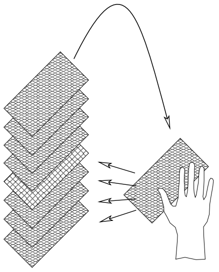
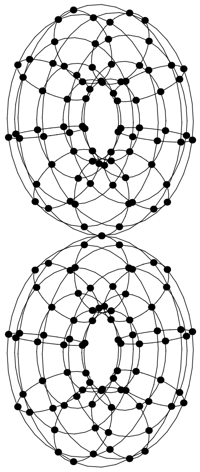
Markov Chains and Mixing Times

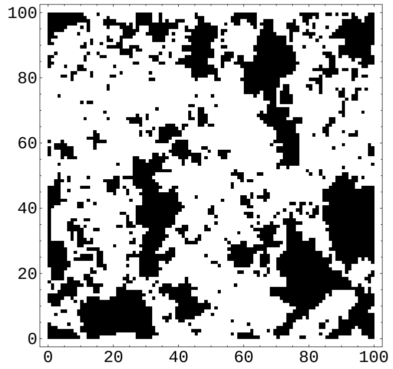
David A. Levin Yuval Peres Elizabeth L. Wilmer

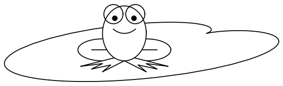
with a chapter on coupling from the past by James G. Propp and David B. Wilson

DRAFT, version of September 15, 2007.









David A. Levin
Department of Mathematics
University of Oregon
dlevin@uoregon.edu
http://www.uoregon.edu/~dlevin/

Yuval Peres
Microsoft Research
and University of California, Berkeley
peres@stat.berkeley.edu
http://stat-www.berkeley.edu/~peres/

Elizabeth L. Wilmer
Department of Mathematics
Oberlin College
Elizabeth.Wilmer@oberlin.edu
http://www.oberlin.edu/math/faculty/wilmer.html

Acknowledgements

The authors thank the Mathematical Sciences Research Institute, the National Science Foundation VIGRE grant to the Department of Statistics at the University of California, Berkeley, and National Science Foundation grants DMS-0244479 and DMS-0104073 for support. We also thank Hugo Rossi for suggesting we embark on this project. Thanks to Jian Ding, Tom Hayes, Itamar Landau, Yun Long, ¡¡¡¡¡¡¡ .mine Karola Meszaros, Shobhana Murali, and Sithparran ======= Karola Meszarosfor, Shobhana Murali, Tomoyuki Shirai, and Sithparran ¿¿¿¿¿¿¿ .r572 Vanniasegaram for corrections to an earlier version and making valuable suggestions. Yelena Shvets made the illustration in Section 7.5.1. The simulations of the Ising model in Chapter 15 are due to Raissa D'Souza. We thank László Lovász for useful discussions. We thank Robert Calhoun for technical assistance.

Contents

CHAPTER 1

Introduction

Consider the following (inefficient) method of shuffling a stack of cards: a card is taken from the top of the deck and placed at a randomly chosen location in the deck. This is known as the *top-to-random* shuffle, not surprisingly.

We want a mathematical model for this type of process. Suppose that several of these shuffles have been performed in succession, each time changing the composition of the deck a little bit. After the next shuffle, the cards will be in some order, and this ordering will depend only on the order of the cards now and the outcome of the next shuffle. This property is important because to describe the evolution of the deck, we need only specify the probability of moving from one ordering of cards to any other ordering of cards in one shuffle.

The proper model for this card shuffling procedure is called a *Markov chain*. From any arrangements of cards, it is possible to get to any other by a sequence of top-to-random shuffles. We may suspect that after many of these moves, the deck should become randomly arranged. Indeed, this is the motivation for performing any kind of shuffle, as we are attempting to randomize the deck. Here, by "randomly arranged," we mean that each arrangement of the cards is equally likely.

Under mild regularity conditions, a Markov chain converges to a unique stationary distribution. Traditional undergraduate treatments of Markov chains examine fixed chains as time goes to infinity. In the past two decades, a different asymptotic analysis has emerged. For a Markov chain with a large state space, we care about the *finite* number of steps needed to get the distribution reasonably close to its limit. This number is known as the *mixing time* of the chain. There are now many methods for determining its behavior as a function of the geometry and size of the state space.

Aldous and Diaconis (1986) presented the concept of mixing times to a wider audience, using card shuffling as a central example. Since then, both the field and its interactions with computer science and statistical physics have grown tremendously. Many of these exciting developments can and should be communicated to undergraduates. We hope to present this beautiful and relevant material in an accessible way. This book is intended for a second undergraduate course in probability and emphasizes current developments in the rigorous analysis of convergence time for Markov chains.

The course will expose students to both key mathematical and probabilistic concepts and the interactions of probability with other disciplines. The models we analyze will largely be "particle systems" arising in statistical physics. Interestingly, many of these models exhibit *phase transitions*: the behavior of the model may change abruptly as a parameter describing local interactions passes through a critical value. For our particle systems, the mixing time may vary from “fast” (polynomial in the instance size n) to “slow” (exponential in n) as interaction parameters pass through a critical value.

CHAPTER 2

Discrete Simulation

{ch:simtechs}

2.1. What Is Simulation?

Let X be a random unbiased bit:

$$\mathbf{P}\{X = 0\} = \mathbf{P}\{X = 1\} = \frac{1}{2}. \tag{2.1}$$

If we assign the value 0 to the “heads” side of a coin, and the value 1 to the “tails” side, we can generate a bit which has the same distribution as X by tossing the coin.

Suppose now the bit is biased, so that

$$\mathbf{P}\{X = 1\} = \frac{1}{4}, \qquad \mathbf{P}\{X = 0\} = \frac{3}{4}. \tag{2.2}$$ {Eq:BiasedBit}

Again using only our (fair) coin toss, we are able to easily generate a bit with this distribution: Toss the coin twice and assign the value 1 to the result “two heads”, and the value 0 to all other possible outcomes. Since the coin cannot remember the result of the first toss when it is tossed for the second time, the tosses are independent and the probability of two heads is 1/4 (ideally, assuming the coin is perfectly symmetric.) This is a recipe for generating observations of a random variable which has the same distribution (2.2) as X. This is called a *simulation* of X.

Consider the random variable U_n which is uniform on the finite set

$$\left\{0, \frac{1}{2^n}, \frac{2}{2^n}, \ldots, \frac{2^n - 1}{2^n}\right\}. \tag{2.3}$$ {Eq:Dyadics}

This random variable is a discrete approximation to the uniform distribution on $[0, 1]$. If our only resource is the humble fair coin, we are still able to simulate U_n: toss the coin n times to generate independent unbiased bits $X_1, X_2, \ldots, X_n$, and output the value

$$\sum_{i=1}^{n} \frac{X_i}{2^i}. \tag{2.4}$$ {Eq:RandomSum}

This random variables has the uniform distribution on the set in (2.3). (See Exercise 2.9.)

Consequently, a sequence of independent and unbiased bits can be used to simulate a random variable whose distribution is close to uniform on $[0, 1]$. A sufficient number of bits should be used to ensure that the error in the approximation is small enough for any needed application. A computer can store a real number only to finite precision, so if the value of the simulated variable is to be placed in computer memory, it will be rounded to some finite decimal approximation. With this in mind, the discrete variable in (2.4) will be just as useful as a variable uniform on the interval of real numbers $[0, 1]$.

{Sec:PseudoRandom}

2.2. About Random Numbers

Because most computer languages provide a built-in capability for simulating random numbers chosen independently from the uniform density on the unit interval $[0, 1]$, we will assume throughout this book that there is a ready source of independent uniform-$[0, 1]$ random variables.

This assumption requires some further discussion, however. Since computers are finitary machines and can work with numbers of only finite precision, it is in fact impossible for a computer to generate a continuous random variable. Not to worry: a discrete random variable which is uniform on, for example, the set in (2.3) is a very good approximation to the uniform distribution on $[0, 1]$, at least when n is large.

A more serious issue is that computers do not produce truly random numbers at all. Instead, they use deterministic algorithms, called *pseudorandom number generators*, to produce sequences of numbers that *appear* random. There are many tests which identify features which are unlikely to occur in a sequence of independent and identically distributed random variables. If a sequence produced by a pseudorandom number generator can pass a battery of these tests, it is considered an appropriate substitute for random numbers.

One technique for generating pseudorandom numbers is a *linear congruential sequence* (LCS). Let x_0 be an integer seed value. Given that x_{n-1} has been generated, let

$$x_n = (ax_{n-1} + b) \mod m. \tag{2.5}$$

Here a, b and m are fixed constants. Clearly, this produces integers in $\{0, 1, \ldots, m\}$; if a number in $[0, 1]$ is desired, divide by m.

The properties of $(x_0, x_1, x_2, \ldots)$ vary greatly depending on choices of a, b and m, and there is a great deal of art and science behind making judicious choices for the parameters. For example, if $a = 0$, the sequence doesn't look random at all!

Any linear congruential sequence is eventually periodic. (Exercise 2.8.) The period of a LCS can be much less than m, the longest possible value.

The goal of any method for generating pseudorandom numbers is to generate output which is difficult to distinguish from truly random numbers using statistical methods. It is an interesting question whether a given pseudorandom number generator is good. We will not enter into this issue here, but the reader should be aware that the "random" numbers produced by today's computers are not in fact random, and sometimes this can lead to inaccurate simulations. For an excellent discussion of these issues, see Knuth (1997).

2.3. Simulating Discrete Distributions and Sampling from Combinatorial Sets

A Poisson random variable X with mean λ has mass function

$$p(k) := \frac{e^{-\lambda}\lambda^k}{k!}.$$

X can be simulated using a uniform random variable U as follows: subdivide the unit interval into adjacent subintervals $I_1, I_2, \ldots$ where the length of I_k is $p(k)$. Because the chance a random point in $[0,1]$ falls in I_k is $p(k)$, the index X for which $U \in I_X$ is a Poisson random variable with mean λ.

In principle, any discrete random variable can be simulated from a uniform random variable using this method. To be concrete, suppose X takes on the values $a_1, \ldots, a_N$ with probabilities $p_1, p_2, \ldots, p_N$. Let $F_k := \sum_{j=1}^{k} p_j$ (and $F_0 := 0$), and define $\phi : [0,1] \to \{a_1, \ldots, a_N\}$ by

$$\phi(u) := a_k \text{ if } F_{k-1} < u \leq F_k. \tag{2.6}$$ {Eq:DiscreteSim}

If $X = \phi(U)$, where U is uniform on $[0,1]$, then $\mathbf{P}\{X = a_k\} = p_k$. (Exercise 2.9.)

Much of this book is concerned with the problem of simulating discrete distributions. This may seem odd, as we just described an algorithm for simulating any discrete distribution!

One obstacle is that this recipe requires that the probabilities $(p_1, \ldots, p_N)$ are known exactly, while in many applications these are only known up to constant multiple. This is a more common situation than the reader may imagine, and in fact many of the central examples treated in this book fall into this category.

A random element of a finite set is called a *uniform sample* if it is equally likely to be any of the members of the set. Many applications require uniform samples from combinatorial sets whose sizes are not known.

{Example:SAW}

Example 2.1 (Self-avoiding walks). A self-avoiding walk in $\mathbb{Z}^2$ of length n is a sequence $(z_0, z_1, \ldots, z_n)$ such that $z_0 = (0,0)$, $|z_i - z_{i-1}| = 1$, and $z_i \neq z_j$ for $i \neq j$. See figure 2.1 for an example of length 6. Let Ξ_n be the collection of all self-avoiding walks of length n. Chemical and physical structures such as molecules and polymers are often modeled as "random" self-avoiding walks, that is, as uniform samples from Ξ_n.

Unfortunately, a formula for the size of Ξ_n is not known. Although the size can be calculated by computer for a fixed n if n is small enough, for sufficiently large n this is not possible. Nonetheless, we still desire (a practical) method for sampling uniformly from Ξ_n. We present a Markov chain in Example 4.23 whose state space is the set of all self-avoiding walks of a given length and whose stationary distribution is uniform.

A nearest-neighbor path $0 = v_0, \ldots, v_n$ is *non-reversing* if $v_k \neq v_{k-2}$ for $k = 2, \ldots, n$. It is simple to generate a non-reversing path recursively. First choose v_1 uniformly at random from $\{(0,1), (1,0), (0,-1), (-1,0)\}$. Given that $v_0, \ldots, v_{k-1}$ is a non-reversing path, choose v_k uniformly from the three sites in $\mathbb{Z}^2$ at distance 1 from v_{k-1} but different from v_{k-2}.

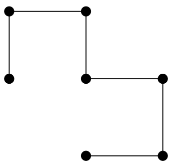


FIGURE 2.1. A self-avoiding path fig:SAW

Let Ξ_n^{nr} be the set of non-reversing nearest-neighbor paths of length n. The above procedure generates a uniform random sample from Ξ_n^{nr}. (Exercise 2.10.)

Exercise 2.11 implies that if we try generating random non-reversing paths until we get a self-avoiding path, the expected number of trials required grows exponentially in the length of the paths.

Many problems are defined for a family of structures indexed by *instance size*. For example, we desire an algorithm for generating uniform samples from self-avoiding paths of length n, for each n. The efficiency of solutions is measured by the growth of *run-time* as a function of instance size. If the run-time grows exponentially in instance size, the algorithm is considered impractical.

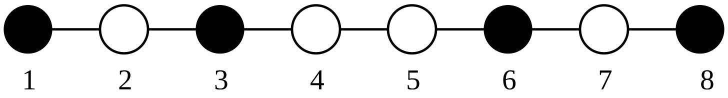


FIGURE 2.2. A configuration of the hard-core gas model with $n = 8$. Colored circles correspond to occupied sites.

{Xmple:1dHC}

EXAMPLE 2.2 (One dimensional hard-core gas). The *hard-core gas* models the random distribution of particles under the restriction that the centers of any two particles are at least a fixed distance apart. In one dimension, the state space Ω_n is functions $\omega : \{1, 2, \dots, n\} \to \{0, 1\}$ satisfying $\omega(j)\omega(j+1) = 0$ for $j = 1, \dots, n-1$. We think of $\{1, 2, \dots, n\}$ as sites arranged linearly, and ω as describing a configuration of particles on $\{1, \dots, n\}$. The condition $\omega(j) = 1$ indicates that site j is occupied by a particle. The constraint $\omega(j)\omega(j+1) = 0$ means that no two adjacent sites are both occupied by particles.

Exercise 2.12 suggests an algorithm for inductively generating a random sample from Ω_n: Suppose you are able to generate random samples from Ω_k for $k \leq n-1$. With probability f_{n-1}/f_{n+1}, put a 1 at location n, a 0 at location $n-1$, and then generate a random element of Ω_{n-2} to fill out the configuration at $\{1, 2, \dots, n-2\}$. With the remaining probability f_n/f_{n+1}, put a 0 at location n and fill out the positions $\{1, 2, \dots, n-1\}$ with a random element of Ω_{n-1}.

{Example:DominoTilings}

EXAMPLE 2.3 (Domino Tilings). A domino tile is a 2×1 or 1×2 rectangle, and, informally speaking, a domino tiling of a region is a partition of the region into domino tiles, disjoint except along their boundaries.

Consider the set $\mathcal{T}_{n,m}$ of all domino tilings of an $n \times m$ checker board. See figure 2.3 for an element of $\mathcal{T}_{6,6}$. Random domino tilings arise in statistical physics,

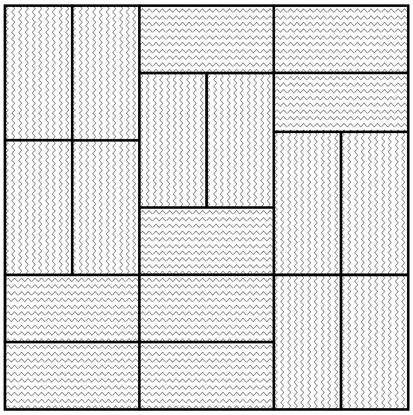


Figure 2.3. A domino tiling of a 6×6 checkerboard. Fig:Domino

and it was a physicist who first completed the daunting combinatorial calculation of the size of $\mathcal{T}_{n,m}$. (See Notes.)

Although the size N of $\mathcal{T}_{n,m}$ is known, the simulation method using (2.6) is not necessarily the best. The elements of $\mathcal{T}_{n,m}$ must be enumerated so that when an integer in $\{1, \dots, N\}$ is selected, the corresponding tiling can be generated.

To summarize, we would like methods for picking at random from large combinatorial sets which do not require enumerating the set or even knowing how many elements are in the set. We will see later that Markov chain Monte Carlo often provides such a method.

2.4. Randomly Ordered Decks Of Cards: Random Permutations

{Sec:SimPerms}

If a game is to be played from a deck of cards, fairness usually requires that the deck is completely random. That is, each of the 52! arrangements of the 52 cards should be equally likely.

An arrangements of cards in a particular order is an example of a *permutation*. Mathematically, a permutation on $[n] := \{1, 2, \dots, n\}$ is a mapping from $[n]$ to itself which is both one-to-one and onto. The collection $\mathcal{S}_n$ of all permutations on $[n]$ is called the *symmetric group*.

We describe a simple algorithm for generating a random permutation. Let σ_0 be the identity permutation. For $k = 1, 2, \dots, n-1$ inductively construct σ_k from σ_{k-1} by swapping the cards at location k and J_k, where J_k is an integer picked uniformly in $[k, n]$, independently of previous picks. More precisely,

$$\sigma_k(k) := \sigma_{k-1}(J_k), \quad \sigma_k(J_k) := \sigma_{k-1}(k), \quad \text{and} \quad \sigma_k(i) := \sigma_{k-1}(i) \text{ for } i \neq k, J_k.$$

The *kth position* refers to the image of k under the permutation. At the kth stage, a particular choice for the kth position has probability $(n-k+1)^{-1}$. Consequently, the probability of generating a particular permutation is $\prod_{k=1}^{n}(n-k+1)^{-1} = (n!)^{-1}$.

This method requires n steps, which is quite efficient. However, this is not how any human being shuffles cards! For a standard deck of playing cards, it would require 52 steps, many more operations than the usual handful of standard shuffles. We will discuss several methods of shuffling cards later, which generate *approximate* random permutations on n things. Our interest will be in how many shuffles need to be applied before the approximation to a random permutation is good.

xercise:RandomFunction}

Exercise 2.1. Suppose that a random function $\sigma : [n] \to [n]$ is created by letting $\sigma(i)$ be a random element of $[n]$, independently for each $i = 1, \dots, n$. If the resulting function σ is a permutation, stop, and otherwise begin anew by generating a fresh random function. *Stirling's Formula* (see Feller (1968, Chapter II, Equation 9.1) or Graham et al. (1994, Table 452)) gives the approximation

{Eq:Stirling}

$$n! \sim \sqrt{2\pi} n^{n+1/2} e^{-n}, \tag{2.7}$$

where $a_n \sim b_n$ means that $\lim_{n\to\infty} a_n/b_n = 1$. Use (2.7) to approximate the expected number of times a random function must be generated before a permutation results.

Exercise:BadPermMethod}

Exercise 2.2. Consider the following variation of our method for generating random permutations: let σ_0 be the identity permutation. For $k = 1, 2, \dots, n$ inductively construct σ_k from σ_{k-1} by swapping the cards at location k and J_k, where J_k is an integer picked uniformly in $[1, n]$, independently of previous picks.

For which values of n does this variant procedure yield a uniform random permutation?

2.5. Random Colorings

A proper k-coloring of $[n] := \{1, 2, \dots, n\}$ is a map $h : [n] \to [k]$ such that

$$h(j) \neq h(j+1) \text{ for } j = 1, 2, \dots, n-1.$$

The reader should imagine each of $\{1, 2, \dots, k\}$ representing a color, and a proper k-coloring as an assignment of colors to $\{1, 2, \dots, n\}$ such that no two consecutive integers share the same color. Let $\Omega_{k,n}$ be the set of all proper k-colorings of $[n]$.

We can generate a random element H from $\Omega_{k,n}$ using a simple recursive procedure.

{Exercise:RandomCol}

Exercise 2.3. Let $H(1)$ be a uniform sample from $[k]$. Given that $H(i)$ has been assigned for $i = 1, \dots, j-1$, choose $H(j)$ uniformly from $[k] \setminus \{H(j-1)\}$. Repeat for $j = 2, \dots, n$. Show that H is a uniform sample from $\Omega_{k,n}$.

Suppose now we want to color the nodes of the grid in figure 2.4 so that no pair of nodes separated by a single link have the same color, and we want to do this so that each proper coloring has the same chance. We describe an approximate way to do this in chapter 14.

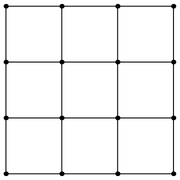


FIGURE 2.4. How can we generate a proper coloring of the nodes uniformly at random? Fig:PlaneGrid

2.6. Von Neumann unbiasing*

Suppose you have available an i.i.d. vector of *biased bits*, $X_1, X_2, \ldots, X_n$. That is, each X_k is a $\{0,1\}$-valued random variable, with $\mathbf{P}\{X_k = 1\} = p \neq 1/2$. Furthermore, suppose that we do not know the value of p. Can we convert this random vector into a (possibly shorter) random vector of independent and *unbiased* bits?

This problem was considered by Von Neumann (1951) in his work on early computers. He described the following procedure: divide the original sequence of bits into pairs, discard pairs having the same value, and for each discordant pair 01 or 10, take the first bit. An example of this procedure is shown in figure 2.5; the extracted bits are shown in the second row.

original bits	00	11	01	01	10	00	10	10	11	10	01	⋯
extracted unbiased	·	·	0	0	1	·	1	1	·	1	0	⋯
discarded bits	0	1	·	·	·	0	·	·	1	·	·	⋯
XORed bits	0	0	1	1	1	0	1	1	0	1	1	⋯

(2.8)

FIGURE 2.5. Extracting unbiased bits from biased bit stream. Fig:VN

Note that the number L of unbiased bits produced from $(X_1, \ldots, X_n)$ is itself a random variable. We denote by $(Y_1, \ldots, Y_L)$ the vector of extracted bits.

{Exercise:VonNeumann}

EXERCISE 2.4. Show that applying Von Neumann's procedure to the vector $(X_1, \ldots, X_n)$ produces a vector $(Y_1, \ldots, Y_L)$ of random length L, which conditioned on $L = m$ is uniformly distributed on $\{0,1\}^m$.

How efficient is this method? For any algorithm for extracting random bits, let N be the number of fair bits generated using the first n of the original bits. The efficiency is measured by the asymptotic rate

$$r(p) = \limsup_{n\to\infty} \frac{\mathbf{E}(N)}{n}. \tag{2.9}$$

Let $q := 1 - p$.

{Exercise:VNEfficiency}

EXERCISE 2.5. Show that for the Von Neumann algorithm, $\mathbf{E}(N) = npq$, and the rate is $r(p) = pq$.

The Von Neumann algorithm throws out many of the original bits, which in fact contain some unexploited randomness. By converting the discarded 00s and 11s to 0s and 1s, we obtain a new vector $Z = (Z_1, Z_2, \ldots, Z_{n/2-L})$ of bits. In the example shown in figure 2.5, these bits are shown on the third line.

Exercise 2.6. Prove: conditioned on $L = m$, the two vectors $Y = (Y_1, \ldots, Y_L)$ and $Z = (Z_1, \ldots, Z_{n/2-L})$ are independent, and the bits $Z_1, \ldots, Z_{n/2-L}$ are independent.

