal

might be moved back together by a later one. However, after  $t$  reversals, we may be sure that the number of conserved edges is at least as large as the number of undisturbed edges.

Start running the  $L$ -reversal chain from the identity permutation, and let  $U$  be the number of undisturbed edges at time  $t = t(n)$ . We can write  $U = U_1 + \cdots + U_n$ , where  $U_k$  is the indicator of the edge  $(k, k + 1)$  being undisturbed. Under the  $L$ -reversal model, each edge has probability  $2/n$  of being disturbed in each step, so

$$\mathbf{E}U = n \left(1 - \frac{2}{n}\right)^t \sim n^\varepsilon.$$

We can also use indicators to estimate the variance of  $U$ . At each step of the chain, there are  $nL$  reversals that can be chosen. Each edge is disturbed by exactly  $2L$  legal reversals, since it can be either the right or the left endpoint of reversals of  $L$  different lengths. If the edges are more than  $L$  steps apart, no legal reversal breaks both. If they are closer than that, exactly one reversal breaks both. Hence

$$\mathbf{P}\{U_i = 1 \text{ and } U_j = 1\} = \begin{cases} \left(\frac{nL - (4L - 1)}{nL}\right)^t & \text{if } 1 \leq j - i \leq L \text{ or } 1 \leq i - j \leq L, \\ \left(\frac{nL - 4L}{nL}\right)^t & \text{otherwise} \end{cases}$$

(in this computation, the subscripts must be interpreted mod  $n$ ).

Write  $p = \mathbf{P}(U_k = 1) = (1 - 2/n)^t \sim n^{\varepsilon - 1}$ . We can now estimate

$$\begin{aligned} \text{Var}U &= \sum_{i=1}^n \text{Var}U_i + \sum_{i \neq j} \text{Cov}(U_i, U_j) \\ &= np(1 - p) + 2nL \left( \left(1 - \frac{4 - 1/L}{n}\right)^t - p^2 \right) \\ &\quad + n(n - 2L) \left( \left(1 - \frac{4}{n}\right)^t - p^2 \right). \end{aligned}$$

Note that the sum of covariances has been split into those terms for which  $i$  and  $j$  are at a distance at most  $L$  apart and those for which they are further apart. Let's examine the resulting pieces individually. For the first one, factoring out  $p^2$  and taking a power series expansion gives

$$p^2 \cdot 2nL \left( \left(1 + \frac{1}{nL} + O\left(\frac{1}{n^2}\right)\right)^t - 1 \right) = O\left(\frac{p^2 n L t}{nL}\right) = o(np),$$

so these terms (which are positive) are negligible compared to  $\mathbf{E}U$ .

Doing the same to the second piece yields

$$n(n - k)p^2 \left( \left(1 - \frac{4}{n^2 - 4n + 4}\right)^t - 1 \right) = O\left(n^2 p^2 \cdot \frac{t}{n^2}\right) = o(np),$$

so that these terms (which are negative) are also negligible compared to  $\mathbf{E}U$ . Since  $p = o(1)$ , we can conclude that

$$\text{Var}U \sim \mathbf{E}U. \tag{16.8}$$

Let  $A \subseteq \mathcal{S}_n$  be the set of permutations with at least  $\mathbf{E}U/2$  conserved edges. Under the uniform distribution on  $\mathcal{S}_n$ , the event  $A$  has probability less than or equal to  $4/\mathbf{E}U$ , by Markov's inequality.

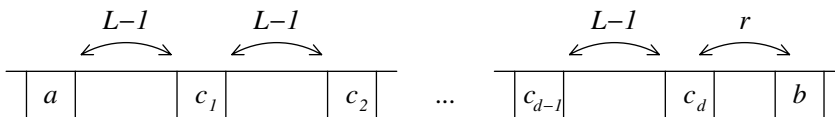


FIGURE 16.4. To express  $(ab)$  in terms of short transpositions, first carry the marker at position  $a$  over to position  $b$ ; then perform all but the last transposition in reverse order to take the marker at position  $b$  over to position  $a$ .