The probability that $Z_i = 1$ is $p' = p^2/(p^2 + q^2)$. We can apply the algorithm again on the independent bits Z. Given that $L = m$, Exercise 2.5 implies that the expected number of fair bits we can extract from Z is

$$\text{(length of } Z)p'q' = \left(\frac{n}{2} - m\right)\left(\frac{p^2}{p^2+q^2}\right)\left(\frac{q^2}{p^2+q^2}\right). \tag{2.10}$$

By Exercise 2.5 again, the expected value of L is npq. Hence the expected number of extracted bits is

$$n[(1/2) - pq]\left(\frac{p^2}{p^2+q^2}\right)\left(\frac{q^2}{p^2+q^2}\right). \tag{2.11}$$

Adding these bits to the original extracted bits yields a rate for the modified algorithm of

$$pq + [(1/2) - pq]\left(\frac{p^2}{p^2+q^2}\right)\left(\frac{q^2}{p^2+q^2}\right). \tag{2.12}$$

A third source of bits is obtained by taking the XOR of adjacent pairs. (The XOR of two bits a and b is 0 if and only if $a = b$.) Call this sequence $U = (U_1, \ldots, U_{n/2})$. This is given on the fourth row in figure 2.5. It turns out that U is independent of Y and Z, and applying the algorithm on U yields independent and unbiased bits. It should be noted, however, that given $L = m$, the bits in U are not independent, as it contains exactly m 1's.

Note that when the Von Neumann algorithm is applied to the sequence Z of discarded bits and to U, it creates a new sequence of discarded bits. The algorithm can be applied again to this sequence, improving the extraction rate.

Indeed, this can be continued indefinitely. This idea is developed in Peres (1992).

2.7. Problems

{Exer:CoinSimUnif}

Exercise 2.7. Check that the random variable in (2.4) has the uniform distribution on the set in (2.3).

{Exer:LCSPeriodic}

Exercise 2.8. Show that if $f : \{1, \ldots, m\} \to \{1, \ldots, m\}$ is any function, and $x_n = f(x_{n-1})$ for all n, then there is an integer k such that $x_n = x_{n+k}$ eventually. That is, the sequence is eventually periodic.

{Exer:UnifDiscDim}

Exercise 2.9. Let U be uniform on $[0, 1]$, and let X be the random variable $\phi(U)$, where ϕ is defined as in (2.6). Show that X takes on the value a_k with probability p_k.

{Exercise:NonRev}

Exercise 2.10. A nearest-neighbor path $0 = v_0, \ldots, v_n$ is *non-reversing* if $v_k \neq v_{k-2}$ for $k = 2, \ldots, n$. It is simple to generate a non-reversing path recursively. First choose v_1 uniformly at random from $\{(0,1), (1,0), (0,-1), (-1,0)\}$. Given that $v_0, \ldots, v_{k-1}$ is a non-reversing path, choose v_k uniformly from the three sites in $\mathbb{Z}^2$ at distance 1 from v_{k-1} but different from v_{k-2}.

Let Ξ_n^{nr} be the set of non-reversing nearest-neighbor paths of length n. Show that the above procedure generates a uniform random sample from Ξ_n^{nr}.

{Exer:SAW}

Exercise 2.11. One way to generate a random self-avoiding path is to generate non-reversing paths until a self-avoiding path is obtained.

(a) Let $c_{n,4}$ be the number of paths in $\mathbb{Z}^2$ which do not contain loops of length 4 at indices $i \equiv 0 \mod 4$. More exactly, these are paths $(0,0) = v_0, v_1, \ldots, v_n$ so that $v_{4i} \neq v_{4(i-1)}$ for $i = 1, \ldots, n/4$. Show that

$$c_{n,4} \leq \left[4(3^3) - 8\right]\left[3^4 - 6\right]^{\lceil n/4 \rceil - 1} \tag{2.13}$$

{Eq:NoLoops}

(b) Conclude that the probability that a random non-reversing path of length n is self-avoiding is bounded above by $e^{-\alpha n}$ for some fixed $\alpha > 0$.

{Exer:HC}

Exercise 2.12. Recall that the *Fibonacci* numbers are defined by $f_0 := f_1 := 1$, and $f_n := f_{n-1} + f_{n-2}$ for $n \geq 1$. Show that the number of configurations in the one-dimensional hard-core model with n sites is f_{n+1}.

{Exer:HC2}

Exercise 2.13. Show that the algorithm described in Example 2.2 generates a uniform sample from Ω_n.

2.8. Notes

Counting the number of self-avoiding paths is an unsolved problem. For more on this topic, see Madras and Slade (1993). Randall and Sinclair (2000) give an algorithm for approximately sampling from the uniform distribution on these walks.

For more examples of sets enumerated by the Fibonacci numbers, see Stanley (1986, Chapter 1, Exercise 14) and Graham et al. (1994, Section 6.6). Benjamin and Quinn (2003) use combinatorial interpretations to prove Fibonacci identities (and many other things).

On random numbers, Von Neumann offers the following:

> "Any one who considers arithmetical methods of producing random digits is, of course, in a state of sin." (von Neumann, 1951)

Iterating the Von Neumann algorithm asymptotically achieves the optimal extraction rate of $-p \log_2 p - (1-p)\log_2(1-p)$, the entropy of a biased random bit (Peres, 1992). Earlier, a different optimal algorithm was given by Elias (1972), although the iterative algorithm has some computational advantages.

Kasteleyn's formula (Kasteleyn, 1961) for the number of tilings of a $n \times m$ grid, when n and m are even (Example 2.3), is

$$2^{nm} \prod_{i=1}^{n/2} \prod_{j=1}^{m/2} \left(\cos^2 \frac{\pi j}{n+1} + \cos^2 \frac{\pi k}{m+1} \right). \tag{2.14}$$

Thorp (1965) proposed Exercise 2.2 as an “Elementary Problem” in the *American Mathematical Monthly*.

CHAPTER 3

Introduction to Finite Markov Chains

{Chapters:MC}

3.1. Finite Markov Chains {Sec:FinMarkChains}

A Markov chain is a system which moves among the elements of a finite set Ω in the following manner: when at $x \in \Omega$, the next position is chosen according to a fixed probability distribution $P(x, \cdot)$. More precisely, a sequence of random variables $(X_0, X_1, \ldots)$ is a *Markov chain with state space* Ω *and transition matrix* P if for each $y \in \Omega$,

$$\mathbf{P}\left\{X_{t+1} = y \mid X_0 = x_0, X_1 = x_1, \ldots, X_{t-1} = x_{t-1}, X_t = x\right\} = P(x, y) \qquad (3.1)$$ {Eq:MarkovDef}

for all $x_0, x_1, \ldots, x_{t-1}, x \in \Omega$ such that

$$\mathbf{P}\{X_0 = x_0, X_1 = x_1, \ldots, X_{t-1} = x_{t-1}, X_t = x\} > 0.$$

Here P is an $|\Omega| \times |\Omega|$ matrix whose xth row is the distribution $P(x, \cdot)$. Thus P is *stochastic*, that is, its entries are all non-negative and

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

Equation (3.1), often called the *Markov property*, means that the conditional probability of proceeding from state x to state y is the same, no matter what sequence $x_0, x_1, \ldots, x_{t-1}$ of states precedes the current state x. This is exactly why the matrix P suffices to describe the transitions.

{xmpl:frog}

Example 3.1. 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. First thing every morning, the frog decides whether to jump by

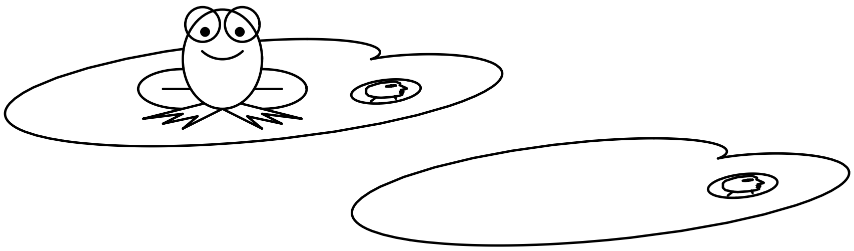


Figure 3.1. fig:frog A randomly jumping frog. Whenever he tosses heads, he jumps to the other lily pad.

tossing the current lily pad's coin. If the coin lands heads up, he jumps to the other lily pad. If the coin lands tails, he remains where he is.

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

{Eq:FrogMatrix}
$$P = \begin{pmatrix} P(e,e) & P(e,w) \\ P(w,e) & P(w,w) \end{pmatrix} = \begin{pmatrix} 1-p & p \\ q & 1-q \end{pmatrix}, \tag{3.2}$$

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 $X_t = w$.

If the frog spends Sunday on the east pad, then 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,

{eq:time1}
$$\mathbf{P}\{X_1 = e \mid X_0 = e\} = 1 - p, \qquad \mathbf{P}\{X_1 = w \mid X_0 = e\} = p. \tag{3.3}$$

What happens Tuesday? The reader should check that, by conditioning on X_1,

{eq:time2a}
$$\mathbf{P}\{X_2 = e \mid X_0 = e\} = (1-p)(1-p) + pq. \tag{3.4}$$

While we could keep writing out formulas like (3.4), there is a more systematic approach. Let's 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 (3.3) becomes $\mu_1 = \mu_0 P$.

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

{eq:frogmatmult}
$$\mu_t = \mu_{t-1} P \qquad \text{for all } t \geq 1. \tag{3.5}$$

Indeed, for any initial distribution μ_0,

{eq:froghiordtrans}
$$\mu_t = \mu_0 P^t \qquad \text{for all } t \geq 0. \tag{3.6}$$

How does the distribution μ_t behave in the long term? Figure 3.2 suggests that μ_t has a limit π (whose value depends on p and q) as $t \to \infty$. Any such limit distribution π must satisfy

$$\pi = \pi P,$$

which implies (after a little algebra)

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

If we define, for $t \geq 0$,

$$\Delta_t = \mu_t(e) - \frac{q}{p+q},$$

then the sequence (Δ_t) satisfies (c.f. Exercise 3.2)

{Eq:FrogRate}
$$\Delta_{t+1} = (1 - p - q)\Delta_t. \tag{3.7}$$

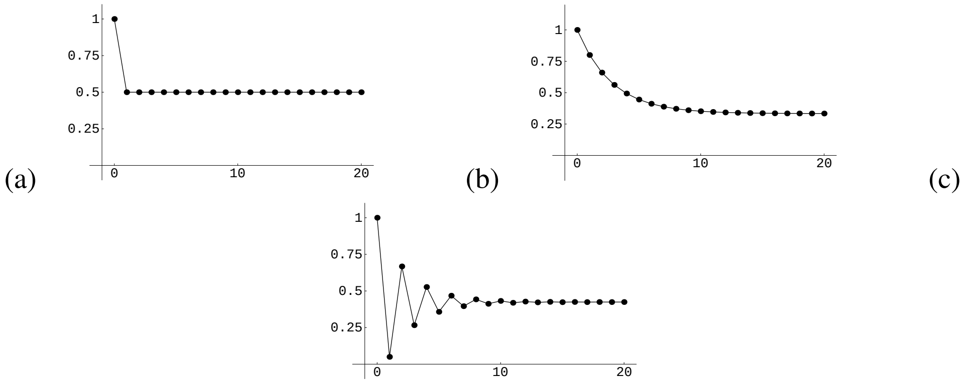


fig:limits

FIGURE 3.2. The probability of being on the east pad (started from the east pad) plotted versus time for (a) $p = q = 1/2$ (b) $p = 0.2$, $q = 0.1$ (c) $p = 0.95$, $q = 0.7$.

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

{eq:froglimit}
$$\lim_{t\to\infty} \mu_t(e) = \frac{q}{p+q} \quad \text{and} \quad \lim_{t\to\infty} \mu_t(w) = \frac{p}{p+q} \tag{3.8}$$

for any initial distribution μ_0.

The traditional theory of finite Markov chains is concerned with convergence statements of the type seen in (3.32), that is, with the rate of convergence as $t \to \infty$ for a *fixed chain*. Note that $1 - p - q$ is an eigenvalue of the frog's matrix P, and from (3.31) this eigenvalue determines the rate of convergence in (3.32):

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

As we explained in the Introduction, our focus in this book is quite different. We are studying families of chains, and we are interested in the asymptotics as the state space grows—not just as time grows.

Fortunately, the computations we just did for a 2-state chain generalize to any finite Markov chain: the distribution at time t can be found by matrix multiplication. Let $(X_0, X_1, \dots)$ be a finite Markov chain with state space Ω and transition matrix P, and let the row vector μ_t be the distribution of X_t:

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

By conditioning on the possible predecessors of the $(t+1)$-st state, we see that for all $y \in \Omega$

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

Rewriting this in vector form gives

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

and hence

$$\mu_t = \mu_0 P^t \quad \text{for } t \geq 0. \tag{3.9}$$
{Eq.Aftertsteps}

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

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

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

Using these definitions and (3.9) shows that

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

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*.)

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

What if we multiply a column vector f by P on the left? Think of f as function on the state space Ω (for the frog of Example 3.1, $f(x)$ might be the average number of flies the frog catches per day at 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 by column vector by P on the left takes us from a function to the expected value of that function tomorrow.*

{Sec:SimMC}

3.2. Simulating a Finite Markov Chain

In Chapter 2, we discussed methods for sampling from various interesting distributions on finite sets, given the ability to produce certain simple types of random variables—coin flips, or uniform samples from the unit interval, say. It is natural to ask: how can we sample from the distribution of a Markov chain which has been run for many steps?

One possible method would be to explicitly compute the vector $\mu_0 P^t$, then use one of the methods from Chapter 2 to sample from this distribution. If our state space is even moderately large, this method will be extremely inefficient, since it requires us to raise the $|\Omega| \times |\Omega|$ matrix P to a large power. There is an even more elementary problem, however: for many chains we study (and would like to simulate), we don't even know $|\Omega|$!

Fortunately, generating a trajectory of a Markov chain can be done one step at a time. Let's look at a simple example.

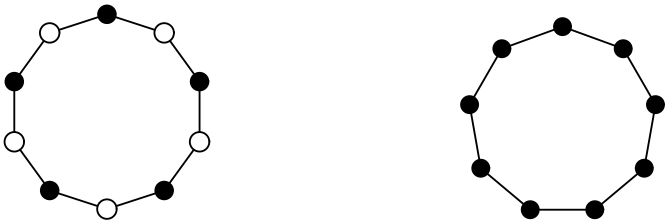


Fig:Cycle
FIGURE 3.3. Random walk on $\mathbb{Z}_{10}$ is periodic, since every step goes from an even state to an odd state, or vice-versa. Random walk on $\mathbb{Z}_9$ is aperiodic.

{Xmpl:Ncycle}

EXAMPLE 3.2 (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

$$P(j,k) = \begin{cases} 1/2 & \text{if } k \equiv j+1 \pmod{n}, \\ 1/2 & \text{if } k \equiv j-1 \pmod{n}, \\ 0 & \text{otherwise.} \end{cases} \tag{3.10}$$

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 3.3). At each time, the walker must either go one step clockwise, or one step counterclockwise.

That description in words translates neatly into a simulation method. Let $Z_1, Z_2, \dots$ be a sequence of independent and identically distributed random variables, each of which is equally likely to be $+1$ or -1. Let's require that our walker starts at 0, i.e. that $X_0 = 0$. Then for each $t \geq 0$ set

$$X_{t+1} = X_t + Z_t \bmod n. \tag{3.11}$$ {eq:randmapxmpl}

The resulting sequence of random variables $X_0, X_1, \dots$ is clearly a Markov chain with transition matrix P.

More generally, we define a *random mapping representation* of a Markov chain on state space Ω with transition matrix P to consist of a function $f : \Omega \times \Lambda \to \Omega$ such that for some sequence of independent and identically distributed random variables $Z_0, Z_1, \dots$, each of which takes values in the set Λ,

$$X_0, f(X_0, Z_0), f(X_1, Z_1), f(X_2, Z_2), \dots$$

is a Markov chain with transition matrix P. The function f takes in the current state and some new random information, and from that information determines the next state of the chain. More explicitly, if we are at state $x \in \Omega$ at time t, and our auxiliary randomness-generating device outputs $z \in \Lambda$, then the next state of the chain will be $f(x, z)$:

$$(X_t = x \text{ and } Z_t = z) \Rightarrow X_{t+1} = f(x, z).$$

In the example above, $\Lambda = \{1, -1\}$, each Z_i is uniform on Λ, and

$$f(x, z) = x + z \bmod n.$$

Proposition 3.3. *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 $Z_1, Z_2, \dots$ will be uniformly chosen in this interval. To determine the function $f : \Omega \times \Lambda \to \Omega$, we use the method of Exercise 2.9 to simulate the discrete distributions $P(x_j, \cdot)$. More specifically, set $F_{j,k} = \sum_{i=1}^{k} P(x_j, x_i)$ and define

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

■

Note that, unlike transition matrices, random mapping representations are far from unique. For instance, replacing the $f(x, z)$ in the previous proof with $f(x, 1-z)$ yields another representation.

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 6, on coupling.

3.3. Irreducibility and Aperiodicity

{Sec:IrrAper}

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 5.6) 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, Sections 4.4 and 4.7. See Section 3.7 for a discussion of all the ways in which a Markov chain can fail to be irreducible.)

The chain P will be called *aperiodic* if $\gcd\{t : P^t(x, x) > 0\} = 1$ for all $x \in \Omega$. If a chain is not aperiodic, we call it *periodic*.

If P is aperiodic and irreducible, then there is an integer r so that $P^r(x, y) > 0$ for all $x, y, \in \Omega$. (See Exercise 3.3.)

According to our definition, a chain in which all paths from x_0 to x_0 have even length is periodic. In such a chain, the states lying on x_0–x_0 paths can be split into those at even distance from x_0 and those at odd distance from x_0; all allowed transitions go from one class to the other. No matter how many steps the chain started at x_0 takes, the distribution at a particular instant will never be spread over all the states. The best we can hope for is that the distribution will alternate between being nearly uniform on the "even" states, and nearly uniform on the "odd" states. Of course, if $\gcd\{t : P^t(x,x) > 0\} > 2$, the situation can be even worse!

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.

{Xmpl:NcyclePer}

Example 3.4 (The n-cycle, revisited). Recall random walk on the n-cycle, defined in Example 3.2. 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 3.3). Random walk on an odd-length cycle is aperiodic.

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

$$Q(j,k) = \begin{cases} 1/4 & \text{if } k \equiv j+1 \pmod{n}, \\ 1/2 & \text{if } k \equiv j \pmod{n}, \\ 1/4 & \text{if } k \equiv j-1 \pmod{n}, \\ 0 & \text{otherwise.} \end{cases} \tag{3.12}$$

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

3.4. Random Walks on Graphs

{Sec:RWG}

The random walk on the n-cycle, shown in Figure 3.3, is a simple case of an important type of Markov chain.

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

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

$$P(x,y) = \begin{cases} \frac{1}{\deg(x)} & \text{if } y \sim x, \\ 0 & \text{otherwise.} \end{cases} \tag{3.13}$$

{Eq:SRW}

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 3.5. Consider the graph G shown in Figure 3.4. The transition matrix

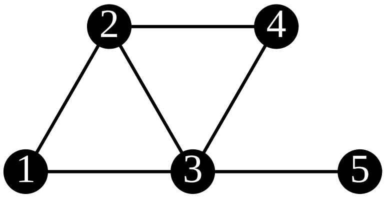


fig:SRW
FIGURE 3.4. An example of a graph with vertex set $\{1, 2, 3, 4, 5\}$ and 6 edges.

of simple random walk on G is

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

We will say much, much more about random walks on graphs throughout this book—but especially in Chapter 10.

{Sec:StatDist}

3.5. Stationary Distributions

3.5.1. Definition. We saw in Example 3.1 that a distribution π on Ω satisfying

{Eq:StationaryEq}

$$\pi = \pi P \tag{3.14}$$

can have another interesting property: in that case, π was the long-term limiting distribution of the chain. We call a probability π satisfying (3.14) a *stationary distribution* of the Markov chain. Clearly, if π is a stationary distribution and $\mu_0 = \pi$ (i.e. the chain is started in a stationary distribution), then $\mu_t = \pi$ for all $t \geq 0$.

Note that we can also write (3.14) elementwise: an equivalent formulation is

{Eq:StationarySystem}

$$\pi(y) = \sum_{x \in \Omega} \pi(x) P(x, y) \quad \text{for all } y \in \Omega. \tag{3.15}$$

{Example:PiForSRW}

EXAMPLE 3.6. Consider simple random walk on a graph $G = (V, E)$. For any vertex $y \in V$,

$$\sum_{x \in V} \deg(x) P(x, y) = \sum_{x \sim y} \frac{\deg(x)}{\deg(x)} = \deg(y). \tag{3.16}$$

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

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

the probability measure proportional to the degrees, is always a stationary distribution for the walk. For the graph in Figure 3.4,

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

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

Our goal for the rest of this chapter and the next is to prove a general yet precise version of the statement that "finite Markov chains converge to their stationary distributions." 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 many other places.

{Sec:FirstReturn}

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

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

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

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

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

{lem:firstreturnintegra

Lemma 3.7. *For any states x and y of an irreducible aperiodic chain, $\mathbf{E}_x(\tau_y^+) < \infty$.*

Proof. By Exercise 3.3, there exists an r such that every entry of P^r is positive. Let $\varepsilon = \min_{z,w \in \Omega} P^r(z, w)$ be its smallest entry. No matter the value of X_t, the probability of hitting state y at time $t+r$ is at least ε. Thus, for $k \geq 0$, the probability that the chain has *not* arrived at y by time kr is no larger than the probability that k independent trials, each with success probability ε, all fail:

$$\mathbf{P}_x\{\tau_y^+ > kr\} \leq \mathbf{P}_x\{X_r \neq y, X_{2r} \neq y, \dots, X_{kr} \neq y\} \leq (1 - \varepsilon)^k. \tag{3.17}$$ {eq:everyrth}

See Exercise 3.12 to complete the proof. ■

3.5.3. Existence of a stationary distribution. The Convergence Theorem (Theorem 5.6 below) implies that the "long-term" fractions of time a finite aperiodic Markov chain spends in each state coincide with the chain's stationary distribution. We, however, 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.

{Prop:PiExists}

Proposition 3.8. *Let P be the transition matrix of an irreducible Markov chain. Then there exists a probability distribution π on Ω such that $\pi = \pi P$.*

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

$$\begin{aligned}\tilde{\pi}(y) &:= \mathbf{E}_z(\text{number of visits to } y \text{ before returning to } z)\\ &= \sum_{t=0}^{\infty} \mathbf{P}_z\{X_t = y, \tau_z^+ > t\}.\end{aligned} \tag{3.18}$$

{eq:pitildedefn}

By Exercise 3.13, $\tilde{\pi}(y) < \infty$ for all $y \in \Omega$. Let's try checking whether $\tilde{\pi}$ is stationary, starting from the definition:

{eq:tildesum}

$$\sum_{x\in\Omega} \tilde{\pi}(x)P(x,y) = \sum_{x\in\Omega}\sum_{t=0}^{\infty} \mathbf{P}_z\{X_t = x, \tau_z^+ > t\}P(x,y). \tag{3.19}$$

Now reverse the order of summation in (3.19). After doing so, we can use the Markov property to compute the sum over x. Essentially we are shifting by one the time slots checked, while at the same time shifting the state checked for by one step of the chain—from x to y:

{eq:almostthere}

$$\begin{aligned}\sum_{t=0}^{\infty}\sum_{x\in\Omega} \mathbf{P}_z\{X_t = x, \tau_z^+ \geq t+1\}P(x,y) &= \sum_{t=0}^{\infty} \mathbf{P}_z\{X_{t+1} = y, \tau_z^+ \geq t+1\} \qquad (3.20)\\ &= \sum_{t=1}^{\infty} \mathbf{P}_z\{X_t = y, \tau_z^+ \geq t\}. \qquad (3.21)\end{aligned}$$

The expression in (3.21) is very similar to (3.18), so we're almost done. In fact,

q:negligibledifference}

$$\begin{aligned}\sum_{t=1}^{\infty} \mathbf{P}_z\{X_t = y, \tau_z^+ \geq t\} &= \tilde{\pi}(y) - \mathbf{P}_z\{X_0 = y, \tau_z^+ > 0\} + \sum_{t=1}^{\infty} \mathbf{P}_z\{X_t = y, \tau_z^+ = t\} \qquad (3.22)\\ &= \tilde{\pi}(y) - \mathbf{P}_z\{X_0 = y\} + \mathbf{P}_z\{X_{\tau_z^+} = y\}. \qquad (3.23)\end{aligned}$$

Now consider two cases:

$y = z$: Since $X_0 = z$ and $X_{\tau_z^+} = z$, the two last terms of (3.23) are both 1, and they cancel each other out.

$y \neq z$: Here both terms are 0.

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

{Eq:pi}

$$\pi(x) = \frac{\tilde{\pi}(x)}{\mathbf{E}_z(\tau_z^+)} \quad \text{satisfies } \pi = \pi P. \tag{3.24}$$

■

Remark. The computation at the heart of the proof of Proposition 3.8 can be generalized. The argument we give above works whenever $X_0 = z$ is a fixed state and the stopping time τ satisfies both $\mathbf{P}_z\{\tau < \infty\} = 1$ and $\mathbf{P}_z\{\tau = z\} = 1$.

{Sec:StatUnique}

3.5.4. Uniqueness of the stationary distribution.

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

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

$$h(x) = \sum_{y \in \Omega} P(x, y) h(y). \tag{3.25}$$ {Eq:HarmonicDefn}

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

{Lem:Liouville}

Lemma 3.9. *A function h which is harmonic at every point of Ω is constant.*

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

$$h(x_0) = P(x_0, z)h(z) + \sum_{y \neq z} P(x_0, y)h(y) < M, \tag{3.26}$$

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

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

{Cor:StatDistUnique}

Corollary 3.10. *Let P be the transition matrix of an irreducible Markov chain. There exists a unique probability distribution π satisfying $\pi = \pi P$.*

Proof. While proving Proposition 3.8, we constructed one such measure. Lemma 3.9 implies that the kernel of $P - I$ has dimension 1, so the column rank of $P - I$ is $|\Omega| - 1$. The row rank equals column rank (equals rank), so 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. Another proof of Corollary 3.10 follows from the Convergence Theorem (Theorem 5.6, proved below).

3.6. Reversibility and time reversals

{Sec:Reversibility}

Suppose a probability π on Ω satisfies

$$\pi(x)P(x, y) = \pi(y)P(y, x). \tag{3.27}$$ {Eq:DetailedBalance}

Exercise 3.22 asks you to check that π is then stationary for P. Furthermore, when (3.27) holds,

$$\pi(x_0)P(x_0, x_1) \cdots P(x_{n-1}, x_n) = \pi(x_n)P(x_n, x_{n-1}) \cdots P(x_1, x_0). \tag{3.28}$$ {Eq:Reversed}

We can rewrite (3.28) in the following suggestive form:

$$\mathbf{P}_\pi\{X_0 = x_0, \ldots, X_n = x_n\} = \mathbf{P}_\pi\{X_0 = x_n, X_1 = x_{n-1}, \ldots, X_n = x_0\}, \tag{3.29}$$ {Eq:Reversed2}

In words: when a chain satisfying (3.27) is run in stationarity, the distribution of finite segments of trajectory is the same no matter whether we run time backwards or forwards. For this reason, a chain satisfying (3.27) is called *reversible*. The equations (3.27) are called the *detailed balance* equations.

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

{Eq:ReversedMatrix}

$$\widehat{P}(x, y) := \frac{\pi(y)P(y, x)}{\pi(x)}. \tag{3.30}$$

Exercise 7.6 shows that the terminology "time-reversal" is reasonable. (Note that when a chain is reversible, as defined in Section 3.6, then $\widehat{P} = P$.)

3.7. Classifying the States of a Markov Chain*

{sec:classification}

We will occasionally need to study chains which are *not* irreducible—see, for instance, Sections 4.1, 4.2 and 4.3.3. 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 Ω. Given $x, y \in \Omega$, we say that x *sees* y, and write $x \to y$, if there exists an $r > 0$ such that

$P^r(x, y) > 0$. That is, x sees y if it's possible for a trajectory of the chain to proceed from x to y. We say that x *communicates with* y, and write $x \leftrightarrow y$, if and only if $x \to y$ and $y \to x$.

The equivalence classes under $\leftrightarrow$ are called *communication classes*. For $x \in \Omega$, let $[x]$ denote the communication class of x.

Example 3.11. When P is irreducible, all the states of the chain lie in a single communication class.

Example 3.12. When a communication class consists of a single state $z \in \Omega$, it follows that $P(z, z) = 1$ and we call z an *absorbing* state. Once a trajectory arrives at z, it is "absorbed" there and can never leave.

It follows from Exercise 3.24(c) that every chain trajectory follows a weakly increasing path in the partial order on communication classes. Once the chain arrives in a class that is maximal in this order, it stays there forever. See Exercise 18.8, which connects this structure to the concepts of *recurrence* and *transience* defined in Chapter 18.

3.8. Problems

{Exer:frogstate}

Exercise 3.1. Can you tell what time of day is shown in Figure 3.1? What are the frog's plans? [Solution]

{ex:froglimit}

Exercise 3.2. Consider the jumping frog chain of Example 3.1, whose transition matrix is given in (3.2). Assume that our frog begins hopping from an arbitrary distribution μ_0 on $\{e, w\}$.

(a) Define, for $t \geq 0$,

$$\Delta_t = \mu_t(e) - \frac{q}{p+q}.$$

Show that

$$\Delta_{t+1} = (1-p-q)\Delta_t. \tag{3.31}$$ {Eq:FrogRate}

(b) Conclude that when $0 < p < 1$ and $0 < q < 1$,

$$\lim_{t\to\infty} \mu_t(e) = \frac{q}{p+q} \quad \text{and} \quad \lim_{t\to\infty} \mu_t(w) = \frac{p}{p+q} \tag{3.32}$$ {eq:froglimit}

for any initial distribution μ_0.

{Exer:Aperiodic}

Exercise 3.3. Show that when P is aperiodic and irreducible, there exists an integer r such that $P^r(x,y) > 0$ for all $x, y \in \Omega$.

{ex:oddcycle}

Exercise 3.4. 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.

{Exer:Connected}

Exercise 3.5. A graph G is *connected* when any two vertices x and y of G can be connected by a path $x = x_0, x_1, \ldots, x_k = y$ of vertices such that $x_i \sim x_{i+1}$, for $0 \leq i \leq k-1$. Show that random walk on G is irreducible if and only if G is connected.

{Exer:TreeTFAE}

Exercise 3.6. 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 $m = n - 1$.

{Exer:TreeBasics}

Exercise 3.7. Let T be a tree.

(a) Prove that T contains a *leaf*, that is, a vertex of degree 1.
(b) Prove that between any two vertices in T there is a unique path.

{Exer:Tree3ColIrr}

Exercise 3.8. 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. [Solution]

{Exer:PermParity}

Exercise 3.9. Consider the following natural (if apparently slow) method of shuffling cards: at each point in time, a pair of distinct cards is chosen, and the positions of those two cards are switched. Mathematically, this corresponds to the following Markov chain: make the state space $\Omega = S_n$, the set of all permutations of $[n]$, and set

$$P(\sigma_1, \sigma_2) = \begin{cases} 1/\binom{n}{2} & \sigma_2 = \sigma_1(ij) \text{ for some transposition } (ij), \\ 0 & \text{otherwise.} \end{cases}$$

(a) Show that this Markov chain is irreducible, but periodic.

1	2	3	4
5	6	7	8
9	10	11	12
13	15	14	

Fig:fifteen
FIGURE 3.5. The "fifteen puzzle".

(b) Modify the shuffling technique so that the two cards to be exchanged are chosen independently and uniformly at random (and if the same card is chosen twice, nothing is done to the deck). Compute the transition probabilities for the modified shuffle, and show that it is both irreducible and aperiodic.

{Exer:fifteen}

EXERCISE 3.10. The long-notorious Sam Loyd "fifteen puzzle" is shown in Figure 3.5. It consists of 15 tiles, numbered with the values 1 through 15, sitting in a 4 by 4 grid; one space is left empty. The tiles are in order, except that tiles 14 and 15 have been switched. The only allowed moves are to slide a tile adjacent to the empty space into the empty space.

Is it possible, using only legal moves, to switch the positions of tiles 14 and 15, while leaving the rest of the tiles fixed?

(a) Show that the answer is "no."

(b) Describe the set of all configurations of tiles that can be reached using only legal moves.

[SOLUTION]

{Exer:SymmTransMat}

EXERCISE 3.11. Let P be a transition matrix satisfying $P(x, y) = P(y, x)$ for all states $x, y \in \Omega$. Show that the uniform distribution on Ω is stationary for P.

:FirstReturnIntegrable}

EXERCISE 3.12.

(a) Prove that if Y is a positive integer-valued random variable, then $\mathbf{E}(Y) = \sum_{t \geq 0} \mathbf{P}\{Y > t\}$.

(b) Use (a) and (3.17) to finish the proof of Lemma 3.7.

[SOLUTION]

{Exer:RetTimeIrr}

EXERCISE 3.13. Prove that if P is irreducible (but not necessarily aperiodic), then $\mathbf{E}_x(\tau_y^+) < \infty$. [SOLUTION]

{Exer:TwoStepsRev}

EXERCISE 3.14. Let P be a transition matrix which is reversible with respect to the probability distribution π on Ω. Show that the transition matrix P^2 corresponding to two steps of the chain is also reversible with respect to π. [SOLUTION]

{Exer:StatDistPos}

EXERCISE 3.15. Let π be a stationary distribution for an irreducible transition matrix P. Prove that $\pi(x) > 0$ for all $x \in \Omega$.

EXERCISE 3.16. Check carefully that equation (3.18) is true.

{ex:periodicstatdist}

EXERCISE 3.17. Let P be the transition matrix of a chain and let $Q = \frac{I+P}{2}$.

(a) Show that for any distribution μ on Ω, $\mu = \mu P$ if and only if $\mu = \mu Q$.
(b) Show that P has a unique stationary distribution if and only if Q does.

{Exer:BolzWeierStatDist

Exercise 3.18. Here we outline another proof, more analytic, of the existence of stationary distributions. Let P be the transition matrix of a Markov chain on state space Ω. For an arbitrary initial distribution μ on Ω and $n > 0$, define the distribution ν_n by

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

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

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

(b) Show that there exists a subsequence $(\nu_{n_k})_{k\geq 0}$ such that $\lim_{k\to\infty} \nu_{n_k}(x)$ exists for every $x \in X$.
(c) For $x \in \Omega$, define $\nu(x) = \lim_{k\to\infty} \nu_{n_k}(x)$. Show that ν is a stationary distribution for P.

[Solution]

Exercise 3.19. Let P be the transition matrix of a Markov chain with state space X. Let $\Delta \subset X$ be a subset of the state space, and assume $h : \Omega \to \mathbb{R}$ is a function harmonic at all states $x \notin \Delta$.

Prove that if $h(y) = \max_{x\in\Omega} h(x)$, then $y \in \Delta$. (Note: this is a discrete version of a *maximum principle*.)

{Exer:RetTime}

Exercise 3.20. Show that for any state x of an irreducible chain, $\pi(x) = \frac{1}{E_x(\tau_x^+)}$.

Exercise 3.21. Check that for any graph G, the simple random walk on G defined by (3.13) is reversible.

{Exer:RevImpliesStat}

Exercise 3.22. Show that when π satisfies (3.27), then π also satisfies (3.14), i.e. π is stationary for P.

The following exercises concern the material in Section 3.7.

{Exer:ClassEquiv}

Exercise 3.23. Show that $\leftrightarrow$ is an equivalence relation on Ω.

{Exer:ClassPartialOrder

Exercise 3.24. The relation "sees" can be lifted to communication classes by defining $[x] \to [y]$ if and only if $x \to y$.

(a) Show that $\to$ is a well-defined relation on the communication classes.
(b) Show that $\to$ is a partial order on communication classes.
(c) Show that if, in some trajectory (X_t) of the underlying Markov chain, $X_r = x$ and $X_s = y$, where $r < s$, then $[x] \to [y]$.

Remark. It is certainly possible for the partial order constructed in Exercise 3.24(b) above to be trivial, in the sense that no class can see any other! In this case the underlying Markov chain consists of non-interacting sets of mutually communicating states; any trajectory is confined to a single communication class.

3.9. Notes

The right-hand side of (3.1) does not depend on t either. We take this as part of the definition of a Markov chain; be warned that other authors sometimes single this out 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 these notes.)

Aldous and Fill (in progress, Chapter 2, Proposition 4) present a version of the key computation for Proposition 3.8 which requires only that the chain be started in the same *distribution* as the stopping time ends. 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).

CHAPTER 4

Some Interesting Markov Chains

{Ch:Classic}

Here we present several basic and important examples of Markov chains. Each chain results from a situation that occurs often in other problems, and 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).

4.1. Gambler's Ruin

{Sec:Gambler}

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, 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.

This situation can be modeled by a random walk on a path with vertices $\{0, 1, \ldots, n\}$. At all interior vertices, the walk is equally likely to go up by 1 or down by 1. Once it arrives at 0 or n, however, it stays forever. In the language of Section 3.7, the states 0 and n are *absorbing*.

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? And what are the probabilities of the two possibilities?

{Prop:Gambler}

Proposition 4.1. *Assume that a gambler making fair unit bets on coin flips will abandon the game when his fortune falls to 0 or rises to n. Let X_t be gambler's fortune at time t and let τ be the time required to be absorbed at one of 0 or n. Assume that $X_0 = k$, where $0 \leq k \leq n$. Then:*

$$\mathbf{E}_k(\tau) = k(n-k), \tag{4.1}$$ {Eq:GRExpTime}

$$\mathbf{P}_k\{X_\tau = n\} = k/n. \tag{4.2}$$ {Eq:GRProb}

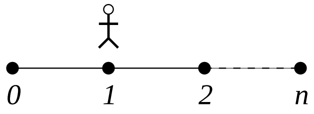


Figure 4.1. How long until the walker reaches either 0 or n? And what is the probability of each? Fig:GamblersRuin

Proof. To solve for the value $\mathbf{E}_k(\tau)$ for a specific k, it is easiest to consider the problem of finding the values $\mathbf{E}_k(\tau)$ for all $k = 0, 1, \dots, n$. To this end, write f_k for the expected time $\mathbf{E}_k(\tau)$ started at position k. Clearly, $f_0 = f_n = 0$; the walk is started at one of the absorbing states. For $1 \le k \le n-1$, it's true that

{Eq:GRR}

$$f_k = \frac{1}{2}(1 + f_{k+1}) + \frac{1}{2}(1 + f_{k-1}) \tag{4.3}$$

Why? When the first step of the walk increases the gambler's fortune, then the conditional expectation of τ is 1 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 4.1 asks you to solve this system of equations, completing the proof of Equation 4.1.

Equation 4.2 is even simpler. Again we try to solve for all the values at once. Let p_k be the probability that the gambler reaches a fortune of n before ruin, given that she starts with k dollars. Then $p_0 = 0$ and $p_n = 1$, while

{Eq:GamblerResult}

$$p_k = \frac{1}{2}p_{k-1} + \frac{1}{2}p_{k+1}, \text{ for } 1 \le k \le n-1. \tag{4.4}$$

Why? If the gambler is at one end or the other, she stays there—the outcome never changes. If she's in between, then the result of the next bet is equally likely to increase her fortune by 1, or decrease it by 1.

Clearly the values p_k must be evenly spaced between 0 and 1, and thus $p_k = k/n$. ■

Remark. See Chapter 10 for powerful generalizations of the simple methods we have just applied.

4.2. Coupon Collecting

{Sec:CouponCollecting}

A card company issues baseball cards, each featuring a single player. There are n players total, and a collector desires a complete set. We suppose each card he acquires is equally likely to be each of the n players. How many cards must he obtain so that his collection contains all n players?

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

{Prop:CouponExpected}

Proposition 4.2. *Consider a collector attempting to collect a complete set of cards. Assume that each new card is chosen uniformly and independently from the set of n possible types, and let τ be the (random) number of cards collected when the set first contains every type. Then*

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

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

$$\tau = \tau_n = \tau_1 + (\tau_2 - \tau_1) + \cdots + (\tau_n - \tau_{n-1}). \tag{4.5}$$

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

$$\mathbf{E}(\tau) = \sum_{k=1}^{n} \mathbf{E}(\tau_k - \tau_{k-1}) = n \sum_{k=1}^{n} \frac{1}{n-k+1} = n \sum_{k=1}^{n} \frac{1}{k}. \tag{4.6}$$ {Eq:CCExp}

■

While Proposition 4.2 is simple and vivid—you should not forget the argument!—we will generally need to know more in about the distribution of τ in future applications. Recall that $\sum_{k=1}^{n} \frac{1}{k} \approx \log n$ (see Exercises 4.5 for more detail). The following estimate says that T is unlikely to be much larger than its expected value.

{Prop:CouponTail}

Proposition 4.3. *Let τ be a coupon collector random variable, as in Proposition 4.2. Then for any $c > 0$*

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

Proof. Let A_i be the event that the ith player does not appear among the first $n \log n + cn$ cards drawn. Then

$$\begin{aligned} \mathbf{P}\{\tau > n \log n + cn\} &= \mathbf{P}\left(\bigcup_{i=1}^{n} A_i\right) \leq \sum_{i=1}^{n} \mathbf{P}(A_i) \\ &= \sum_{i=1}^{n} \left(1 - \frac{1}{n}\right)^{n \log n + cn} \leq n \exp\left(-\frac{n \log n + cn}{n}\right) = e^{-c}. \end{aligned} \tag{4.7}$$ {Eq:CouponTail}

■

4.3. Urn Models

{Sec:Urns}{Sec:BLurn}

4.3.1. The Bernoulli-Laplace model.

{Sec:Ehrenfest}

4.3.2. The Ehrenfest urn model and the hypercube. Suppose n balls are distributed among two urns, I and II. At each move, a ball is selected 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

{Eq:EhrenTM}
$$P(j,k) = \begin{cases} \frac{n-j}{n} & \text{if } k = j+1, \\ \frac{j}{n} & \text{if } k = j-1, \\ 0 & \text{otherwise.} \end{cases} \tag{4.8}$$

Thus, the chain lives on $\Omega = \{0,1,2,\dots,n\}$, moving by ± 1 on each move, and biased towards the middle of the interval.

Exercise 4.6 asks you to check that the stationary distribution is binomial with parameters n and $1/2$.

The Ehrenfest urn is a projection of the random walk on the n-dimensional hypercube. The n-dimensional *hypercube* is the graph which has vertex set $\{0,1\}^n$ and has edges connecting vectors which differ in exactly one coordinate. See Figure 4.2 for an illustration of the 3-dimensional hypercube.

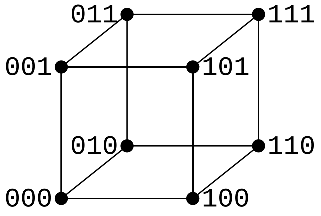


FIGURE 4.2. The 3-dimensional hypercube. Fig:HypercubeA

The simple random walk on $\{0,1\}^n$ 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 $(x^1, \dots, 1-x^j, \dots, x^n)$. That is, the bit at the chosen coordinate is flipped.

It will be convenient to often consider instead the *lazy* random walker. This walker remains at its current position with probability $1/2$, and moves 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 everything.

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

{Eq:HammingDefn}
$$W(\boldsymbol{x}) = \sum_{j=1}^{n} x^j. \tag{4.9}$$

Let $(\boldsymbol{X}_t)$ be the simple random walk on $\{0,1\}^n$, and let $W_t = W(\boldsymbol{X}_t)$ be the Hamming weight of the walker 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 it is clear that (W_t) is a Markov chain with transition probabilities given by (4.8).

{Sec:Polya}

4.3.3. 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 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 $\{1, 2, \ldots\}^2$.

{Lem:PUUniform}

Lemma 4.4. *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* $\{1, 2, \ldots, k + 1\}$.

Proof. Let $U_0, U_1, \ldots, U_n$ be independent and identically distributed random variables, each uniformly distributed on the interval $[0, 1]$. Let L_k be the number of $U_1, U_2, \ldots, U_k$ which lie to the left of U_0.

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

$$\mathbf{P}\{L_k = j - 1,\ L_{k+1} = j\} = \frac{j(k!)}{(k + 2)!} = \frac{j}{(k + 2)(k + 1)}. \tag{4.10}$$ {Eq:JointLk}

Clearly $\mathbf{P}\{L_k = j - 1\} = 1/(k + 1)$, which with (4.10) shows that

$$\mathbf{P}\{L_{k+1} = j \mid L_k = j - 1\} = \frac{j}{k + 2}. \tag{4.11}$$ {Eq:Lk1}

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

$$\mathbf{P}\{L_{k+1} = j - 1 \mid L_k = j - 1\} = \frac{k + 2 - j}{k + 2}. \tag{4.12}$$ {Eq:Lk2}

Equation 4.11 and Equation 4.12 show that the sequences $(L_k + 1)_{k=1}^n$ and $(B_k)_{k=1}^n$ have the same distribution; in particular, $L_k + 1$ and B_k have the same distribution.

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

4.4. Random Walks on Groups

{Sec:RWGroups}

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

(i) $e \cdot g = g$ and $g \cdot e = g$, and

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

{Xmpl:SnCycleNot}

Example 4.5. The set $\mathcal{S}_n$ of all permutations of the standard n-element set $\{1, 2, \dots, n\}$, introduced in Section 2.4, forms a group under the operation of functional composition. The identity element of $\mathcal{S}_n$ is the identity function $\mathrm{id}(k) = k$. Every $\sigma \in \mathcal{S}_n$ has a well-defined inverse function, which is its inverse in the group.

We will sometimes find it convenient to use *cycle notation* for permutations. In this notation, a string such as (abc) refers to the permutation which sends the element a to b, the element b to c, and the element c to a. When several cycles are written consecutively, they are performed one at a time, *from right to left* (as is consistent with ordinary function composition). For example,

$$(13)(12) = (123)$$

and

$$(12)(23)(34)(23)(12) = (14).$$

Given a probability measure μ on a group $(G, \cdot)$, we can define a *random walk on G with increment distribution* μ 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 μ. Equivalently, the transition matrix P of this chain has entries

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

Example 4.6 (The n-cycle). Let μ assign probability $1/2$ to each of to 1 and $n - 1 \equiv -1 \pmod{n}$ in the additive cyclic group $\mathbb{Z}_n = \{0, 1, \dots, n-1\}$. Then the *simple random walk on the n-cycle* first introduced in Example 3.2 is the random walk on $\mathbb{Z}_n$ with increment μ. Similarly, if ν assigns 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 3.4, is the random walk on $\mathbb{Z}_n$ with increment ν.

Example 4.7 (The hypercube). The hypercube random walks discussed in Section 4.3.2 can be viewed as a 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 measure is uniform on the set $\{\mathbf{e}_i \mid 1 \leq i \leq n\}$, where $\mathbf{e}_i$ has a 1 in the i-th place and 0 in all other entries. For the lazy version, the increment measure gives the vector $\mathbf{0}$ (with all zero entries) weight $1/2$ and each $\mathbf{e}_i$ weight $1/2n$.

Remark. We are multiplying the current state by the increment *on the left* because it is often more natural in the symmetric group, which is our most important non-commutative example. (For commutative examples, such as $\mathbb{Z}_n$ or the hypercube, it of course doesn't matter on which side we multiply.)