By Chebyshev’s inequality and (16.8), for sufficiently large  $n$  we have

$$P^t(\text{id}, A^c) \leq \mathbf{P}\{|U - \mathbf{EU}| > \mathbf{EU}/2\} \leq \frac{\text{Var}U}{(\mathbf{EU}/2)^2} < \frac{5}{\mathbf{EU}}.$$

By the definition (4.1) of total variation distance,

$$\|P^t(\text{id}, \cdot) - U\|_{\text{TV}} \geq \left(1 - \frac{5}{\mathbf{EU}}\right) - \frac{4}{\mathbf{EU}} = 1 - \frac{9}{\mathbf{EU}}.$$

Since  $\mathbf{EU} \sim n^\varepsilon$ , we are done. ■

**16.2.2. Upper bound.** We now give an upper bound on the mixing time of the  $L$ -reversal chain via the comparison method, using the same inefficient methods as we did for random adjacent transpositions in Section 16.1.1. To avoid problems with negative values, we consider a lazy version of the  $L$ -reversal chain: at each step, with probability  $1/2$ , perform a uniformly chosen  $L$ -reversal, and with probability  $1/2$ , do nothing.

Again, our exemplar chain for comparison will be the random transposition chain.

To bound the relaxation time of the  $L$ -reversal chain, we must expand each transposition  $(ab) \in \mathcal{S}_n$  as a product of  $L$ -reversals. To show the effect of choice of paths, we try three different strategies and compare the resulting congestion ratios.

We can normalize our presentation of the transposition  $(ab)$  so that the distance around the cycle from  $a$  to  $b$  in the positive direction is at most  $n/2$ . Call the transposition  $(ab)$  **short** when  $b = a + k$  for some  $k < L$  (interpreted mod  $n$  if necessary); call a transposition **long** if it is not short. When  $b = a + 1$ , we have  $(ab) = \rho_{a,b}$ . When  $a + 2 \leq b \leq a + k$ , we have  $(ab) = \rho_{a+1,b-1} \rho_{a,b}$ . We use these paths of length 1 or 2 for all short transpositions. We will express our other paths below in terms of short transpositions; to complete the expansion, we replace each short transposition with two  $L$ -reversals.

*Paths for long transpositions, first method.* Let  $(ab)$  be a long transposition. We build  $(ab)$  by taking the marker at position  $a$  on maximal length leaps for as long as we can, then finishing with a correctly-sized jump to get to position  $b$ ; then take the marker that was at position  $b$  over to position  $a$  with maximal length leaps. More precisely, write

$$b = a + d(L - 1) + r,$$

with  $0 \leq r < L - 1$ , and set  $c_i = a + i(L - 1)$  for  $1 \leq i \leq d$ . Then

$$(ab) = [(a \ c_1)(c_1 \ c_2) \dots (c_{d-1} \ c_d)] (b \ c_d) [(c_d \ c_{d-1}) \dots (c_2 \ c_1)(c_1 \ a)].$$

See Figure 16.4.

Consider the congestion ratio

$$B = \max_{s \in \tilde{S}} \frac{1}{\mu(s)} \sum_{\tilde{s} \in \tilde{S}} \tilde{\mu}(\tilde{s}) N(s, \tilde{s}) |\tilde{s}| \leq \max_{\rho_{i,j} \in S} \frac{4L}{n} \sum_{(a,b) \in \tilde{S}} O\left(\frac{n}{L}\right)$$

of Corollary 13.27. Here  $S$  and  $\mu$  come from the  $L$ -reversal walk, while  $\tilde{S}$  and  $\tilde{\mu}$  come from the random transpositions walk. The initial estimate goes through because the length of all generator paths is at most  $O(n/L)$ , while any single  $L$ -reversal can be used at most twice in a single generator path.