{Prop:RWGStat}

Proposition 4.8. *Let P be the transition matrix of a random walk on a finite group G. Then the uniform measure on G is a stationary distribution for P.*

Proof. Let μ be the increment distribution of the random walk, and let U denote the uniform measure on G. For any $g \in G$,

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

For the first equality, we reindexed by setting $k = gh^{-1}$. (The key point is that, just as it is possible to step away from g using any element in the support of μ, it is possible to arrive at G using any element in the support of μ.) ■

4.4.1. Generating sets and irreducibility. 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 4.9. *Let μ be a probability measure on a finite group G. The random walk on G with increments μ is irreducible if and only if S=$\{g \in G \,|\, \mu(g) > 0\}$ generates G.*

Proof. When the random walk is irreducible, then for any $a, b \in G$ there exists an $r > 0$ s.t. $P^r(a, b) > 0$. In order for this to occur, there must be a sequence $s_1, \dots, s_r \in G$ such that $b = s_r s_{r-1} \dots s_1 a$ and $g_i \in S$ for $i = 1, \dots, r$.

Now assume S generates G, and consider $a, b \in G$. Let $g = ba^{-}1$. We know that g can be written as a word in the elements of S and their inverses. Since every element of G has finite order, any inverses appearing in the expression for g can be rewritten as positive powers of elements of S. If the resulting expression is $g = s_m s_{m-1} \dots {}_1$ where $s_i \in S$ for $i = 1, \dots, m$, then

$$P^m(a, b) = P^m(a, ga) = P(a, s_1 a)P(s_1 a, s_2 s_1 a) \dots P(s_{m-1} \dots {}_1\, a, ga) > 0.$$

■

Let G be a group and let J be a set which generates G. The *directed Cayley graph* associated to G and J is the directed graph with vertex set G in which (v, w) is an edge if and only if $v = gw$ for some generator $g \in J$.

We call a set J of generators of G *symmetric* if $g \in J$ implies $g^{-1} \in J$. When J is symmetric, all edges in the directed Cayley graph are bidirectional, and it may be viewed as an ordinary graph.

Example 4.10 (Random transpositions, version 1). A *transposition* is an element of $\mathcal{S}_n$ that interchanges two elements and leaves all others fixed. Let $T \subseteq \mathcal{S}_n$ be the set of all transpositions. In Section 2.4, we gave a method for generating a uniform random permutation that started with the sorted sequence and used only transpositions. Hence $\langle T \rangle = S_n$, and the corresponding random walk is irreducible.

Suppose that G is finite with generators $\{g_1, \dots, g_n\}$. The simple random walk on the Cayley graph of G is the random walk on G with μ taken to be the uniform distribution on the generators.

{Sec:PermParity}

4.4.2. Parity of permutations and periodicity. For contrast, consider the set T' of all three-cycles in $\mathcal{S}_n$. The set T' does *not* generate all of $\mathcal{S}_n$, but we must introduce an important property of the permutation group $\mathcal{S}_n$ to see why. Given a permutation $\sigma \in \mathcal{S}_n$, consider the sign of the product

$$M(\sigma) = \prod_{1 \le i < j \le n} (\sigma(j) - \sigma(i)) \,.$$

Clearly $M(\mathrm{id}) > 0$, since every term is positive. For every $\sigma \in \mathcal{S}_n$ and every transposition (ab), we have

$$M((ab)\sigma) = -M(\sigma).$$

Why? We may assume that $a < b$. Then for every c such that $a < c < b$, two factors change sign, while the single factor containing both a and b also changes sign.

Call a permutation σ *even* if $M(\sigma) > 0$, and otherwise call σ *odd*. Note that a permutation is even (odd) if and only if every way of writing it as a product of transpositions contains an even (odd) number of factors. Furthermore, under composition of permutations, evenness and oddness follow the same rules as they do for integer addition. Hence the set of all even permutations in S_n forms a subgroup, known as the *alternating group A_n*.

4.4.3. Reversibility and random walks on groups.

{Sec:Transitive}

4.4.4. Transitive chains. A Markov chain is called *transitive* if for each pair $(x, y) \in \Omega \times \Omega$ there is a function $\phi = \phi_{(x,y)}$ mapping Ω to itself such that

$$\phi(x) = y \quad \text{and} \quad P(z, w) = P(\phi(z), \phi(w)). \tag{4.13}$$

Roughly, this mean the chain "looks the same" from any point in the state-space Ω.

4.5. 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 exactly, if (Δ_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 (Δ_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 called the *simple random walk*, and if $p = q = 1/4,\ r = 1/2$ the chain is called the *lazy simple random walk*.

{Thm:SRWHitBound}

Theorem 4.11. *If (X_t) is the simple random walk on $\mathbb{Z}$ and τ_0 is the first time that the walk visits 0, then for $k = 1, 2, \ldots$,*

{Eq:SRWHitBound}

$$\mathbf{P}_k\{\tau_0 > r\} \leq \frac{12k}{\sqrt{r}} \tag{4.14}$$

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

{Lem:RP0}

Lemma 4.12 (Reflection Principle). *Let (X_t) be either the simple random walk or the lazy simple random walk on $\{-B, \ldots, B\}$, and let*

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

be the first time when the walk hits 0. *For* $k \in \{1, 2, \ldots\}$,

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

Summing over $j \geq 1$ *shows that*

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

Proof. The walker "starts afresh" from 0 when he hits 0, meaning that the walk viewed from the first time it hits zero has the same distribution as a walk started from zero and is independent of the past. From this, for $j = 1, 2, \ldots$,

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

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

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

Summing over $s < r$,

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

The last equality follows since the random walk must past through 0 before hitting a negative integer. ■

Remark 4.1. There is also a simple combinatorial proof of Lemma 4.12. 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 4.3: to obtain a bijection from the former set of paths to the latter set, reflect a path after the first time it hits 0.

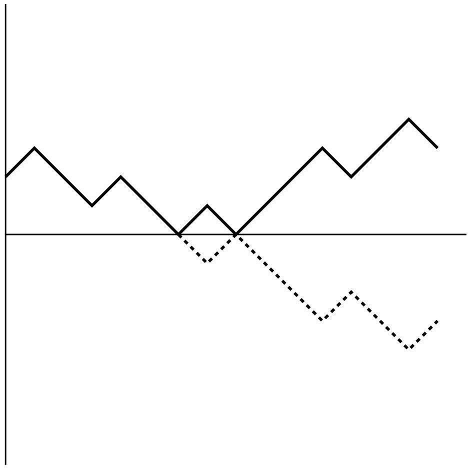


Figure 4.3. A path hitting zero and ending above zero can be transformed, by reflection, into a path ending below zero. Fig:RP

Example 4.13 (First passage time for simple random walk). A nice application of Lemma 4.12 gives the distribution of τ_0 when starting from 1. We have

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

The second to the last equality follows since the conditional probability of hitting 0 at time $2m+1$, given that at time $2m$ the walker is at 1 and has not previously visited 0, is simply the probability of moving from 1 to 0 in one move (by the Markov property). Rewriting and using Lemma 4.12 yields

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

Calculating using the Binomial distribution shows that

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

{Lem:RP1}

Lemma 4.14. *For simple random walk or lazy simple random walk (X_t) on $\mathbb{Z}$, for $k = 1, 2, \ldots$,*

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

Proof. We can write

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

By Lemma 4.12,

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

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

$$\mathbf{P}_k\{\tau_0 > r\} = \mathbf{P}_k\{X_r > 0\} - \mathbf{P}_k\{X_r > 2k\} = \mathbf{P}_k\{0 < X_r \le 2k\} = \mathbf{P}_0\{-k < X_r \le k\}.$$

■

{Lem:RP2}

Lemma 4.15. *For the simple random walk (X_t) on $\mathbb{Z}$,*

{Eq:SRWStirling}

$$\mathbf{P}_0\{X_t = k\} \le \frac{3}{\sqrt{t}}. \tag{4.15}$$

Remark 4.2. By applying Stirling's formula a bit more carefully than we do in the proof below, one can see that in fact

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

when is k not too far away from 0. 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 (4.15) 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 $k = 0, 1, \ldots, r$,

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

By Stirling's Formula (use the bounds $1 \leq e^{1/(12n+1)} \leq e^{1/(12n)} \leq 2$ in Equation B.11), we obtain the bound

$$\mathbf{P}_0\{X_{2r} = 2k\} \leq \sqrt{\frac{8}{\pi}} \frac{1}{\sqrt{2r}}. \tag{4.16}$$ {Eq:BoundEven}

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

$$\mathbf{P}_0\{X_{2r+1} = 2k + 1\} \leq \frac{4}{\sqrt{\pi}} \frac{1}{\sqrt{2r+1}}. \tag{4.17}$$ {Eq:BoundOdd}

Putting together (4.16) and (4.17), establishes (4.15), since $4/\sqrt{\pi} \leq 3$.

■

Proof of Theorem 4.11. Combining Lemma 4.14 and Lemma 4.15, we obtain (4.14). ■

{Thm:NoReturn}

Theorem 4.16. *Let (Δ_i) be i.i.d. integer-valued variables with mean zero and variance σ^2, and let $X_t = \sum_{i=1}^{t} \Delta_i$.*

$$\mathbf{P}\{X_t \neq 0 \textit{ for } 1 \leq t \leq r\} \leq \frac{4\sigma}{\sqrt{r}}. \tag{4.18}$$ {Eq:NoReturn}

Remark 4.3. The constant in this estimate is not sharp, but we will give a very elementary proof, only using Chebyshev's inequality.

Proof. Let

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

be the set of times up to and including r when the walk visits v, and let

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

be those 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

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

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

$$1 \geq \mathbf{E}|A_r| = \mathbf{E}|L_r(0)|\alpha_r. \tag{4.19}$$ {Eq:LocTimeRF}

It remains to estimate $\mathbf{E}|L_n(0)|$ 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 to write

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

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 $\mathbf{E}|L_r(I)| \geq r/2$. The strong Markov property (at the first visit to v) shows that $\mathbf{E}|L_r(v)| \leq \mathbf{E}|L_r(0)|$ for any v, so that $r/2 \leq \mathbf{E}|L_r(D)| \leq 2\sigma\sqrt{r}\mathbf{E}|L_r(0)|$. Thus $\mathbf{E}|L_r(0)| \geq \sqrt{r}/(4\sigma)$. In conjunction with (4.19) this proves (4.18). ■

{Cor:LRWNoZero}

Corollary 4.17. *For the lazy simple random walk on $\mathbb{Z}$ started at height k,*

{Eq:NoHitZero}

$$\mathbf{P}_k\{\tau_0^+ > r\} \leq \frac{8k}{\sqrt{r}}. \tag{4.20}$$

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,

{Eq:LRWOneMove}

$$\mathbf{P}_0\{\tau_0^+ > r\} = \frac{1}{4}\mathbf{P}_1\{\tau_0^+ > r-1\} + \frac{1}{4}\mathbf{P}_{-1}\{\tau_0^+ > r-1\} = \frac{1}{2}\mathbf{P}_1\{\tau_0^+ > r-1\}. \tag{4.21}$$

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 (4.21) and Theorem 4.16,

{Eq:PenultAvoid}

$$\mathbf{P}_1\{\tau_k < \tau_0\}\mathbf{P}_k\{\tau_0^+ > r\} \leq \mathbf{P}_1\{\tau_0 > r-1\} = 2\mathbf{P}_0\{\tau_0^+ > r\} \leq \frac{8}{\sqrt{r}}. \tag{4.22}$$

(The variance σ^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 Equation 4.2, the chance 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 (4.22) this implies (4.20). ■

4.5.1. The Ballot Theorem.

4.6. Metropolis Chains and Glauber Dynamics

4.6.1. Metropolis chains. In Section 3.5, given an irreducible transition matrix P, we constructed a unique stationary distribution π satisfying $\pi = \pi P$. We now consider the inverse problem: given a probability distribution π on Ω, can we find a transition matrix P for which π is its stationary distribution?

Suppose that Ψ is a symmetric transition matrix. In this case, Ψ is reversible with respect to the uniform distribution on Ω. We now show how to modify transitions made according to Ψ to obtain a chain with stationary distribution π, where π is any probability distribution on Ω.

The new chain evolves as follows: when at state x, a candidate move is generated from the distribution $\Psi(x, \cdot)$. If the proposed new state is y, then the move is censored with probability $1 - a(x, y)$. That is, with probability $a(x, y)$, the state y is accepted as the new state, and with the remaining probability, the chain remains at x. Rejecting moves is wasteful, but may be necessary to achieve a specified stationary distribution. The transition matrix of this chain is P, where

$$P(x, y) = \begin{cases} \Psi(x, y)a(x, y) & y \neq x, \\ 1 - \sum_{z \in \Omega \setminus \{x\}} \Psi(x, z)a(x, z) & y = x. \end{cases}$$

P has stationary distribution π if

$$\pi(x)\Psi(x, y)a(x, y) = \pi(y)\Psi(y, x)a(y, x). \tag{4.23}$$

{Eq:RevMetrop}

Since we have assumed Ψ is symmetric, equation (4.23) holds if and only if

$$b(x, y) = b(y, x), \tag{4.24}$$

{Eq:RevMetrop2}

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

$$\begin{aligned} b(x, y) &\leq \pi(x), \\ b(x, y) = b(y, x) &\leq \pi(y). \end{aligned} \tag{4.25}$$

{Eq:bCons}

Since rejecting the moves of the original chain Ψ is wasteful, a solution b to (4.24) and (4.25) should be chosen which is as large as possible. Clearly, all solutions are bounded above by $b(x, y) = \pi(x) \wedge \pi(y)$. For this choice, the acceptance probability $a(x, y)$ equals $(\pi(y)/\pi(x)) \wedge 1$.

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

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

{Rmk:MetRat}

Remark 4.4. A very important feature of the Metropolis chain is that it only depends on the ratios $\pi(x)/\pi(y)$. Frequently $\pi(x)$ is only be explicitly known up to a normalizing constant. The optimization chains described below are examples of this type. The normalizing constant is not needed to run the Metropolis chain.

Example 4.18 (Optimization). Let f be a real-valued function defined on the vertex set Ω of a graph. In many applications it is desired to find a vertex x where f is largest; if the domain Ω is very large, then an exhaustive search many too expensive.

A *hill climb* is an algorithm which attempts to locate the maximum values of f as follows: when at x, if a neighbor y of x has $f(y) > f(x)$, move to y. The reader will quickly see that if f has a local maximum, then the climber may become trapped before discovering a global maximum.

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

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

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

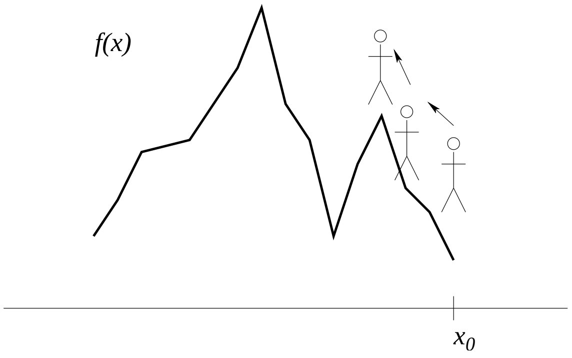


FIGURE 4.4. A hill climb may become trapped at a local maximum. Fig:HillClimb

where $Z(\lambda) := \sum_{x\in\Omega} \lambda^{f(x)}$ is a normalizing constant making μ a probability measure. Note that $\pi(x)$ is increasing in $f(x)$, so that π favors x with large values of $f(x)$.

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

Suppose that

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

Then

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

That is, as $\lambda \to \infty$, the stationary distribution converges to the uniform distribution over the global maximum of f.

As mentioned in Remark 4.4, running the Metropolis chain does not require computation of $Z(\lambda)$, which may be prohibitively expensive to compute.

The Metropolis chain can be defined when the underlying chain is not symmetric.

{Example:MetroplisSRW0}

EXAMPLE 4.19. Suppose you know neither the vertex set V or the edges et of a graph, but are however able to perform a random walk on the graph. You desire a uniform sample from V. Many computer and social networks are of this form. If the graph is not regular, then the stationary distribution is not uniform, so the distribution of the walk will not converge to uniform.

For a general (irreducible) transition matrix Ψ, and an arbitrary probability distribution π on Ω, the Metropolized chain is executed as follows: When at state x, generate a state y from $\Psi(x,\cdot)$. Move to y with probability

{Eq:MetropAcceptProb}

$$\frac{\pi(y)\Psi(y,x)}{\pi(x)\Psi(x,y)} \wedge 1, \tag{4.26}$$

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

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

The reader should check that P is reversible with respect to the probability distribution π.

{Example:MetropolisSRW}

Example 4.20. Consider the set-up in Example 4.19. The Metropolis algorithm can modify the simple random walk to ensure a uniform stationary distribution. The acceptance probability in (4.26) reduces in this case to

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

This biases the walk against moving to higher degree vertices, giving a uniform stationary distribution. Note that the size of the graph is not needed to perform this modification, an important consideration in applications.

4.6.2. Glauber Dynamics. A *proper q-coloring* of the vertices V of a graph assigns to each vertex one among q possible colors so that no two neighboring vertices share a common color. We will represent the colors monochromatically by the integers $\{1, 2, \ldots, q\}$. A proper q-coloring is an element x of $\{1, 2, \ldots, q\}^V$, the functions from V to $\{1, 2, \ldots, q\}$, so that $x(v) \neq x(w)$ for all edges $\{v, w\}$.

A *hardcore configuration* is a placement of particles on the vertices V of a graph so that no two particles are adjacent. A hardcore configuration x is an element of $\{0, 1\}^V$, the functions from V to $\{0, 1\}$, satisfying $x(v)x(w) = 0$ for all edges $\{v, w\}$.

In general, suppose that Ω is a subset of S^V, where V is the vertex set of a graph and S is a finite set, and let μ be a probability distribution on Ω. Both the set of proper q-colorings and the set of hardcore configurations are of this form. In this section, we describe *Glauber dynamics* for μ, which is a reversible Markov chain with stationary distribution μ.

In words, the Glauber chain moves from state x as follows: a vertex w is chosen uniformly at random from V, and a new state is chosen according to the measure μ conditioned to equal x at all vertices different from w.

{Ex:GlCol}

Example 4.21 (Glauber dynamics for uniform proper q-colorings). Suppose that μ is the uniform distribution on proper q-colorings. To understand how the Glauber chain transitions from x, we must determine the distribution of μ conditioned on the set

$$A_{x,w} := \{z \in \Omega \,:\, z(v) = x(v) \text{ for } v \neq w\}.$$

Call a color *feasible* at w in configuration x if it is *not* among the set $\{x(z) \,:\, z \sim w\}$. A configuration x can be changed at vertex w only to a feasible color. The set $A_{x,w}$ consists of all configurations agreeing with x away from w and having a feasible color at w. Since μ is uniform on Ω,

$$\mu(y \mid A_{x,w}) = \frac{1}{|A_{x,w}|}.$$

Thus, the Glauber chain moves from x by selecting a vertex w at random and updating the color at w to a uniform sample from the feasible colors at w.

Example 4.22 (Glauber dynamics for uniform Hardcore configuration). Let μ be the uniform distribution on the Hardcore configurations. The reader should check that the Glauber dynamics for μ updates x at vertex w by leaving w unoccupied if a neighbor of w is occupied, and by placing a particle at w with probability $1/2$ if no neighbor is occupied.

Consider the Markov chain on $\{0,1\}^V$ which moves by picking a vertex w at random and then updating w by placing a particle there with probability $1/2$. Note that this chain does not live on the space of hardcore configurations, as nothing restricts moves placing two particles on adjacent vertices. The Metropolis chain for the uniform distribution on hardcore configurations accepts a move $x \mapsto y$ with probability 0 if y is not a hardcore configuration, and with probability 1 if y is a hardcore configuration. Thus, the Metropolis chain and the Glauber dynamics agree in this example.

4.7. The Pivot Chain for Self-Avoiding Random Walk*

{ExampSAWpSAW2}

Example 4.23 (Pivot chain for self-avoiding paths). The space Ξ_n of self-avoiding lattice paths of length n was described in Example 2.1. These are paths in $\mathbb{Z}^2$ of length n which never intersect themselves.

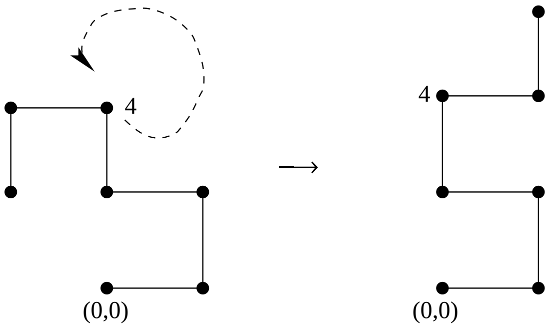
We describe now a Markov chain on Ξ_n and show that it is irreducible. If the current state of the chain is the path $(0, v_1, \ldots, v_n) \in \Xi_n$, the next state is chosen by the following:

(1) Pick a value k from $\{0, 1, \ldots, n\}$ uniformly at random.
(2) Pick uniformly at random from the following transformations of $\mathbb{Z}^2$: Rotations clockwise by $\pi/2$, π, $3\pi/2$, reflection across the x-axis, and reflection across the y-axis.
(3) Take the path from vertex k on, $(v_k, v_{k+1}, \ldots, v_n)$, and apply the transformation chosen in the previous step to this subpath only, taking v_k as the origin.
(4) If the resulting path is self-avoiding, this is the new state. If not, repeat.

An example move is shown in Figure 4.5.

We now show that this chain is irreducible by proving that any self-avoiding path can be unwound to a straight line by a sequence of possible transitions. Since the four straight paths starting at $(0,0)$ are rotations of each other, and since any transition can also be undone by a dual transition, any self-avoiding path can be transformed into another. The proof below follows Madras and Slade (1993, Theorem 9.4.4).

For a path $\xi \in \Xi_n$, put around ξ as small a rectangle as possible, and define $D = D(\xi)$ to be the sum of the length and the width of this rectangle. The left-hand diagram in Figure 4.6 shows an example of this bounding rectangle. Define also $A = A(\xi)$ to be the number of interior vertices v of ξ where the two edges incident at v form an angle of π, that is, which look like either —•— or $\overset{|}{\underset{|}{\bullet}}$. We first observe



current path

path after rotating by π from vertex 4

FIGURE 4.5. Example of a single move of pivot chain for self-avoiding walk. fig:SAWTransition

that $D(\xi) \leq n$ and $A(\xi) \leq n-1$ for any $\xi \in \Xi_n$, and $D(\xi)+A(\xi) = 2n-1$ if and only if ξ is a straight path. We show now that if ξ is any path different from the straight path, we can make a legal move —that is, a move having positive probability—to another path ξ' which has $D(\xi') + A(\xi') > D(\xi) + A(\xi)$.

There are two cases which we will consider separately.

Case 1. Suppose that at least one side of the bounding box does not contain either endpoint, 0 or v_n, of $\xi = (0, v_1, \ldots, v_n)$. This is the situation for the path on the left-hand side in Figure 4.6. Let $k \geq 1$ be the smallest index so that v_k lies on this side. Obtain ξ' by taking ξ and reflecting its tail $(v_k, v_{k+1}, \ldots, v_n)$ across this box side. Figure 4.6 shows an example of this transformation. The new path ξ' satisfies $D(\xi') > D(\xi)$ and $A(\xi') = A(\xi)$ (the reader should convince herself this is indeed true!)

Case 2. Suppose every side of the bounding box contains an endpoint of ξ. This implies that the endpoints are in opposing corners of the box. Let k be the largest index so that the edges incident to v_k form a right angle. The path ξ from v_k to v_n forms a straight line segment, and must lie along the edge of the bounding box. Obtain ξ' from ξ by rotating this straight portion of ξ so that it lies outside the original bounding box. See Figure 4.7.

This operation reduces one dimension of the bounding box by at most the length of the rotated segment, but increases the other dimension by this length. This shows that $D(\xi') \geq D(\xi)$. Also, we have strictly increased the number of straight angles, so $D(\xi') + A(\xi') > D(\xi) + A(\xi)$.

In either case, $D + A$ is strictly increased by the transformation, so continuing this procedure eventually leads to a straight line segment. This establishes that the pivot Markov chain is irreducible.

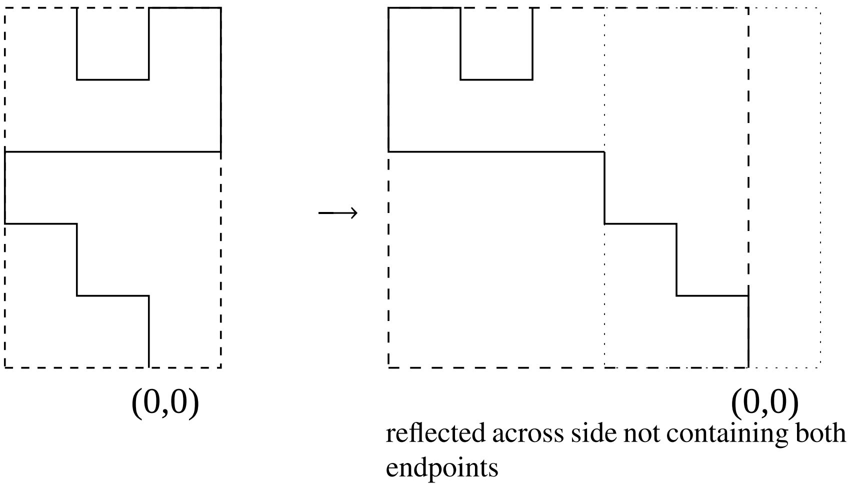


FIGURE 4.6. A SAW without both endpoints in corners of bounding box. Fig:SawCase1

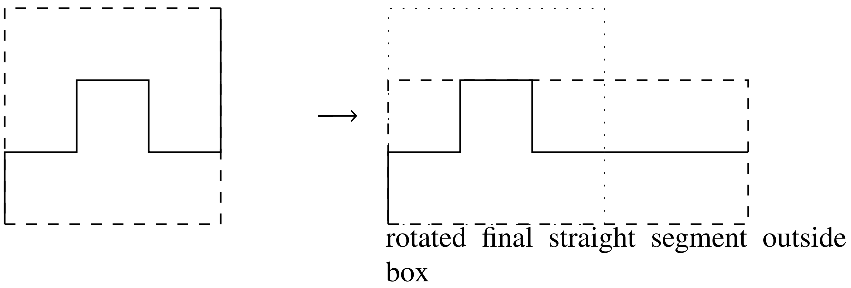


FIGURE 4.7. A SAW with endpoints in opposing corners. Fig:SawCase2

4.8. Problems

{Exer:GamblersRuin}

EXERCISE 4.1. Show that the system of equations

$$f_k = \frac{1}{2}(1 + f_{k+1}) + \frac{1}{2}(1 + f_{k-1}) \tag{4.27}$$

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 Δ_k's form an arithmetic progression) and that $\sum_k \Delta_k = 0$.

{Exer:LazyGambler}

EXERCISE 4.2. Consider a lazy 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. Find the expected value of the time required for her to be absorbed at (either) endpoint in terms of n, k, and p.

{Exer:HitOtherEnd}

Exercise 4.3. Consider a random walk on the path $\{0, 1, \dots, n\}$ in which the walker moves left or right with equal probability; if he tries to move above n, he stays put for that round, and if he hits 0, he stays there forever. Compute the expected time of the walker's absorption at state 0, given that he starts at state n.

{Exercise:EhrenStat}

Exercise 4.4. Let P be the transition matrix for the Ehrenfest chain described in Equation 4.8. Show that the Binomial distribution with parameters d and $1/2$ is the stationary distribution for this chain.

{Exer:HarmonicSum}

Exercise 4.5.

(a) By comparing the integral of $1/x$ with its Riemann sums, show that

$$\log n \leq \sum_{k=1}^{n} k^{-1} \leq \log n + 1. \tag{4.28}$$

(b) In the set-up of Proposition 4.2, prove that

$$\mathbf{P}\{\tau > cn(\log n + 1)\} \leq \frac{1}{c}.$$

{Exercise:EhrenStat}

Exercise 4.6. Let P be the transition matrix for the Ehrenfest chain described in Equation 4.8. Show that the Binomial distribution with parameters d and $1/2$ is the stationary distribution for this chain.

4.9. Notes

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 featured in problem B1 of the 2002 Putnam mathematical competition.

It is an open problem to analyze the convergence behavior of the pivot chain on self-avoiding walks. The algorithm of Randall and Sinclair (2000) uses a different underlying Markov chain to approximately sample from the uniform distribution on these walks.

Rigorous results for simulated annealing were obtained in Hajek (1988).

CHAPTER 5

Introduction to Markov Chain Mixing

{Chapters:Mixing}

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

First we define the *total variation distance* and give several characterizations of it, all of which will be useful in our future work. Next we prove the Convergence Theorem (Theorem 5.6), which says that for an irreducible and aperiodic chain the distribution after many steps approaches the chain's stationary distribution, in the sense that the total variation distance between them approaches 0. In the rest of the chapter we examine the effects of the initial distribution on distance from stationarity, define the *mixing time* of a chain, and prove a version of the Ergodic Theorem (Theorem 5.11) for Markov chains.

5.1. Total Variation Distance

{Sec:TVDist}

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

$$\|\mu - \nu\|_{\mathrm{TV}} = \max_{A \subset \Omega} |\mu(A) - \nu(A)| \,. \tag{5.1}$$

{eq:TVdef}

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

{Xmpl:FrogTV}

Example 5.1. Recall the coin-tossing frog of Example 3.1, who has probability p of jumping from east to west, and probability q of jumping from west to east. His transition matrix is $\begin{pmatrix} 1-p & p \\ q & 1-q \end{pmatrix}$ and his stationary distribution is $\pi = \left(\frac{q}{p+q}, \frac{p}{p+q}\right)$. Assume the frog starts at the east pad (that is, $\mu_0(t) = (1,0)$) and define

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

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

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

We pointed out in Example 3.1 that $\Delta_t = (1-p-q)^t \Delta_0$. Hence for this two-state chain, the total variation distance decreases exponentially fast at t increases. (Note that $(1-p-q)$ is an eigenvalue of P; we will discuss connections between eigenvalues and mixing in Chapter 12.)

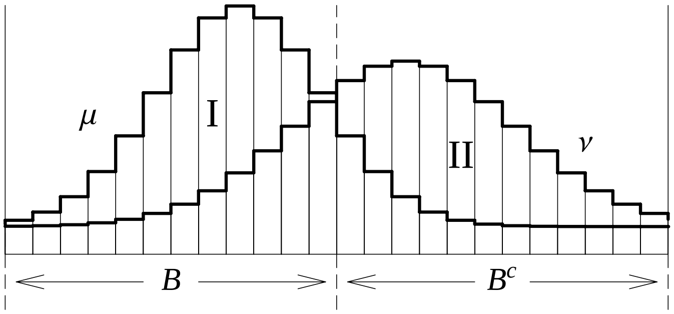


fig:TV1

FIGURE 5.1. Recall that $B = \{x : \mu(x) > \nu(x)\}$. Region I has area $\mu(B) - \nu(B)$. Region II has area $\nu(B^c) - \mu(B^c)$. Since the total area under each of μ and ν is 1, regions I and II must have the same area—and that area is $\|\mu - \nu\|_{TV}$.

It is not immediately clear from (5.1) how to compute the total variation distance between two given distributions. We now give three extremely useful alternative characterizations. Proposition 5.2 reduces total variation distance to a simple sum over the state space. Proposition 5.3 describes total variation distance in terms of integrating a single function with respect to both underlying measures. Proposition 5.5 uses *coupling* to give another probabilistic interpretation: $\|\mu - \nu\|_{TV}$ measures how close to identical we can force two random variables realizing μ and ν to be.

{Prop:TotalVariation}

PROPOSITION 5.2. *Let μ and ν be two probability distributions on Ω. Then*

{Eq:TVisL1}

$$\|\mu - \nu\|_{\mathrm{TV}} = \frac{1}{2} \sum_{x \in \Omega} |\mu(x) - \nu(x)| . \tag{5.2}$$

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

$$\begin{aligned} \mu(A) - \nu(A) &\leq \mu(A \cap B) - \nu(A \cap B) \qquad (5.3)\\ &\leq \mu(B) - \nu(B). \qquad (5.4) \end{aligned}$$

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

By exactly parallel reasoning,

$$\nu(A) - \mu(A) \leq \nu(B^c) - \mu(B^c). \tag{5.5}$$

Fortunately, these two upper bounds are actually the same (as can be seen by subtracting them; see Figure 5.1). Furthermore, when we take $A = B$ (or B^c), then $|\mu(A) - \nu(A)|$ is equal to the upper bound. Thus

$$\|\mu - \nu\|_{TV} = \frac{1}{2}\left[\mu(B) - \nu(B) + \nu(B^c) - \mu(B^c)\right] = \frac{1}{2} \sum_{x \in \Omega} |\mu(x) - \nu(x)|. \tag{5.6}$$

■ {Rmk:TVSet}

Remark 5.1. The proof of Proposition 5.2 also shows that

$$\|\mu - \nu\|_{TV} = \sum_{\substack{x \in \Omega \\ \mu(x) \geq \nu(x)}} [\mu(x) - \nu(x)], \tag{5.7}$$ {Eq:TVHalfSum}

which is a useful identity.

{Prop:TVFunction}

Proposition 5.3. *Let μ and ν be two probability distributions on Ω. Then the total variation distance between them satisfies*

$$\|\mu - \nu\|_{TV} \\ = \frac{1}{2} \sup \left\{ \sum_{x \in \Omega} f(x)\mu(x) - \sum_{x \in \Omega} f(x)\nu(x) \ : \ f \textit{ satisfying } \max_{x \in \Omega} |f(x)| \leq 1 \right\}. \tag{5.8}$$ {Eq:TVLInf}

Proof. We have

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

This shows that the right-hand side of (5.8) is not more than $\|\mu - \nu\|_{TV}$. Define

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

Then

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

Using (5.7) shows that the right-hand side above equals $\|\mu - \nu\|_{TV}$. This shows that the right-hand side of (5.8) is at least $\|\mu - \nu\|_{TV}$. ■

5.2. Coupling and Total Variation Distance

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

Coupling is a general and powerful technique; it can be applied in many different ways. Indeed, Chapters 6 and 14 use couplings of entire chain trajectories to bound rates of convergence to stationarity. Here, we offer a gentle introduction by showing the close connection between couplings of two random variables and the total variation distance between those variables.

{Xmpl:CoupleFairCoin}

Example 5.4. Let μ and ν both be the "fair coin" measure giving weight 1/2 to the elements of $\{0, 1\}$.

{It:IndCoin}

(i) One way to couple μ and ν is to define (X, Y) to be a pair of independent coins, so that $\mathbf{P}\{X = x,\ Y = y\} = 1/4$ for all $x, y \in \{0, 1\}$.

{It:SameCoin}

(ii) Another way to couple μ and ν is to let X be a fair coin toss, and define $Y = X$. In this case, $\mathbf{P}\{X = Y = 0\} = 1/2$, $\mathbf{P}\{X = Y = 1\} = 1/2$, and $\mathbf{P}\{X \neq Y\} = 0$.

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

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

and

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

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

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

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

Returning to Example 5.4, the coupling in part (i) could equivalently be specified by the probability distribution q_1 on $\{0, 1\}^2$ given by

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

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

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

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

{Prop:TVCouple}

Proposition 5.5. *Let μ and ν be two probability distributions on Ω. Then*

{Eq:TVCoupling}

$$\|\mu - \nu\|_{TV} = \inf\left\{\mathbf{P}\{X \neq Y\} \,:\, (X,Y) \textit{ is a coupling of } \mu \textit{ and } \nu\right\}. \tag{5.9}$$

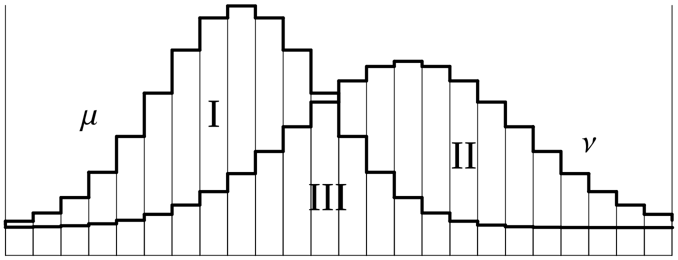


fig:TVcouple

FIGURE 5.2. Since each of regions I and II has area $\|\mu - \nu\|_{TV}$, and μ and ν are probability measures, region III has area $1 - \|\mu - \nu\|_{TV}$.

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

$$\begin{aligned} \mu(A) - \nu(A) &= \mathbf{P}\{X \in A\} - \mathbf{P}\{Y \in A\} && (5.10) \\ &\leq \mathbf{P}\{X \in A, Y \notin A\} && (5.11) \\ &\leq \mathbf{P}\{X \neq Y\}. && (5.12) \end{aligned}$$

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

$$\|\mu - \nu\|_{TV} \leq \inf\left\{\mathbf{P}\{X \neq Y\} \,:\, (X, Y) \text{ is a coupling of } \mu \text{ and } \nu\right\}. \tag{5.13}$$ {Eq:TVLessC}

If we can construct a coupling for which $\mathbf{P}\{X \neq Y\}$ is actually equal to $\|\mu - \nu\|_{TV}$, we'll be done. We will do so by forcing X and Y to be equal as often as they possibly can be. Consider Figure 5.2. Region III, bounded by $\mu(x) \wedge \nu(x) = \min\{\mu(x), \nu(x)\}$, can be seen as the overlap between the two distributions. We construct our coupling so that, whenever we "land" in region III, $X = Y$. Otherwise, we accept that X must be in region I and Y must be in region II; since those regions have disjoint support, X and Y cannot be equal.

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

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

Write

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

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

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

By Equation 5.7 and the immediately preceding equation,

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

We can thus define the probability distribution $\gamma_{\mathrm{III}}(x) = p^{-1}[\mu(x) \wedge \nu(x)]$.

Flip a coin with probability of heads equal to p.

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

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

and set $X = Y = Z$.

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

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

and independently choose Y according to the probability distribution

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

γ_{I} and γ_{II} are probability distributions by (5.7).

Clearly,

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

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

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

■

We call a coupling *optimal* if it attains the infimum in (5.9). The above proof shows that in fact an optimal coupling always exists.

5.3. Convergence Theorem

We are now ready to prove that irreducible, aperiodic Markov chains converge to their stationary distributions—a key step, as much of the rest of the book will be devoted to estimating the rate at which this convergence occurs. The assumption of aperiodicity is indeed necessary—recall the even n-cycle of Example 3.2.

As is often true of such fundamental facts, there are many proofs of the Convergence Theorem. The one given here decomposes the chain into a mixture of repeated independent sampling from its own stationary distribution and another Markov chain. The argument is finished via a little matrix algebra; we've put the details in an exercise. See Exercise 6.1 for another proof using two coupled copies of the chain.

{Thm:ConvergenceThm}

THEOREM 5.6. *Suppose that P is irreducible and aperiodic, with stationary distribution π. Then there exists $1 > \alpha > 0$ such that*

{Eq:ConvThm}

$$\max_{x\in\Omega} \left\|P^t(x,\cdot) - \pi\right\|_{\mathrm{TV}} \leq \alpha^t. \tag{5.14}$$

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

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

for all $x, y \in \Omega$. Once we fix such a δ, the equation

$$P^r = \delta\Pi + (1 - \delta)Q \tag{5.15}$$

{Eq:PmDecomp}

defines a stochastic matrix Q.

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

Next, we use induction to demonstrate that

$$P^{rk} = \left[1 - (1 - \delta)^k\right]\Pi + (1 - \delta)^k Q^k. \tag{5.16}$$

{Eq:PmGeo}

for $k \geq 1$. If $k = 1$, this holds by (5.15). Assuming that (5.16) holds for $k = n$,

$$P^{r(n+1)} = P^{rn}P^r = \{[1 - (1 - \delta)^n]\,\Pi + (1 - \delta)^n Q^n\}\, P^r. \tag{5.17}$$

Distributing and expanding P^r in the second term gives

$$P^{r(n+1)} = [1 - (1 - \delta)^n]\,\Pi P^r + \delta(1 - \delta)^n Q^n\Pi + (1 - \delta)^{n+1} Q^n Q. \tag{5.18}$$

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

$$P^{r(n+1)} = \left[1 - (1 - \delta)^{n+1}\right]\Pi + (1 - \delta)^{n+1} Q^{n+1}. \tag{5.19}$$

This establishes (5.16) for $k = n + 1$ (assuming it holds for $k = n$), and hence it holds for all k.

Multiplying by P^j and rearranging terms now yields

$$P^{rk+j} - \Pi = (1 - \delta)^k \left[Q^k P^j - \Pi\right]. \tag{5.20}$$

{Eq:MatrixDiff}

To complete the proof, examine the x_0th row of (5.20). Take the L^1 norm of both sides and divide by 2. On the right, the second factor is at most the largest possible total variation distance between distributions, which is 1. Hence for any x_0 we have

$$\left\|P^{rk+j}(x_0, \cdot) - \pi\right\|_{TV} \leq (1 - \delta)^k. \tag{5.21}$$

■

Remark. Because of Theorem 5.6, the distribution π is also called the *equilibrium distribution*.

5.4. Standardizing distance from stationarity

Bounding the maximal distance between $P^t(x_0, \cdot)$ and π appearing in the Convergence Theorem (Theorem 5.6) is among our primary objectives. It would simplify analysis to eliminate the dependence on the initial state, so that "distance from stationarity" depends on the transition matrix and the number of steps. In view of this, we define

$$d(t) := \max_{x \in \Omega} \left\|P^t(x, \cdot) - \pi\right\|_{TV}. \tag{5.22}$$

{Eq:dDefn}

We will see in Chapter 6 that it is often possible to bound the maximum distance between the distribution of the chain started from x and the distribution of the chain started at y, over all pairs of states (x, y). Thus it is convenient to define

{Eq:dbarDefn}

$$\bar{d}(t) := \max_{x,y\in\Omega} \left\| P^t(x, \cdot) - P^t(y, \cdot) \right\|_{\mathrm{TV}} . \tag{5.23}$$

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

{Lem:StationaryVsState}

Lemma 5.7.

{Eq:StationaryVsState}

$$d(t) \le \bar{d}(t). \tag{5.24}$$

Proof. As π is stationary, $\pi(A) = \sum_y \pi(y)P^t(y, A)$ for any set A. (This is the definition of stationarity if A is a singleton $\{x\}$. To get this for arbitrary A, just sum over the elements in A.) Using this shows that

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

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

{Eq:LemStepTV}

$$\max_A \sum_{y\in\Omega} \pi(y)|P^t(x, A) - P^t(y, A)| \le \sum_{y\in\Omega} \pi(y) \max_A |P^t(x, A) - P^t(y, A)|. \tag{5.25}$$

Finally, a weighted average of a set of numbers is never larger than the maximum element, so the right-hand side in (5.25) is bounded by $\max_{y\in\Omega} \left\| P^t(x, \cdot) - P^t(y, \cdot) \right\|_{\mathrm{TV}}$. ■

Exercise 5.1 asks the reader to prove the following equalities:

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

{Lem:TVSubMult}

Lemma 5.8. *The function $\bar{d}$ is submultiplicative: $\bar{d}(s + t) \le \bar{d}(s)\bar{d}(t)$.*

Proof. Fix $x, y \in \Omega$, and let (X_s, Y_s) be the optimal coupling of $P^s(x, \cdot)$ and $P^s(y, \cdot)$ whose existence is guaranteed by Proposition 5.5. Hence

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

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

$$P^{s+t}(x, w) = \sum_z P^s(x, z)P^t(z, w) = \sum_z \mathbf{P}\{X_s = z\}P^t(z, w) = \mathbf{E}\left(P^t(X_s, w)\right). \tag{5.26}$$

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

{Eq:SandTCoupling}

$$\begin{aligned} P^{s+t}(x, w) - P^{s+t}(y, w) &= \mathbf{E}\left(P^t(X_s, w)\right) - \mathbf{E}\left(P^t(Y_s, w)\right) \\ &= \mathbf{E}\left(P^t(X_s, w) - P^t(Y_s, w)\right). \end{aligned} \tag{5.27}$$

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

Summing (5.27) over $w \in \Omega$ and applying Proposition 5.2 shows that

$$\left\|P^{s+t}(x,\cdot) - P^{s+t}(y,\cdot)\right\|_{\mathrm{TV}} = \frac{1}{2}\sum_{w}\left|\mathbf{E}\left(P^t(X_s,w) - P^t(Y_s,w)\right)\right|. \tag{5.28}$$

Since $|\mathbf{E}(Z)| \leq \mathbf{E}(|Z|)$ for any random variable Z and expectation is linear, the right-hand side above is less than or equal to

$$\mathbf{E}\left(\frac{1}{2}\sum_{w}\left|P^t(X_s,w) - P^t(Y_s,w)\right|\right). \tag{5.29}$$

Applying Proposition 5.2 again, we see that the quantity inside the expectation is exactly the distance $\left\|P^t(X_s,\cdot) - P^t(Y_s,\cdot)\right\|_{TV}$, which is zero whenever $X_s = Y_s$. Moreover, this distance is always bounded by $\bar{d}(t)$. This shows that

$$\left\|P^{s+t}(x,\cdot) - P^{s+t}(y,\cdot)\right\|_{\mathrm{TV}} \leq \bar{d}(t)\mathbf{E}\left(\mathbf{1}_{\{X_s \neq Y_s\}}\right) = \bar{d}(t)\mathbf{P}\{X_s \neq Y_s\}. \tag{5.30}$$

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

Exercise 5.3 implies that $\bar{d}(t)$ is non-increasing in t. From this and Lemma 5.8 it follows that when c is any non-negative real number and t is any non-negative integer, we have

$$\bar{d}(ct) \leq \bar{d}(\lfloor c \rfloor t) \leq \bar{d}(t)^{\lfloor c \rfloor}. \tag{5.31}$$ {Eq:TimeMult}

5.5. Mixing Time

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

$$t_{\mathrm{mix}}(\varepsilon) := \min\{t \ : \ d(t) \leq \varepsilon\}, \tag{5.32}$$ {Eq:MixingTimeDefnEp}

$$t_{\mathrm{mix}} := t_{\mathrm{mix}}(1/4). \tag{5.33}$$ {Eq:MixingTimeDefn}

Together Lemma 5.7 and Exercise 5.5 show that $d(t) \leq \bar{d}(t) \leq 2d(t)$. This, with Equation 5.31, shows that if c is a non-negative real number,

$$d(\,ct_{\mathrm{mix}}(\varepsilon)\,) \leq \bar{d}(\,ct_{\mathrm{mix}}(\varepsilon)\,) \leq \bar{d}(\,t_{\mathrm{mix}}(\varepsilon)\,)^{\lfloor c \rfloor} \leq (2\varepsilon)^{\lfloor c \rfloor}. \tag{5.34}$$ {Eq:dTimeMult}

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

$$d(\,ct_{\mathrm{mix}}\,) \leq (1/2)^{\lfloor c \rfloor} \tag{5.35}$$ {Eq:MTMult}

$$t_{\mathrm{mix}}(\varepsilon) \leq \left\lceil \log_2 \varepsilon^{-1} \right\rceil t_{\mathrm{mix}}. \tag{5.36}$$ {Eq:TMixMult}

Thus, although the choice of $1/4$ is arbitrary in the definition of t_{mix} (Equation 5.33), a value of ε less than $1/2$ is needed to make the inequality $d(\,ct_{\mathrm{mix}}(\varepsilon)\,) \leq (2\varepsilon)^{\lfloor c \rfloor}$ in (5.34) non-trivial and to achieve an inequality like (5.36).