We must still bound the number of different paths in which a particular reversal might appear. This will clearly be maximized for the reversals of length  $L-1$ , which are used in both the “leaps” of length  $L-1$  and the final positioning jumps. Given a reversal  $\rho = \rho_{i,i+L-1}$ , there are at most  $(n/2)/(L-1)$  possible positions for the left endpoint  $a$  of a long transposition whose path includes  $\rho$ . For each possible left endpoint, there are fewer than  $n/2$  possible positions for the right endpoint  $b$  (we could bound this more sharply, but it would only save us a factor of 2 to do so). The reversal  $\rho$  is also used for short transpositions, but the number of those is only  $O(1)$ . Hence for this collection of paths we have

$$B = O\left(\frac{n^2}{L}\right).$$

*Paths for long transpositions, second method.* We now use a similar strategy for moving markers long distances, but try to balance the usage of short transpositions of all available sizes. Write

$$b = a + c \left( \frac{L(L-1)}{2} \right) + r,$$

with  $0 \leq r < L(L-1)/2$ .

To move the marker at position  $a$  to position  $b$ , do the following  $c$  times: apply the transpositions that move the marker by  $L-1$  positions, then by  $L-2$  positions, and so on, down to moving 1 position. To cover the last  $r$  steps, apply transpositions of lengths  $L-1, L-2, \dots$  until the next in sequence hits exactly or would overshoot; if necessary, apply one more transposition to complete moving the marker to position  $b$ . Reverse all but the last transposition to move the marker from position  $b$  to position  $a$ .

Estimating the congestion ratio works very similarly to the first method. The main difference arises in estimating the number of transpositions  $(a,b)$  whose paths use a particular reversal  $\rho = \rho_{i,j}$ . Now the left endpoint  $a$  can fall at one of at most  $2 \left( \frac{n/2}{L(L-1)/2} \right)$  positions (the factor of 2 comes from the possibility that  $\rho$  is the final jump), since there are at most this number of possible positions for a transposition of the same length as  $\rho$  in one of our paths. The right endpoint  $b$  again has at most  $n/2$  possible values (again, an overestimate that only affects the lead constant). We get

$$B = O\left(\frac{n^2}{L^2}\right). \tag{16.9}$$

That is, we have asymptotically reduced the congestion ratio by a factor of  $L$  by changing the paths to use reversals of all sizes evenly.

*Paths for long transpositions, third method: randomized.* We can use the method described in Remark 13.28 of choosing random, rather than canonical,

paths to match the bound of (16.9). We again describe the paths in terms of short transpositions; to complete the expansion, replace each short transposition with two short reversals.

Fix a transposition  $(bc)$ . Take  $b$  on jumps towards  $c$  of size uniformly chosen between  $L/2 + 1$  and  $L - 1$  until it is within distance  $L - 1$  of  $c$ ; then make the last jump the required size. To take  $c$  back, use the same sequence of jumps, but in reverse.

We must estimate the congestion ratio of (13.28):

$$B = \max_{s \in S} \frac{1}{\mu(s)} \sum_{a \in \tilde{S}} \tilde{\mu}(a) \sum_{\Gamma \in \mathcal{P}_a} \nu_a(\Gamma) N(s, \Gamma) |\Gamma|.$$

Since all but the last step of our paths are reversals of length at least  $L/2$ , for all  $\Gamma$  of positive measure we have  $n/L + O(1) < |\Gamma| < 2n/L + O(1)$ . Any single reversal can appear at most twice in a single path. Hence

$$B \leq \max_{s \in \tilde{S}} \frac{2nL}{n^2} \left( \frac{2n}{L} + O(1) \right) \sum_{a \in \tilde{S}} \sum_{\Gamma \in \mathcal{P}_a} \nu_a(\Gamma).$$

For any  $a \in \tilde{S}$ , the number of pairs  $(b, c)$  for which  $a$  can be used is certainly at most  $n^2$ . Once we fix  $b$  and  $c$ , the probability of hitting exactly the span of  $a$  while choosing the random path is at most  $2(2/L)^2$ . (Why? The reversal  $a$  is used by at most 2 short transpositions. The probability of choosing the correct left endpoint for one of those transpositions is at most  $(2/L)$  (to make this clearer, consider conditioning on all possible partial paths long enough that the left endpoint could possibly be hit). Once the correct left endpoint is hit, the probability of hitting the correct right endpoint is bounded by  $2/L$ .) Hence for this construction of random paths, we have  $B = O(n^2/L^2)$ .

REMARK 16.3. Notice that when  $L = 2$ , all three methods reduce to the paths used in Section 16.1.1 for random adjacent transpositions.

To finish bounding the mixing time, we follow the method of our low-quality estimate (16.1) of  $O(n \log n)$  for the relaxation time of the random transposition chain. By Corollary 13.27 and the laziness of the  $L$ -reversal chain, we have

$$t_{\text{rel}} = O\left(\frac{n^3 \log n}{L^2}\right)$$

for the  $L$ -reversal chain. Finally, as in Section 16.1.1, we use Theorem 12.3 to bound the mixing time by the relaxation time, obtaining

$$t_{\text{mix}} \leq \log(4n!) t_{\text{rel}} = O\left(\frac{n^4 \log^2 n}{L^2}\right).$$

### Exercise

EXERCISE 16.1. Modify the argument of Proposition 16.2 to cover the case  $n/2 < L < n - 1$ . (Hint: there are now pairs of edges both of which can be broken by two different allowed reversals.)

## Notes

Random adjacent transpositions are among the examples analyzed by Diaconis and Saloff-Coste (1993b), who introduced the comparison method for groups. While our presentation uses the same paths and gets the same inequality between the underlying Dirichlet forms, our final bound on the mixing time is much weaker because we apply this inequality only to the spectral gap. Diaconis and Shahshahani (1981) derived very precise information on the spectrum and convergence behavior of the random transpositions walk, and Diaconis and Saloff-Coste (1993b) exploited this data to obtain an  $O(n^3 \log n)$  upper bound on the mixing time of the random adjacent transpositions chain.

Diaconis and Saloff-Coste (1993b) proved the first lower bound we present for this chain and conjectured that the upper bound is of the correct asymptotic order. That it is was shown in Wilson (2004a).

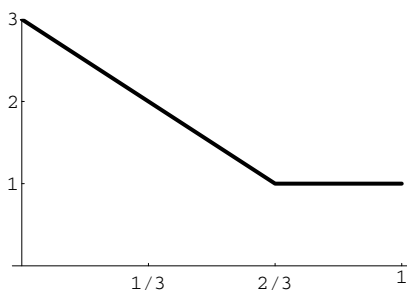


FIGURE 16.5. When  $L = n^\alpha$  and  $0 < \alpha < 1$ , the mixing of the  $L$ -reversal chain takes at least  $\Omega(n^{1 \vee (3-3\alpha)} \log n)$  steps. This plot shows  $1 \vee (3 - 3\alpha)$ .

Durrett (2003) introduced the  $L$ -reversal chain and proved both bounds we present. For the upper bound, our presentation has again significantly weakened the result by considering only the spectral gap; Durrett proved an upper bound of order  $O\left(\frac{n^3 \log n}{L^2}\right)$ .

Durrett (2003) also used Wilson's method to give another lower bound, of order  $\Omega\left(\frac{n^3 \log n}{L^3}\right)$ , when  $L \sim n^\alpha$  for some  $0 < \alpha < 1$ . Taking the maximum of the two lower bounds for  $L$  in this range tells us that the mixing of the  $L$ -reversal chain takes at least  $\Omega(n^{1 \vee (3-3\alpha)} \log n)$  steps—see Figure 16.5. Durrett conjectured that this lower bound is, in fact, sharp.

Cancrini, Caputo, and Martinelli (2006) showed that the relaxation time of the  $L$ -reversal chain is  $\Theta(n^{1 \vee (3-3\alpha)})$ . Morris (2008) has proved an upper bound on the mixing time that is only  $O(\log^2 n)$  larger than Durrett's conjecture.

Kandel, Matias, Unger, and Winkler (1996) discuss shuffles relevant to a different problem in genomic sequence analysis.