5.6. Reversing Symmetric Chains

For a distribution R on $\mathcal{S}_n$, the *inverse distribution* $\overline{R}$ is defined by $\overline{R}(\rho) = R(\rho^{-1})$.

{Lem:InvTVSame}

Lemma 5.9. *Let P be the transition matrix of the random walk on $\mathcal{S}_n$ generated by a distribution R, and let $\overline{P}$ be that of the walk generated by $\overline{R}$. Let U be the uniform distribution on $\mathcal{S}$. Then*

$$\left\|P^t(\mathrm{id},\cdot) - U\right\|_{\mathrm{TV}} = \left\|\overline{P}^t(\mathrm{id},\cdot) - U\right\|_{\mathrm{TV}}$$

Proof. Let $X_0 = \mathrm{id}, X_1, \ldots$ be a Markov chain with transition matrix P. We can write $X_k = \pi_1\pi_2\ldots\pi_k$, where the random permutations $\pi_1, \pi_2, \cdots \in \mathcal{S}_n$ are independent choices from the distribution R. Similarly, let (Y_t) be a chain with transition matrix $\overline{P}$, with increments $\rho_1, \rho_2, \cdots \in \mathcal{S}_n$ chosen independently from $\overline{R}$.

For any fixed elements $\sigma_1, \ldots, \sigma_t \in \mathcal{S}_n$,

$$\mathbf{P}(\pi_1 = \sigma_1, \ldots, \pi_t = \sigma_t) = \mathbf{P}(\rho_1 = \sigma_t^{-1}, \ldots, \rho_t = \sigma_1^{-1}),$$

by the definition of $\overline{P}$. Summing over all strings such that $\sigma_1\sigma_2\ldots\sigma_t = \sigma$ yields

$$P^t(\mathrm{id},\sigma) = \overline{P}^t(\mathrm{id},\sigma^{-1}).$$

Hence

$$\sum_{\sigma\in\mathcal{S}_n}\left|P^t(\mathrm{id},\sigma) - \frac{1}{n!}\right| = \sum_{\sigma\in\mathcal{S}_n}\left|\overline{P}^t(\mathrm{id},\sigma^{-1}) - \frac{1}{n!}\right| = \sum_{\sigma\in\mathcal{S}_n}\left|\overline{P}^t(\mathrm{id},\sigma) - \frac{1}{n!}\right|$$

which is the desired result. ■

The result of Lemma 5.9 generalizes to slightly less symmetric Markov chains.

{Lem:TimeReversal}

Lemma 5.10. *Let P be a transitive transition matrix and let $\widehat{P}$ be the time-reversed matrix defined in (3.30). Then*

{Eq:TimeReversedTV}

$$\left\|\widehat{P}^t(x,\cdot) - \pi\right\|_{\mathrm{TV}} = \left\|P^t(x,\cdot) - \pi\right\|_{\mathrm{TV}}. \tag{5.37}$$

Proof. Since our chain is transitive, it has a uniform stationary distribution (see Exercise 7.5). For $x, y \in \Omega$, let $\phi_{(x,y)}$ be a permutation carrying x to y and preserving the structure of the chain. For any $x, y \in \Omega$ and any t,

$$\sum_{z\in\Omega}\left|P^t(x,z) - |\Omega|^{-1}\right| = \sum_{z\in\Omega}\left|P^t(\phi_{(x,y)}(x), \phi_{(x,y)}(z)) - |\Omega|^{-1}\right| \tag{5.38}$$

$$= \sum_{z\in\Omega}\left|P^t(y,z) - |\Omega|^{-1}\right|. \tag{5.39}$$

Averaging both sides over y yields

{Eq:TransDoubleSumTV}

$$\sum_{z\in\Omega}\left|P^t(x,z) - |\Omega|^{-1}\right| = \frac{1}{|\Omega|}\sum_{y\in\Omega}\sum_{z\in\Omega}\left|P^t(y,z) - |\Omega|^{-1}\right|. \tag{5.40}$$

Because π is uniform, we have $P(y,z) = \widehat{P}(z,y)$, and thus $P^t(y,z) = \widehat{P^t}(z,y)$. It follows that the right-hand side above is equal to

$$\frac{1}{|\Omega|}\sum_{y\in\Omega}\sum_{z\in\Omega}\left|\widehat{P^t}(z,y) - |\Omega|^{-1}\right| = \frac{1}{|\Omega|}\sum_{z\in\Omega}\sum_{y\in\Omega}\left|\widehat{P^t}(z,y) - |\Omega|^{-1}\right| \tag{5.41}$$

By Exercise 7.7, $\widehat{P}$ is also transitive, so (5.40) holds with $\widehat{P}$ replacing P (and z and y interchanging roles). We conclude that

$$\sum_{z\in\Omega}\left|P^t(x,z) - |\Omega|^{-1}\right| = \sum_{y\in\Omega}\left|\widehat{P^t}(x,y) - |\Omega|^{-1}\right|. \tag{5.42}$$

Dividing by 2 and applying Proposition 5.2 completes the proof. ■

5.7. Ergodic Theorem*

The idea of the ergodic theorem for Markov chain is that "time averages equal space averages".

If f is a real-valued function defined on Ω, and μ is any probability distribution on Ω, then we define

$$E_\mu(f) = \sum_{x\in\Omega} f(x)\mu(x).$$

{Thm:ErgodicThm}

THEOREM 5.11. *Let f be a real-valued function defined on Ω. If (X_t) is an irreducible Markov chain, then for any starting distribution μ,*

$$\mathbf{P}_\mu\left\{\lim_{t\to\infty}\frac{1}{t}\sum_{s=0}^{t-1} f(X_s) = E_\pi(f)\right\} = 1. \tag{5.43}$$

{Eq:ErgodicThm}

PROOF. Suppose that the chain starts at x, define $\tau^+_{x,0} := 0$ and

$$\tau^+_{x,k} = \min\{t > \tau^+_{x,(k-1)} \,:\, X_t = 0\}.$$

Since the chain "starts afresh" every time it visits x, the blocks $(X_{\tau^+_{x,k}}, X_{\tau^+_{x,k}+1}, \ldots, X_{\tau^+_{x,(k+1)}-1})$ are independent of one another. Thus if

$$Y_k := \sum_{s=\tau^+_{x,(k-1)}}^{\tau^+_{x,k}-1} f(X_s),$$

then the sequence (Y_k) is i.i.d. If $S_t = \sum_{s=0}^{t-1} f(X_s)$, then $S_{\tau^+_{x,n}} = \sum_{k=1}^n Y_k$, by the Strong Law of Large Numbers (Theorem B.4),

$$\mathbf{P}_x\left\{\lim_{n\to\infty}\frac{S_{\tau^+_{x,n}}}{n} = \mathbf{E}_x(Y_1)\right\} = 1.$$

Again by the Strong Law of Large Numbers, since $\tau^+_{x,n} = \sum_{k=1}^n(\tau^+_{x,k} - \tau^+_{x,(k-1)})$, writing simply τ^+_x for $\tau^+_{x,1}$,

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

Thus,

$$\mathbf{P}_x\left\{\lim_{n\to\infty}\frac{S_{\tau^+_{x,n}}}{\tau^+_{x,n}} = \frac{\mathbf{E}_x(Y_1)}{\mathbf{E}_x(\tau^+_x)}\right\} = 1. \tag{5.44}$$ {Eq:Erg1}

Note that

$$\mathbf{E}_x(Y_1) = \mathbf{E}_x\left(\sum_{s=0}^{\tau^+_x-1} f(X_s)\right) = \mathbf{E}_x\left(\sum_{x\in\Omega} f(x)\sum_{s=0}^{\tau^+_x-1}\mathbf{1}_{\{X_s=x\}}\right) = \sum_{x\in\Omega} f(x)\mathbf{E}_x\left(\sum_{s=0}^{\tau^+_x-1}\mathbf{1}_{\{X_s=x\}}\right).$$

Using (3.24) shows that

{Eq:Erg2}

$$\mathbf{E}_x(Y_1) = E_\pi(f)\mathbf{E}_x(\tau^+_x). \tag{5.45}$$

Putting together (5.44) and (5.45) shows that

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

Exercise 5.2 shows that (5.43) holds when $\mu = \delta_x$, the probability distribution with unit mass at x. Averaging over the starting state completes the proof. ■

Taking $f(y) = \delta_x(y) = \mathbf{1}_{\{y=x\}}$ in Theorem 5.11 shows that

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

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

5.8. Problems

{Exer:MaxMeas}

EXERCISE 5.1. Prove that

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

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

{Exercise:SubSeqSum}

EXERCISE 5.2. Let (a_n) be a bounded (deterministic) sequence. If for a sequence of integers (n_k) satisfying $\lim_{k\to\infty} n_k/n_{k+1} = 1$

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

then

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

{Ex:TVDistMonotone}

EXERCISE 5.3. Let P by the transition matrix of a Markov chain with state space Ω, and let μ and ν be any two distributions on Ω. Prove that

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

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

{Exer:TVDecreasing}

Exercise 5.4. Let P be the transition matrix of a Markov chain with stationary distribution π. Prove that for any $t \geq 0$,

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

where $d(t)$ is defined by (5.22).

{Exer:StationaryVsState

Exercise 5.5. Let P be the transition matrix of a Markov chain with stationary distribution π. Prove that for any $t \geq 0$,

$$\bar{d}(t) \leq 2d(t),$$

where $d(t)$ is defined by (5.22) and $\bar{d}(t)$ is defined by (5.23). [Solution]

5.9. Notes

One standard approach to proving the Convergence Theorem for ergodic finite Markov chains is to study the eigenvalues of the transition matrix. See, for instance, Seneta (2006). Eigenvalues are often useful for bounding mixing times, particularly for reversible chains, and we will study them in Chapter 12.

Aldous (1983) (in Lemma 3.5) gives versions of our Lemma 5.8 and Exercises 5.4 and 5.5. He says all these results "can probably be traced back to Doeblin."

CHAPTER 6

Coupling

{Ch:Coupling}

6.1. Definition

{Sec:CouplingDefn}

As we defined in Section 5.1, a *coupling* of two probability distributions μ and ν is a pair of random variables (X, Y), defined on the same probability space, such that the marginal distribution of X is μ and the marginal distribution of Y is ν.

Couplings are useful because we can often make comparisons between distributions by constructing a coupling and comparing the random variables. Proposition 5.5 characterized $\|\mu - \nu\|_{TV}$ as the minimum, over all couplings (X, Y) of μ and ν, of the probability that X and Y are different. This provides a very useful way to get upper bounds on the distance by finding a "good" coupling (X, Y) for which X and Y agree as much as possible.

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

Example 6.1. A simple random walk on $\{0, 1, \dots, n\}$ is a Markov chain which moves either up or down at each move with equal probability. If the walk attempts to move outside the interval when at a boundary point, it stays put. It is intuitively clear that $P^t(x, n) \leq P^t(y, n)$ whenever $x \leq y$, as this says that the chance of being at the "top" value n after t steps doesn't decrease as you increase the height of the starting position.

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

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

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

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

$$P^t(x, n) = \mathbf{P}\{X_t = n\} \leq \mathbf{P}\{Y_t = n\} = P^t(y, n). \tag{6.1}$$

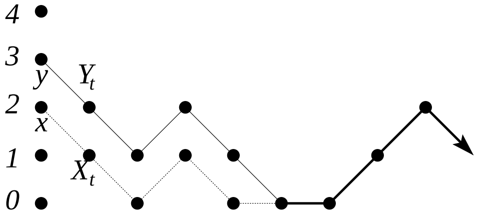


FIGURE 6.1. Coupled random walks on $\{0, 1, 2, 3, 4\}$. The walks stay together after meeting. Fig:CoupledRW

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

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

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

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

{Eq:StayTogether}

$$\text{if } X_s = Y_s \text{ then } X_t = Y_t \text{ for } t \geq s. \tag{6.2}$$

To construct a coupling satisfying (6.2), simply run the chains according to the original coupling until they meet; then run them together.

6.2. Bounding Total Variation Distance

First, we show that the distance between the distributions of the chain started from any two states can be bounded by the meeting time distribution of coupled chains started from those same states. As usual, we will fix a Markov chain with state space Ω, transition matrix P and stationary distribution π.

Thm:CouplingFromStates}

THEOREM 6.2. *Let $\{(X_t, Y_t)\}$ be a coupling satisfying (6.2) for which $X_0 = x$ and $Y_0 = y$. Let τ_{couple} be the first time the chains meet:*

{Eq:CouplingTimeDef}

$$\tau_{\text{couple}} := \min\{t : X_t = Y_t\}. \tag{6.3}$$

Then

$$\left\| P^t(x, \cdot) - P^t(y, \cdot) \right\|_{TV} \leq \mathbf{P}\{\tau_{\text{couple}} > t\}. \tag{6.4}$$

PROOF. Notice that $P^t(x, z) = \mathbf{P}\{X_t = z\}$ and $P^t(y, z) = \mathbf{P}\{Y_t = z\}$. Breaking up the events in these probabilities according whether or not $\tau_{\text{couple}} \leq t$ gives

$$\begin{aligned} P^t(x,z) - P^t(y,z) = \mathbf{P}\{X_t = z, \tau_{\text{couple}} \leq t\} + \mathbf{P}\{X_t = z, \tau_{\text{couple}} > t\} \\ - \mathbf{P}\{Y_t = z, \tau_{\text{couple}} \leq t\} - \mathbf{P}\{Y_t = z, \tau_{\text{couple}} > t\} \end{aligned} \tag{6.5}$$

Now since $X_t = Y_t$ when $\tau_{\text{couple}} \leq t$, the difference $\mathbf{P}\{X_t = z, \tau_{\text{couple}} \leq t\} - \mathbf{P}\{Y_t = z, \tau_{\text{couple}} \leq t\}$ vanishes, and

$$P^t(x,z) - P^t(y,z) = \mathbf{P}\{X_t = z, \tau_{\text{couple}} > t\} - \mathbf{P}\{Y_t = z, \tau_{\text{couple}} > t\}. \tag{6.6}$$

Taking absolute values and summing over z yields

$$\begin{aligned} \left\| P^t(x,\cdot) - P^t(y,\cdot) \right\|_{TV} &\leq \frac{1}{2} \sum_z \left[\mathbf{P}\{X_t = z, \tau_{\text{couple}} > t\} + \mathbf{P}\{Y_t = z, \tau_{\text{couple}} > t\} \right] && (6.7) \\ &= \mathbf{P}\{\tau_{\text{couple}} > t\}. && (6.8) \end{aligned}$$

■

Lemma 5.7, combined with Theorem 6.2 proves the following corollary:

{Cor:Coupling}

Corollary 6.3. *Suppose that for each pair of states x, y there is a coupling (X_t, Y_t) with $X_0 = x$ and $Y_0 = y$. For each such coupling, let τ_{couple} be the first time the chains meet, as defined in* (6.3). *Then*

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

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

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

Clearly any Markovian coupling is indeed a coupling of Markov chains, as we defined in Section 6.1.

Remark. All couplings used in this book will be Markovian.

6.3. Random Walk on the Torus

{Sec:RWTorus}

We defined *random walk on the n-cycle* in Example 3.2. The underlying graph of this walk is called $\mathbb{Z}_n$. It has vertex set $\{1, 2, \ldots, n\}$, with an edge between j and k if $j \equiv k \pm 1 \mod n$. See Figure 3.3. The *d-dimensional torus* is the Cartesian product

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

Vertices $x = (x^1, \ldots, x^d)$ and $y = (y^1, y^2, \ldots, y^d)$ are neighbors in $\mathbb{Z}_n^d$ if for some $j \in \{1, 2, \ldots, n\}$, we have $x^i = y^i$ for all $i \neq j$ and $x^j \equiv y^j \pm 1 \mod n$. See Figure 6.2 for an example.

When n is even, the graph $\mathbb{Z}_n^d$ is bipartite and the associated random walk is periodic. To avoid this complication, we consider the lazy random walk on $\mathbb{Z}_n^d$, defined in Section 3.3, which remains still with probability $1/2$ at each move.

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

{Thm:RWTorus}

Theorem 6.4. *For the lazy random walk on the d-dimension torus $\mathbb{Z}_n^d$,*

$$t_{\text{mix}}(\varepsilon) = O\left(c(d) n^2 \log_2(\varepsilon^{-1}) \right), \tag{6.9}$$

{Eq:Tau1RWTorus}

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

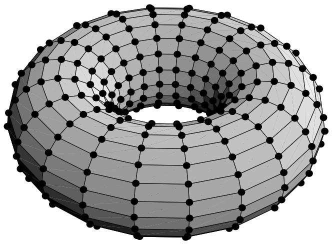


FIGURE 6.2. The 2-torus $\mathbb{Z}_{20}^2$. Fig:Torus

In order to apply Corollary 6.3 to prove Theorem 6.4, we construct a coupling for each pair (x, y) of starting states and bound the coupling time $\tau_{\text{couple}} = \tau_{x,y}$.

To couple together a random walk (X_t) started at x with a random walk (Y_t) started at y, first pick one of the d coordinates at random. If the two chains agree in the chosen coordinate, we move both of the chains by $+1$, -1, or 0 in that coordinate. If the two chains differ in the chosen coordinate, we randomly choose one of the chains to move, leaving the other fixed. We then move the selected chain by $+1$ or -1 in the chosen coordinate.

Let τ_i be the time required for coordinate i to agree in both chains. Each time coordinate i is selected, the clockwise distance of the chain started at x to the chain started at y either increases or decreases by 1, with equal probability. This distance, when observed at the times that coordinate i is selected, is then a random walk on $\{0, 1, 2, \dots, n\}$, with absorption at 0 and n. You should recognize this situation as the "gambler's ruin" discussed in Section 4.1. Proposition 4.1 implies that the expected time to couple is at most $n^2/4$, regardless of starting distance.

Since coordinate i is selected with probability $1/d$ at each move, there is a geometric waiting time between moves with expectation d. Exercise 6.3 implies that

$$\mathbf{E}(\tau_i) \leq \frac{dn^2}{4}. \tag{6.10}$$

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

{Eq:CoupleTimeForTorus}

$$\mathbf{E}(\tau_{\text{couple}}) \leq \frac{d^2 n^2}{4}. \tag{6.11}$$

This time is independent of the starting states, and we can use Markov's inequality to get

$$\mathbf{P}\{\tau_{\text{couple}} > t\} \leq \frac{\mathbf{E}(\tau_{\text{couple}})}{t} \leq \frac{1}{t}\frac{d^2 n^2}{4} \tag{6.12}$$

Taking $t_0 = d^2 n^2$ shows that $d(t_0) \leq 1/4$, and so $t_{\text{mix}} \leq d^2 n^2$. By Equation 5.36,

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

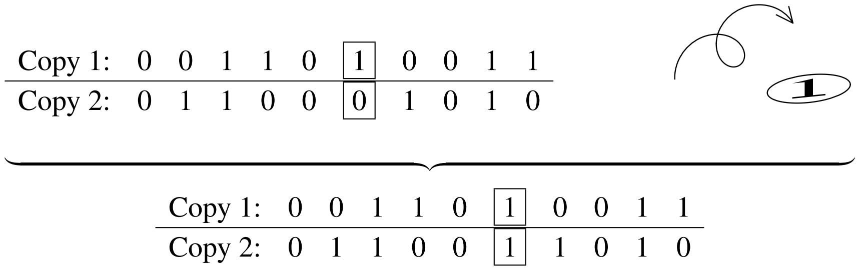

Copy 1: 0 0 1 1 0 1 0 0 1 1
Copy 2: 0 1 1 0 0 0 1 0 1 0
1
Copy 1: 0 0 1 1 0 1 0 0 1 1
Copy 2: 0 1 1 0 0 1 1 0 1 0

Fig:HCCoup

FIGURE 6.3. One step in two coupled lazy walks on the hypercube. First, choose a coordinate to update—here, the sixth. Then, flip a 0/1 coin and use the result to update the chosen coordinate to the same value in both walks.

and we have proved Theorem 6.4.

Exercise 6.4 shows that the bound on $c(d)$ can be improved.

6.4. Random Walk on the Hypercube

{Sec:CouplingRWHC}

The simple random walk hypercube $\{0,1\}^n$ was defined in Section 4.3.2: this is the simple walker on the graph having vertex set $\{0,1\}^n$ – the binary words of length n – and with edges connecting words differing in exactly one letter.

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

As remarked in Section 4.3.2, a convenient way to generate the lazy walk is as follows: pick one of the n coordinates uniformly at random, and *refresh* the bit at this coordinate with a random fair bit (one which equals 0 or 1 each with probability $1/2$).

This algorithm for running the walk leads to the following coupling of two walks with possibly different starting positions: First, pick among the n coordinates uniformly at random; suppose that coordinate i is selected. *In both walks*, replace the bit at coordinate i *with the same* random fair bit. (See Figure 6.3.) From this time onwards, both walks will agree in the ith coordinate. A moment's thought reveals that individually each of the walks is indeed a lazy random walker on the hypercube.

If τ is the first time when all of the coordinates have been selected at least once, then the two walkers agree with each other from time τ onwards. (If the initial states agree in some coordinates, the first time the walkers agree could be strictly before τ.) The distribution of τ is exactly the same as the coupon collector random variable studied in Section 4.2. In particular, $\mathbf{E}(\tau) = n\sum_{k=1}^{n} k^{-1} \leq n(\log n + 1)$. Using Corollary 6.3 shows that

$$d(t) \leq \mathbf{P}\{\tau > t\} \leq \frac{\mathbf{E}(\tau)}{t} \leq \frac{n(\log n + 1)}{t}.$$

Thus, (5.36) yields

$$t_{\text{mix}}(\varepsilon) \leq 4n(\log n + 1)\left\lceil \log_2(\varepsilon^{-1}) \right\rceil. \tag{6.13}$$ {Eq:BadHCBound}

Simply, $t_{\text{mix}} = O(n \log n)$. The bound in (6.13) will be sharpened in Section 8.5 via a more complicated coupling.

6.5. Problems

er:CouplingConvergence}

EXERCISE 6.1. A mild generalization of Theorem 6.2 can be used to give an alternative proof of the Convergence Theorem.

(a) Show that when (X_t, Y_t) is a coupling satisfying (6.2) for which $X_0 \sim \mu$ and $Y_0 \sim \nu$, then

{Eq:CplCnvThm}
$$\left\|\mu P^t - \nu P^t\right\|_{TV} \leq \mathbf{P}\{\tau_{\text{couple}} > t\}. \tag{6.14}$$

(b) If in (a) we take $\nu = \pi$, where π is the stationary distribution, then (by definition) $\pi P^t = \pi$, and (6.14) bounds the difference between μP^t and π. The only thing left to check is that there exists a coupling guaranteed to coalesce, that is, for which $\mathbf{P}\{\tau_{\text{couple}} < \infty\} = 1$. Show that if the chains (X_t) and (Y_t) are taken to be independent of one another then they are assured to eventually meet.

cise:MarkovianCoupling}

EXERCISE 6.2. Let (X_t, Y_t) be a Markovian coupling such that for some $0 < \alpha < 1$ and some $t_0 > 0$, the coupling time $\tau_{\text{couple}} = \min\{t \geq 0 : X_t = Y_t\}$ satisfies $\mathbf{P}\{\tau_{\text{couple}} \leq t_0\} \geq \alpha$ for *all* pairs of initial states (x, y). Prove that

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

{Exer:WeakWald}

EXERCISE 6.3. Show that if $X_1, X_2, \ldots$ are independent and each have mean μ, and τ is a $\mathbb{Z}^+$-valued random variable independent of all the X_i's, then

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

{Exer:BetterTorusBound}

EXERCISE 6.4. We can get a better bound on the mixing time for the lazy walker on the d-dimensional torus by sharpening the analysis of the "coordinate-by-coordinate" coupling given in the proof of Theorem 6.4.

Let $t \geq kdn^2$.

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

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

6.6. Notes

For many examples of coupling, a good reference is Lindvall (2002).

CHAPTER 7

Strong Stationary Times

{Ch:SST}

7.1. Two Examples

{Sec:T2R1}

7.1.1. The top-to-random shuffle. Consider the following (slow) method of shuffling a deck of n cards: Take the top card and insert it uniformly at random in the deck. This process will eventually mix up the deck – the successive arrangements of the deck is a Markov chain on the $n!$ possible orderings of the cards, with uniform stationary distribution. (See Exercise 7.1.)

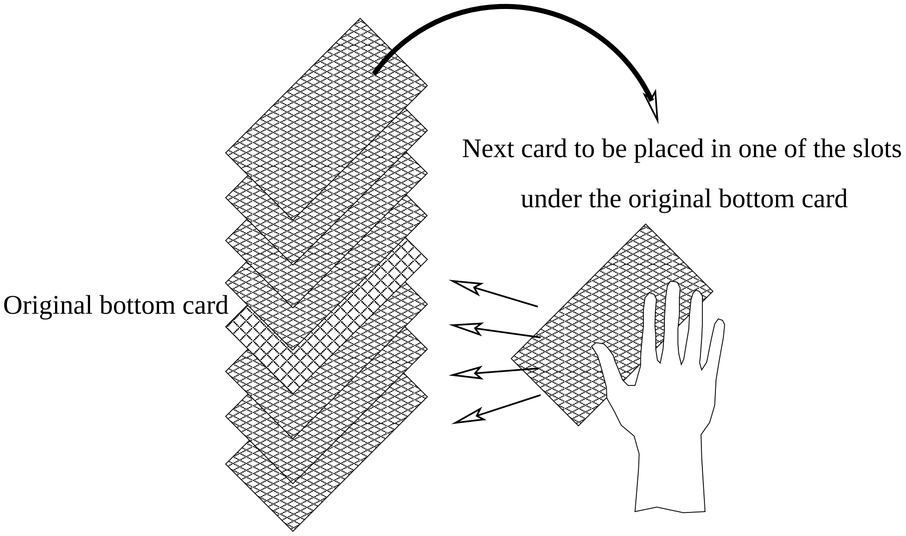

Next card to be placed in one of the slots
under the original bottom card
Original bottom card

FIGURE 7.1. The top-to-random shuffle. Fig:TopToRandom

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

Let τ be the time one move after the first occasion when the original bottom card has moved to the top of the deck. We show now that the arrangement of cards at time τ is distributed uniformly on the set of all permutations of $\{1, \ldots, n\}$. More generally, we argue that when there are k cards under the original bottom card, then all $k!$ orderings of these k cards are equally likely.

This can be seen by induction. When $k = 1$, the conclusion is obvious. Suppose that there are $(k - 1)$ cards under the original bottom card, and that each of the $(k - 1)!$ arrangements are equally probable. The next card to be inserted below the original bottom card is equally likely to land in any of the k possible positions, and by hypothesis, the remaining $(k - 1)$ cards are in random order. We conclude that all $k!$ arrangements are equally likely.

{Sec:RWHC_ST_I}

7.1.2. Random walk on the hypercube. We have met already the lazy random walk on the hypercube $\{0, 1\}^n$ in Section 4.3.2. Recall that a move of this walk can be executed by choosing among the n coordinate at random, and replacing the bit at the selected location by an independent fair bit.

Let τ be the first time that each of the coordinates has been selected at least once. Since all the bits at this time have been replaced by independent fair coin tosses, the distribution of the state of the chain at τ is uniform on $\{0, 1\}^n$, and independent of the value of τ.

In both of these examples, we found an "online algorithm" for when to stop the chain so that the stopping state is distributed exactly according to the stationary distribution π.

It should not be too surprising that bounding the size of τ (in distribution) bounds the mixing time of the chain, the *fixed* time required before the distribution of the chain is near the stationary distribution.

The random times τ in these two examples are both *strong stationary times*. Before we can give a precise definition, we first must understand stopping times.

7.2. Stopping in the Stationary Distribution

{Sec:Stop}

7.2.1. Stopping times. A friend gives you directions to his house, telling you to take Main street and to turn left at the first street after City Hall. These are acceptable directions, because you are able to determine when to turn using landmarks you have already encountered before the turn. This is an example of a *stopping time*, which is an instruction for when to "stop" depending only on information up until the turn.

On the other hand, his roommate also provides directions to the house, telling you to take Main street and turn left at the last street *before* you reach a bridge. You have never been down Main street, so not knowing where the bridge is located, you unfortunately must drive past the turn before you can identify it. Once you reach the bridge, you must backtrack. This is *not* a stopping time, you must go past the turn before recognizing it.

We now provide a precise definition for a stopping time. Let $(Y_t)_{t=0}^{\infty}$ be a sequence of random variables taking values in the space Λ, which we assume to be either a finite set or $\mathbb{R}^d$. Another sequence (Z_t) with values in Λ' is said to be *adapted* to (Y_t) if for each t there exists a function $f_t : \Lambda^{t+1} \to \Lambda'$ so that

$$Z_t = f_t(Y_0, Y_1, \ldots, Y_t).$$

Example 7.1. Let (Y_t) be an i.i.d. sequence of mean-zero $\{-1, +1\}$-valued random variables, and let $S_t = \sum_{s=1}^{t} Y_s$. The sequence (S_t) is adapted to (Y_t).

In this set-up, the sequence (Y_t) is the fundamental source of noise, while we may be primarily interested in a sequence (Z_t) which is built from this source of randomness.

A *stopping time* for (Y_t) is a random time τ with values in $\{0, 1, 2, \ldots, \} \cup \{\infty\}$ such that $(\mathbf{1}_{\{\tau=t\}})$ is adapted to (Y_t). (The random variable $\mathbf{1}_A$ is the *indicator random variable* for the event A, i.e. the $\{0, 1\}$-valued random variable which equals 1 if and only if A occurs.) For a stopping time τ, the event $\{\tau = t\}$ is determined by the vector $(Y_0, Y_1, \ldots, Y_n)$.

Example 7.2 (Hitting times). Let A be a subset of Ω. The history up to time t suffices to determine whether a site in A is visited for the first time at time t. That is, if

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

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

$$\mathbf{1}_{\{\tau_A=t\}} = \mathbf{1}_{\{Y_0 \notin A, Y_1 \notin A, \ldots, Y_{t-1} \notin A, Y_t \in A\}}.$$

The right-hand side is a function of $(Y_0, Y_1, \ldots, Y_t)$, whence $(\mathbf{1}_{\{\tau_A=t\}})$ is adapted to (Y_t) and τ_A is a stopping time.

An example of a random time which is *not* a stopping time is the first time that the sequence reaches its maximum value over a time interval $\{0, 1 \ldots, t_1\}$:

$$M = \min\left\{t \, : \, Y_t = \max_{1 \leq s \leq t_1} Y_s\right\}. \tag{7.1}$$

It is impossible to check whether $M = t$ by looking only at the first t values of the sequence. Indeed, any investor hopes to sell a stock at the time M when it achieves its maximum value. Alas, this would require clairvoyance—the ability to see into the future—and is *not* a stopping time.

7.2.2. Achieving equilibrium. Let (X_t) be a Markov chain which is adapted to the sequence of random variables (Y_t). A *strong stationary time* for a Markov chain (X_t) is a stopping time τ for (Y_t) such that X_τ, the chain sampled at τ, has two properties: first, the law of X_τ is exactly the stationary distribution of the chain, and second, the value of X_τ is independent of τ. That is, for all $t = 0, 1, 2, \ldots$,

$$\mathbf{P}\{X_t = x, \ \tau = t\} = \pi(x)\mathbf{P}\{\tau = t\}. \tag{7.2}$$

{Eq:SSTDefn}

Strong stationary times were introduced in Aldous and Diaconis (1987); see also Aldous and Diaconis (1986).

We will later need the following strengthening of equation (7.2): If τ is a strong stationary time, then

$$\mathbf{P}\{X_t = x, \ \tau \leq t\} = \pi(x)\mathbf{P}\{\tau \leq t\} \tag{7.3}$$

{Eq:SSTFuture}

To see this, if $s \leq t$ and $\mathbf{P}\{T = s\} > 0$, then

$$\begin{aligned} \mathbf{P}\{X_t = x, \ T = s\} &= \sum_{y \in \Omega} \mathbf{P}\{X_t = x \mid X_s = y, \ T = s\}\mathbf{P}\{X_s = y, \ T = s\} \\ &= \sum_{y \in \Omega} P^{t-s}(y, x)\pi(y)\mathbf{P}\{T = s\}. \end{aligned} \tag{7.4}$$

{Eq:SSTStrong}

Since π satisfies $\pi = \pi P^{t-s}$, the right-hand side of (7.4) equals $\pi(x)\mathbf{P}\{T = s\}$. Summing over $s \leq t$ establishes (7.3).

7.3. Bounding Convergence using Strong Stationary Times

Throughout this section, we discuss a Markov chain (X_t) with transition matrix P and stationary distribution π. The route from strong stationary times to bounding convergence time is the following proposition:

{Prop:SSTBound}

Proposition 7.3. *If τ is a strong stationary time, then*

{Eq:TVSST}

$$d(t) = \max_{x \in \Omega} \|P^t(x, \cdot) - \pi\|_{\mathrm{TV}} \leq \max_{x \in \Omega} \mathbf{P}_x\{\tau > t\}. \tag{7.5}$$

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

{Eq:SepDef}

$$s(t) := \max_{x,y \in \Omega} \left[1 - \frac{P^t(x,y)}{\pi(y)} \right]. \tag{7.6}$$

The relationship between $s(t)$ and τ is:

{Lem:SepAndSST}

Lemma 7.4. *If τ is a strong stationary time, then*

{Eq:SepUB}

$$s(t) \leq \max_{x \in \Omega} \mathbf{P}_x\{\tau > t\}. \tag{7.7}$$

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

$$1 - \frac{P^t(x,y)}{\pi(y)} = 1 - \frac{\mathbf{P}_x\{X_t = y\}}{\pi(y)} \leq 1 - \frac{\mathbf{P}_x\{X_t = y,\ \tau \leq t\}}{\pi(y)}. \tag{7.8}$$

By Equation 7.3, the right-hand side is bounded above by

$$1 - \frac{\pi(y)\mathbf{P}_x\{\tau \leq t\}}{\pi(y)} = \mathbf{P}_x\{\tau > t\}. \tag{7.9}$$

■

The next lemma along with Lemma 7.4 proves (7.5).

{Lem:TVSep}

Lemma 7.5. $d(t) \leq s(t)$.

Proof. Writing

$$\|P^t(x,\cdot) - \pi\|_{\mathrm{TV}} = \sum_{\substack{y \in \Omega \\ P^t(x,y) < \pi(y)}} \left[\pi(y) - P^t(x,y) \right] = \sum_{\substack{y \in \Omega \\ P^t(x,y) < \pi(y)}} \pi(y) \left[1 - \frac{P^t(x,y)}{\pi(y)} \right], \tag{7.10}$$

we conclude that

$$\|P^t(x,\cdot) - \pi\|_{\mathrm{TV}} \leq \max_{y \in \Omega} \left[1 - \frac{P^t(x,y)}{\pi(y)} \right]. \tag{7.11}$$

■

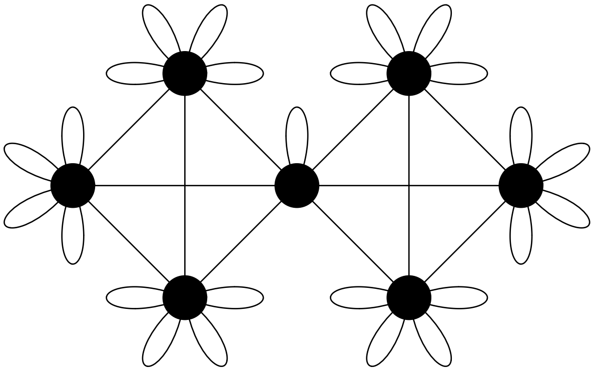


FIGURE 7.2. Two complete graphs (on 4 vertices), “glued” at a single vertex. Loops have been added so that every vertex has the same degree (count each loop as one edge). Fig:TwoComplete

7.4. Examples

{Sec:TwoComplete}

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

Let $v^\star$ be the vertex where the two complete graphs meet. The degree at $v^\star$ has degree $2n-2$, while the degree at every other vertex has degree $n-1$. We modify the graph to make it regular and to have holding probabilities, by adding 1 loop at $v^\star$ and n loops at at all other vertices. See Figure 7.2 for an illustration when $n = 4$. The degree of every vertex is $2n - 1$. Since the graph is regular, the stationary distribution is uniform.

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

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

$$\mathbf{E}(\tau) = 2n \tag{7.12}$$ {Eq:TwoCompExp}

Using Markov’s inequality and (7.12) shows that

$$\mathbf{P}_x\{\tau \geq t\} \leq \frac{2n}{t}. \tag{7.13}$$ {Eq:TTwoK}

Taking $t = 8n$ in (7.13) and applying Proposition 7.3 shows that

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

A lower bound on t_{mix} of order n is obtained in Exercise 7.11.

{Section:HC}

7.4.2. Random walk on the hypercube. We return to the lazy random walker on $\{0,1\}^n$, discussed in Section 7.1.2. The time τ when each coordinate has been selected at least once for the first time is a strong stationary time. This stopping time and the coordinate-by-coordinate coupling used in Section 6.4 are closely related: the coupon collector's time of Section 4.2 dominates the coupling time and has the same distribution as τ. It is therefore not surprising that we obtain here exactly the same upper bound for $t_{\rm mix}$ as was found using the coupling method. In particular, combining Proposition 4.3 and Proposition 7.3 give the bound $t_{\rm mix}(\varepsilon) \leq n \log n + \log(\varepsilon)n$.

{Sec:TtoRUpper}

7.4.3. Top-to-random shuffle. Revisiting the top-to-random shuffle introduced in Section 7.1.1, the time τ when the original bottom card is first placed in the deck after rising to the top is a strong stationary time.

Consider the motion of the original bottom card. When there are k cards beneath it, the chance that it rises one card remains k/n until a shuffle puts the top card underneath it. Thus, the distribution of τ is the same as the coupon collector's time. As above for the lazy hypercube walker, combining Proposition 7.3 and Proposition 4.3 yields

{eq.t2rdub}

$$d(n \log n + \alpha n) \leq e^{-\alpha} \quad \text{for all } n. \tag{7.14}$$

Consequently,

{eq.t2rmtub}

$$t_{\rm mix}(\varepsilon) \leq n \log n + \log(\varepsilon)n. \tag{7.15}$$

{Sec:Transitive}{Sect:MTF}

7.5. The Move-to-Front Chain

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

If his choice of book is random, this is an example of the *move-to-front* chain. It is a very natural chain which arises in many applied contexts. Any setting where items are stored in a stack, removed at random locations, and placed on the top of the stack can be modeled by the move-to-front chain.

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

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

By Lemma 5.9, the mixing time for move-to-front will be identical to that of the top-to-random shuffle. Consequently, the mixing time for move-to-front is not more than $n \log n - log(\varepsilon)n$.

{Exercise:T2R}

7.6. Problems

Exercise 7.1. Show that the top-to-random shuffle just described is a Markov chain with stationary distribution uniform on the $n!$ card arrangements. [Solution]

Exercise 7.2. Show that the time until the card initially one card from the bottom rises to the top, plus one more move, is a strong stationary time, and find its expectation.

Drawing by Yelena Shvets

{Exercise:GluedKn}

EXERCISE 7.3. Show that for the Markov chain on two complete graphs in Section 7.4.1, the stationary distribution is uniform on all $2n - 1$ vertices.

{Exer:TorusTransitive}

EXERCISE 7.4. Show the lazy random walk on the torus (Section 6.3) is transitive. [SOLUTION]

{Exer:TransitiveUniform

EXERCISE 7.5. Show that the stationary distribution of a transitive chain must be uniform.

{Exercise:ReversedChain

EXERCISE 7.6. Let (X_t) be a Markov chain with transition matrix P, and write $(\widehat{X}_t)$ for the time-reversed chain with the matrix $\widehat{P}$ defined in (3.30).

(a) Check that π is stationary for $\widehat{P}$.

(b) Show that

$$\mathbf{P}_\pi\{X_0 = x_0, \ldots, X_t = x_t\} = \mathbf{P}_\pi\{\widehat{X}_0 = x_t, \ldots, \widehat{X}_t = x_0\}. \tag{7.16}$$

[SOLUTION]

{Exercise:RevTrans}

EXERCISE 7.7. Show that if P is transitive, then $\widehat{P}$ is also transitive.

{Exercise:SepIsSubM}

EXERCISE 7.8. Let $s(t)$ be defined as in (7.6).

{It:Decomp}

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

{It:SepSubM1}

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

{Eq:SepSubM1}

$$P^{t+u}(x,y) = [1 - s(t)s(u)]\,\pi(y) + s(t)s(u)\sum_{z\in\Omega} Q^t(x,z)Q^u(z,y). \tag{7.17}$$

(c) Using (7.17) establish that s is submultiplicative: $s(t+u) \leq s(t)s(u)$.

[Solution]

{Exer:SSTGeo}

Exercise 7.9. Show that if $\max_{x\in\Omega} \mathbf{P}_x\{\tau > t_0\} \leq \varepsilon$, then $d(t) \leq \varepsilon^{t/t_0}$. [Solution]

{Exercise:WaldFull}

Exercise 7.10 (Wald's Identity). Let (Y_t) be a sequence of independent and identically distributed random variables.

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

{Eq:WaldFull}

$$\mathbf{E}\left(\sum_{t=1}^{\tau} Y_t\right) = \mathbf{E}(\tau)\mathbf{E}(Y_1). \tag{7.18}$$

Hint: Write $\sum_{t=1}^{\tau} Y_t = \sum_{t=1}^{\infty} Y_t \mathbf{1}_{\{\tau \geq t\}}$.

(ii) Let τ be a stopping time for the sequence (Y_t). Show that $\{\tau \geq t\}$ is independent of Y_{t+1}, so (7.18) holds provided that $\mathbf{E}(\tau) < \infty$.

[Solution]

ercise:TwoKnLowerBound}

Exercise 7.11. Consider the Markov chain of Section 7.4.1 defined on two glued complete graphs. By considering the set $A \subset \Omega$ of all vertices in one of the two complete graphs, show that $t_{\mathrm{mix}} \geq (n/2)\,[1 + o(1)]$.

7.7. Notes

References on strong uniform times are Aldous and Diaconis (1986) and Aldous and Diaconis (1987).

A state x is a *halting state* for a stopping time τ if $X_t = x$ implies $\tau \leq t$. Lovasz and Winkler showed that a stationary time has minimal expectation among all stationary times if and only if it has a halting state.

CHAPTER 8

Lower Bounds on Mixing Times and Cut-Off

{Ch:Lower}

8.1. Diameter Bound

Suppose that (X_t) is a random walk on a graph with vertex set Ω. If the possible locations of the walker after t steps are not a significant fraction of Ω, then the distribution of her position at time t cannot be close to stationary. We can make this precise.

Define the *diameter* of a graph with vertex set Ω to be the maximum distance between two vertices:

$$\text{diam} = \max_{x,y\in\Omega} \rho(x,y). \tag{8.1}$$

{Eq:DiamDefn}

(The distance $\rho(x,y)$ between vertices x and y in a graph is the minimum length of a path connecting x and y.) Note that if x_0 and y_0 are vertices with $\rho(x_0,y_0) =$ diam, then $P^{(\text{diam}-1)/2}(x_0,\cdot)$ and $P^{(\text{diam}-1)/2}(y_0,\cdot)$ are positive on disjoint vertex sets. Consequently, $\bar{d}((\text{diam}-1)/2) = 1$ and for any $\varepsilon < 1/2$,

$$t_{\text{mix}}(\varepsilon) \geq \frac{\text{diam}}{2}. \tag{8.2}$$

{Eq:TMixDiameter}

8.2. Bottleneck Ratio

{Sec:BR}

Bottlenecks in the state-space Ω of a Markov chain are geometric features that control mixing time. A bottleneck makes portions of Ω difficult to reach from some starting locations, limiting the speed of convergence. See Figure 8.1 for the illustration of a graph having an obvious bottleneck.

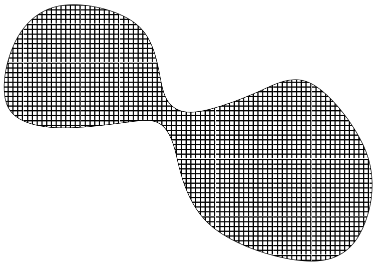


FIGURE 8.1. A graph with a bottleneck. Fig:Bottleneck

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

The *edge measure* Q is defined by

$$Q(x,y) := \pi(x)P(x,y), \quad Q(A,B) = \sum_{x \in A, y \in B} Q(x,y). \tag{8.3}$$

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

The *bottleneck ratio* of the set S is defined as

$$\Phi(S) := \frac{Q(S,S^c)}{\pi(S)}, \tag{8.4}$$

and the bottleneck ratio of the whole chain is

{Eq:BNDefn}

$$\Phi_\star := \min_{S \,:\, \pi(S) \leq \frac{1}{2}} \Phi(S). \tag{8.5}$$

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

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

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

{Eq:BNRSRW}

$$\Phi(S) = \frac{|\partial S|}{\sum_{x \in S} \deg(x)}. \tag{8.6}$$

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

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

{Thm:CheegerLower}

Theorem 8.1. *If $\Phi_\star$ is the bottleneck ratio defined in* (8.5)*, then*

{Eq:CheegerLower}

$$t_{\text{mix}} = t_{\text{mix}}(1/4) \geq \frac{1}{4\Phi_\star}. \tag{8.7}$$

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

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

From Remark 5.1,

{Eq:Scompliment}

$$\pi(S)\,\|\mu_S P - \mu_S\|_{\text{TV}} = \pi(S) \sum_{\substack{y \in \Omega, \\ \mu_S P(y) \geq \mu_S(y)}} \left[\mu_S P(y) - \mu_S(y)\right]. \tag{8.8}$$

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

{Eq:SCompliment2}

$$\pi(S)\,\|\mu_S P - \mu_S\|_{\text{TV}} = \sum_{\substack{y \in \Omega, \\ \pi_S P(y) \geq \pi_S(y)}} \left[\pi_S P(y) - \pi_S(y)\right]. \tag{8.9}$$

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

{Eq:PiSForward}

$$\pi_S P(y) = \sum_{x \in \Omega} \pi_S(x) P(x, y) = \sum_{x \in S} \pi(x) P(x, y) \leq \sum_{x \in \Omega} \pi(x) P(x, y) = \pi(y). \tag{8.10}$$

Again using that $\pi(y) = \pi_S(y)$ for $y \in S$, from (8.10) follows the inequality

$$\pi_S P(y) \leq \pi_S(y) \quad \text{for } y \in S. \tag{8.11}$$

{Eq:Bigger}

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

$$\pi_S P(y) \geq 0 = \pi_S(y) \quad \text{for } y \in S^c. \tag{8.12}$$

{Eq:BiggerSc}

Combining (8.11) and (8.12) shows the the sum on the right in (8.9) can be taken over S^c:

$$\pi(S) \|\mu_S P - \mu_S\|_{TV} = \sum_{y \in S^c} [\pi_S P(y) - \pi_S(y)]. \tag{8.13}$$

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

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

Dividing by $\pi(S)$,

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

By Exercise 5.3, for any $u \geq 0$,

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

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

$$\left\|\mu_S P^t - \mu_S\right\|_{TV} \leq t\Phi(S). \tag{8.14}$$

{Eq:TVtPhi}

Assume that $\pi(S) \leq \frac{1}{2}$. In this case,

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

Also,

$$\frac{1}{2} \leq \|\mu_S - \pi\|_{TV} \leq \left\|\mu_S - \mu_S P^t\right\|_{TV} + \left\|\mu_s P^t - \pi\right\|_{TV}. \tag{8.15}$$

{Eq:CLB1}

Taking $t = t_{\mathrm{mix}} = t_{\mathrm{mix}}(1/4)$ in (8.15), by definition of t_{mix} and using (8.14),

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

Rearranging and minimizing over S establishes (8.7). ■

{Example:TwoTorLB}

Example 8.2 (Two glued tori). Consider the graph of two tori "glued" together at a single vertex. This graph is a pair of two-dimensional tori sharing exactly one common node, which we label $v^\star$; see Figure 8.2. Denote by V_1 and V_2 the vertices in the right and left tori, respectively.

The set ∂V_1 consists of all edges $\{v^\star, v\}$, where $v \in V_2$. The size of ∂V_1 is $2d$. Also, $\sum_{x \in V_1} \deg(x) = 2dn^2$. Consequently,

$$\Phi_\star \leq \Phi(V_1) = \frac{2d}{2dn^2} = n^{-2}.$$

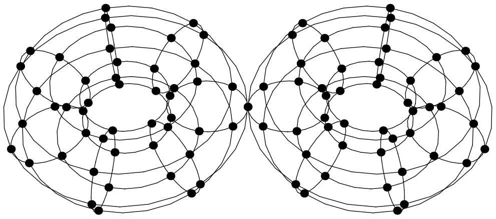


FIGURE 8.2. Two "glued" tori. Fig:TwoTori

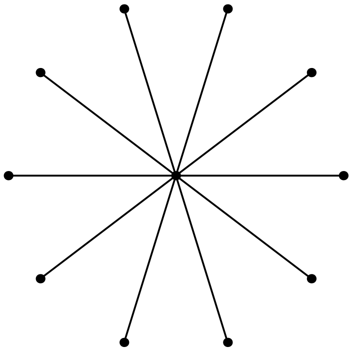


FIGURE 8.3. The star graph with 11 vertices. Fig:StarGraph

Theorem 8.1 implies that $t_{\text{mix}} \geq n^2/4$. We return to this example in Section 11.7, where it is proved that $t_{\text{mix}} \asymp n^2 \log n$. Thus the lower bound here does not give the correct order.

{Example:ColorStar}

EXAMPLE 8.3 (Coloring the star). Recall that a proper q-coloring of a graph G with vertex set V and edge set E is a map $x : V \to \{1, 2, \ldots, q\}$ so that $x(v) \neq x(w)$ for all $\{v, w\} \in E$. (See Section 14.3.1.) Ω is the set of all proper q-colorings of G, and π is the uniform distribution on Ω. The Glauber dynamics for π is the Markov chain which makes transitions as follows: At each unit of time, a vertex is chosen from V uniformly at random, and the color at this vertex is chosen uniformly at random from all *feasible colors*. The feasible colors at vertex v are all colors *not* present among the neighbors of v.

In Chapter 14, an upper bound on t_{mix} is proven when there are an abundance of colors relative to the maximum degree of the graph. (Cf. Section 14.3.3.) In that case, the Glauber dynamics is *fast mixing*, meaning that t_{mix} is polynomial in $|V|$. (Note that $|V|$ is much smaller than $|\Omega|$.)

Here we show by example that if the maximum degree is growing in n while the number of colors q is fixed, then the mixing time grows at least exponentially in $|V|$.

The graph we study here is the *star*, as shown in Figure 8.3. This graph is a tree of depth 1.

Let $v_\star$ denote the root vertex, and let S be defined as the set of proper colorings so that $v_\star$ has color 1:

$$S = \{x \in \Omega \ : \ x(v_\star) = 1\}.$$

Since the constraint $x(v_\star) = 1$ means that each leaf can be colored with any of the remaining $q - 1$ colors, $|S| = (q-1)^{n-1}$. For $(x, y) \in S \times S^c$, the transition probability $P(x, y)$ is non-zero if and only if all leaves v satisfy $x(v) = y(v)$ and $x(v) \notin \{x(v_\star), y(v_\star)\}$. It follows that

$$\sum_{x \in S, y \in S^c} Q(x, y) \le \frac{1}{|\Omega|}\frac{1}{n}(q-1)(q-2)^{n-1},$$

and

$$\frac{Q(S, S^c)}{\pi(S)} \le \frac{(q-1)^2}{n(q-2)}\left(1 - \frac{1}{q-1}\right)^n \le \frac{(q-1)^2}{n(q-2)} e^{-n/(q-1)}.$$

Consequently, the mixing time is at least of exponential order:

$$t_{\text{mix}} \ge \frac{n(q-2)}{4(q-1)^2} e^{n/(q-1)}.$$

Remark 8.1. In fact, this argument shows that if $n/(q \log q) \to \infty$, then t_{mix} is super-polynomial in n.

{Xmple:BinTreeLB}

Example 8.4 (Binary Tree). A *rooted binary tree of depth* k, denoted by $\mathcal{T}_{2,k}$, is a tree with a distinguished vertex v_0, the root, so that

- v_0 has degree 2,
- every vertex at distance j from the root, where $1 \le j \le k - 1$, has degree 3,
- the vertices at distance k from v_0, called *leaves*, have degree 1.

There are $n = 2^{k+1} - 1$ vertices in $\mathcal{T}_{2,k}$.

In this example, we consider the lazy random walk on $\tilde{\mathcal{T}}_{2,k}$; this walk remains at its current position with probability $1/2$.

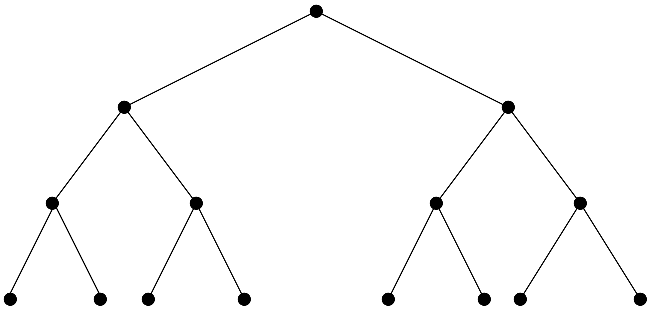


Figure 8.4. A binary tree of height 3. Fig:BTree

Label the vertices adjacent to v_0 as v_r and v_ℓ. Call w a *descendent* of v if the shortest path from w to v_0 passes through v. Let S consist of the right-hand side of the tree, that is, v_r and all of its descendants.

We write $|v|$ for the length of the shortest path from v to v_0. The stationary distribution is

$$\pi(v) = \begin{cases} \frac{2}{2n-1} & \text{for } v = v_0, \\ \frac{3}{2n-1} & \text{for } 0 < |v| < k, \\ \frac{1}{2n-1} & \text{for } |v| = k. \end{cases}$$

Adding $\pi(v)$ over $v \in S$ shows that $\pi(S) = (n-2)/(2n-1)$. Since there is only one edge from S to S^c,

$$Q(S, S^c) = \pi(v_r)P(v_r, v_0) = \left(\frac{3}{2n-1}\right)\frac{1}{6} = \frac{1}{2(2n-1)},$$

and so

$$\Phi(S) = \frac{1}{2n-4}.$$

Applying Theorem 8.1 establishes the lower bound

$$t_{\text{mix}} \geq \frac{n-2}{2}.$$

8.3. Distinguishing Statistics

One way to produce a lower bound on the mixing time t_{mix} is to find a statistic f (a real-valued function on Ω) so that the distance between the distribution of $f(X_t)$ and the distribution of f under the stationary distribution π can be bounded from below.

Let μ and ν be two probability distributions on Ω, and let f be a real-valued function defined on Ω. We write E_μ to indicate expectations of random variables (on sample space Ω) with respect to the probability distribution μ:

$$E_\mu(f) = \sum_{x \in \Omega} f(x)\mu(x).$$

(Note the distinction between E_μ with $\mathbf{E}_\mu$, the expectation operator corresponding to the Markov chain (X_t) started with μ.) Likewise $\text{Var}_\mu(f)$ indicates variance computed with respect to the probability distribution μ.

op:ChebyshevLowerBound}

Proposition 8.5. *Let μ and ν be two probability distributions on Ω, and f a real-valued function on Ω. If*

{Eq:SepBySDs}

$$|E_\mu(f) - E_\nu(f)| \geq r\sigma, \tag{8.16}$$

where $\sigma^2 = \max\{\text{Var}_\mu(f), \text{Var}_\nu(f)\}$, *then*

{Eq:TVLB}

$$\|\mu - \nu\|_{TV} \geq 1 - \frac{4}{4 + r^2}. \tag{8.17}$$

Before proving this, we provide a useful lemma:

{Lem:TVProj}

Lemma 8.6. *Let μ and ν be probability distributions on Ω, and let $f : \Omega \to \Lambda$ be a function on Ω, where Λ is a finite set. Write μf^{-1} for the probability distribution on Λ defined by $(\mu f^{-1})(A) := \mu(f^{-1}(A))$ for $A \subset \Lambda$. Then*

$$\|\mu - \nu\|_{\mathrm{TV}} \geq \left\|\mu f^{-1} - \nu f^{-1}\right\|_{\mathrm{TV}} .$$

Remark 8.2. When X is a Ω-valued random variable with distribution μ and $f : \Omega \to \Lambda$ is a function, then $f(X)$ has distribution μf^{-1} on Λ.

Proof. Since

$$|\mu f^{-1}(B) - \nu f^{-1}(B)| = |\mu(f^{-1}(B)) - \nu(f^{-1}(B))|,$$

it follows that

$$\max_{B \subset \Lambda} |\mu f^{-1}(B)) - \nu f^{-1}(B)| \leq \max_{A \subset \Omega} |\mu(A) - \nu(A)|.$$

■

If α is a probability distribution on a finite subset Λ of $\mathbb{R}$, the translation of α by c is the probability distribution α_c on $\Lambda + c$ defined by $x \mapsto \alpha(x - c)$. Total variation distance is *translation invariant*: If α and β are two probability distributions on a finite subset Λ of $\mathbb{R}$, then $\|\alpha_c - \beta_c\|_{TV} = \|\alpha - \beta\|_{TV}$.

Proof of Proposition 8.5. Suppose that α and β are probability distributions on a finite subset Λ of $\mathbb{R}$. Let

$$m_\alpha := \sum_{x \in \Lambda} x\alpha(x), \quad m_\beta := \sum_{x \in \Lambda} x\beta(x)$$

be the mean of α and β, respectively; assume that $m_\alpha > m_\beta$. Define M by $m_\alpha - m_\beta = 2M$. By translating, we can assume that $m_\alpha = M$ and $m_\beta = -M$. Let $\eta = (\alpha + \beta)/2$, and define

$$f(x) := \frac{\alpha(x)}{\eta(x)}, \quad g(x) := \frac{\beta(x)}{\eta(x)}.$$

By Cauchy-Schwarz,

$$4M^2 = \left[\sum_{x \in \Lambda} x[f(x) - g(x)]\eta(x)\right] \leq \sum_{x \in \Lambda} x^2\eta(x) \sum_{x \in \Lambda} [f(x) - g(x)]^2\eta(x). \tag{8.18}$$ {Eq:CM1}

Note that

$$\sum_{x \in \lambda} x^2\eta(x) = \frac{m_\alpha^2 + \mathrm{Var}(\alpha) + m_\beta^2 + \mathrm{Var}(\beta)}{2} = M^2 + v, \tag{8.19}$$ {Eq:CM2}

where $v := (\mathrm{Var}(\alpha) + \mathrm{Var}(\beta))/2$. Since

$$|f(x) - g(x)| = 2\frac{|\alpha(x) - \beta(x)|}{\alpha(x) + \beta(x)} \leq 2,$$

we have

$$\sum_{x \in \Lambda} [f(x) - g(x)]^2\eta(x) \leq 4\frac{1}{2}\sum_{x \in \Lambda} |f(x) - g(x)|\eta(x) = 4\frac{1}{2}\sum_{x \in \Lambda} |\alpha(x) - \beta(x)|. \tag{8.20}$$ {Eq:CM3}

Putting together Equations (8.18) - (8.20) shows that

$$M^2 \leq (M^2 + \nu) \|\alpha - \beta\|_{\mathrm{TV}} ,$$

and rearranging shows that

$$\|\alpha - \beta\|_{\mathrm{TV}} \geq 1 - \frac{\nu}{\nu + M^2}.$$

If $2M \geq r\sqrt{\nu}$, then

{Eq:ForAB}

$$\|\alpha - \beta\|_{\mathrm{TV}} \geq 1 - \frac{4}{4 + r^2}. \tag{8.21}$$

If $\alpha = \mu f^{-1}, \beta = \nu f^{-1}$, and $\Lambda = f(\Omega)$, then $m_{\mu f^{-1}} = E_\mu(f)$, and (8.16) implies that $2M \geq r\sigma \geq r\sqrt{\nu}$. Using (8.21) in this case shows that

$$\left\|\mu f^{-1} - \nu f^{-1}\right\|_{\mathrm{TV}} \geq 1 - \frac{4}{4 + r^2}.$$

This together with Lemma 8.6 establishes (8.17). ■

8.3.1. Random walk on hypercube. We use Proposition 8.5 to bound below the mixing time for the random walk on the hypercube, studied in Section 7.4.2.

First we record a simple lemma concerning the coupon collector problem.

{Lem:MeanVarCC}

Lemma 8.7. *Consider the coupon collecting problem with n distinct coupon types (c.f. Section 4.2), and let $I_j(t)$ be the indicator of the event that the jth coupon has* not *been collected by time t. The random variables $I_j(t)$ are negatively correlated, and if $R_t = \sum_{j=1}^n I_j(t)$ is the number of coupon types not collected by time t, then*

{Eq:CCRemExp}

$$\mathbf{E}(R_t) = n\left(1 - \frac{1}{n}\right)^t, \tag{8.22}$$

{Eq:CCRemVar}

$$\mathrm{Var}(R_t) \leq \frac{n}{4}. \tag{8.23}$$

Proof. For $j \neq k$,

$$\mathbf{E}\left(I_j(t)\right) = \left(1 - \frac{1}{n}\right)^t$$
$$\mathbf{E}\left(I_j(t) I_k(t)\right) = \left(1 - \frac{2}{n}\right)^t.$$

Thus, for $j \neq k$,

$$\mathrm{Cov}(I_j(t), I_k(t)) = \left(1 - \frac{1}{n}\right)^{2t} - \left(1 - \frac{2}{n}\right)^t \leq 0.$$

From this Equation 8.22 and Equation 8.23 follow. ■

{Prop:LowerBoundRWHC}

Proposition 8.8. *For the lazy random walk on the n-dimensional hypercube,*

{Eq:LBHC}

$$d\left(\frac{1}{2}n \log n - \alpha n\right) \geq 1 - 8e^{-2\alpha+1} \tag{8.24}$$

Proof. Let $\mathbf{1}$ denote the vector of ones $(1, 1, \dots, 1)$, and let $W(\boldsymbol{x}) = \sum_{i=1}^{n} x^i$ be the Hamming weight of $\boldsymbol{x} = (x^1, \dots, x^n) \in \{0, 1\}^n$. We will apply Proposition 8.5 with $f = W$. The position of the walker at time t, started at $\mathbf{1}$, is denoted by $\boldsymbol{X}_t = (X_t^1, \dots, X_t^n)$.

As π is uniform on $\{0, 1\}^n$, the distribution of the random variable W under π is binomial with parameters n and $p = 1/2$. In particular,

$$E_\pi(W) = \frac{n}{2}, \quad \mathrm{Var}_\pi(W) = \frac{n}{4}.$$

Let R_t be the number of coordinates not update at least once by time t. When starting from $\mathbf{1}$, the conditional distribution of $W(\boldsymbol{X}_t)$ given $R_t = r$ is the same as $r + B$, where B is a binomial random variable with parameters $n - r$ and $1/2$. Consequently,

$$\mathbf{E}_{\mathbf{1}}(W(\boldsymbol{X}_t) \mid R_t) = R_t + \frac{(n - R_t)}{2} = \frac{1}{2}(R_t + n),$$

and using Equation 8.22,

$$\mathbf{E}_{\mathbf{1}}(W(\boldsymbol{X}_t)) = \frac{n}{2}\left[1 + \left(1 - \frac{1}{n}\right)^t\right].$$

Using the identity $\mathrm{Var}(W(\boldsymbol{X}_t)) = \mathrm{Var}(\mathbf{E}(W(\boldsymbol{X}_t) \mid R_t)) + \mathbf{E}(\mathrm{Var}(W(\boldsymbol{X}_t) \mid R_t))$,

$$\mathrm{Var}_{\mathbf{1}}(W(\boldsymbol{X}_t)) = \frac{1}{4}\mathrm{Var}(R_t) + \frac{1}{4}[n - \mathbf{E}_{\mathbf{1}}(R_t)].$$

By Lemma 8.7, R_t is the sum of negatively correlated indicators and consequently $\mathrm{Var}(R_t) \le \mathbf{E}(R_t)$. We conclude that

$$\mathrm{Var}_{\mathbf{1}}(W(\boldsymbol{X}_t)) \le \frac{n}{4}$$

Setting

$$\sigma = \sqrt{\max\{\mathrm{Var}_\pi(W), \mathrm{Var}(W(\boldsymbol{X}_t))\}} = \frac{\sqrt{n}}{2},$$

we have

$$\begin{aligned}
|E_\pi(W) - \mathbf{E}_{\mathbf{1}}(W(\boldsymbol{X}_t))| &= \frac{n}{2}\left(1 - \frac{1}{n}\right)^t \\
&= \sigma\sqrt{n}\left(1 - \frac{1}{n}\right)^t \\
&= \sigma \exp\left\{-t[-\log(1 - n^{-1})] + \frac{\log n}{2}\right\} \\
&\ge \sigma \exp\left\{-\frac{t}{n}\left(1 + \frac{1}{n}\right) + \frac{\log n}{2}\right\}.
\end{aligned}$$

The inequality follows since $\log(1 - x) \ge -x - x^2$ for $0 \le x \le 1/2$. By Proposition 8.5,

$$\left\|P^t(\mathbf{1}, \cdot) - \pi\right\|_{\mathrm{TV}} \ge 1 - 8\exp\left\{\frac{2t}{n}\left(1 + \frac{1}{n}\right) + \log n\right\}. \tag{8.25}$$ {Eq:HCDLowerBound}

The inequality (8.24) follows because

$$\frac{1}{2}n\log n - \alpha n \le t_n = \left[\frac{1}{2}n\log n - \left(\alpha - \frac{1}{2}\right)n\right]\left[1 - \frac{1}{n+1}\right],$$

and the right-hand side of (8.25) evaluated at $t = t_n$ equals $1 - 8e^{-2\alpha+1}$. ■

{Sec:TtoRLower}

8.4. Top-to-random shuffle

The top-to-random shuffle was introduced in Section 7.1.1, and upper bounds on $d(t)$ and t_{mix} were obtained in Section 7.4.3. Here we now obtain matching lower bounds.

The bound below, from Aldous and Diaconis (1986), uses only the definition of total variation distance.

{Prop:TtoRlower}

Proposition 8.9. *Let (X_t) be the top-to-random chain on n cards. For any $\varepsilon > 0$, there exists α_0 so that for $\alpha > \alpha_0$,*

{Eq:TtoRTV}

$$d_n(n\log n - \alpha n) \ge 1 - \varepsilon \quad \textit{for all } n \textit{ large enough.} \tag{8.26}$$

In particular, there is a constant α_1 so

{Eq:TtoRLowerMix}

$$t_{\mathrm{mix}} \ge n\log n - \alpha_1 n, \tag{8.27}$$

provided n is sufficiently large.

Proof. The bound is based on the events

$$A_j = \{\text{the original bottom } j \text{ cards are in their original relative order}\}. \tag{8.28}$$

Let id be the identity permutation; we will bound below $\left\|P^t(\mathrm{id},\cdot) - \pi\right\|_{TV}$.

Let τ_j be the time for the card initially jth from the bottom to reach the top. Then

$$\tau_j = \sum_{i=j}^{n-1} \tau_{j,i},$$

where $\tau_{j,i}$ is the time it takes the card initially jth from the bottom to ascend from position i (from the bottom) to position $i+1$. The variables $\{\tau_{j,i}\}_{i=j}^{n-1}$ are independent and $\tau_{j,i}$ has a geometric distribution with parameter $p = j/n$, whence $\mathbf{E}(\tau_{j,i}) = n/i$ and $\mathrm{Var}(\tau_{j,i}) < n^2/i^2$. We obain the bounds

{eq.ttre1}

$$\mathbf{E}(\tau_j) = \sum_{i=j}^{n-1} \frac{n}{j} \ge n(\log n - \log j - 1) \tag{8.29}$$

and

{eq.ttre2}

$$\mathrm{Var}(\tau_j) \le n^2 \sum_{i=j}^{\infty} \frac{1}{i(i-1)} \le \frac{n^2}{j-1}. \tag{8.30}$$

Let α_0 satisfy $\alpha_0 > \log j - 1$. Using the bounds (8.29) and (8.30), together with Chebyshev's inequality,

$$\begin{aligned} \mathbf{P}\{\tau_j < n \log n - \alpha_0 n\} &\le \mathbf{P}\{\tau_j - \mathbf{E}(\tau_j) < -n(\alpha_0 - \log j - 1)\} \\ &\le \frac{1}{(j-1)}. \end{aligned}$$

Define $t_n(\alpha) = n \log n - \alpha n$. If $\tau_j \ge t_n(\alpha)$, then the original j bottom cards remain in their original relative order at time $t_n(\alpha)$, so for $\alpha > \alpha_0$,

$$P^{t_n(\alpha)}(\mathrm{id}, A_j) \ge \mathbf{P}\{\tau_j \ge t_n(\alpha)\} \ge \mathbf{P}\{\tau_j \ge t_n(\alpha_0)\} \ge 1 - \frac{1}{(j-1)}.$$

On the other hand, for the uniform stationary distribution $\pi(A_j) = 1/(j!)$, whence for $\alpha > \alpha_0$,

$$d_n(\, t_n(\alpha)\,) \ge \left\| P^{t_n(\alpha)}(\mathrm{id}, \cdot) - \pi \right\|_{\mathrm{TV}} \ge P^{t_n(\alpha)}(\mathrm{id}, A_j) - \pi(A_j) > 1 - \frac{2}{j-1}. \tag{8.31}$$ {Eq:TtoR2}

Taking $j = e^{\alpha-2}$, provided $n \ge e^{\alpha-2}$, we have

$$d_n(\, t_n(\alpha))\,) > 1 - \frac{2}{e^{\alpha-2}}.$$

That is,

$$\liminf_{n\to\infty} d_n(\, t_n(\alpha)\,) \ge g_{\mathrm{lower}}(\alpha),$$

where $g_{\mathrm{lower}}(\alpha) \to 1$ as $\alpha \to \infty$. ■

8.5. The Cut-Off Phenomenon

{Sec:CO}

For the top-to-random shuffle on n cards, we obtained in Section 7.4.3 that

$$\limsup_{n\to\infty} d_n(n \log n + \alpha n) \le e^{-\alpha}, \tag{8.32}$$ {eq.t2raup}

while in Section 8.4, we showed that

$$\liminf_{n\to\infty} d_n(n \log n - \alpha n) \ge 1 - 2[e^{\alpha} - 2]^{-1}. \tag{8.33}$$ {eq.t2ralb}

In particular, the upper bound in (8.32) tends to 0 as $\alpha \to \infty$ and the lower bound in (8.33) tends to 1 as $\alpha \to \infty$. Thus, in the *window* $(n \log n - \alpha n, n \log n + \alpha n)$ centered at $n \log n$, the total variation distance drops from close to 1 to close to 0. Note that the window size is of order n, which grows slower than its center, $n \log n$.

If we rescale time by $n \log n$, so we consider the function $\delta_n(t) = d_n(tn \log n)$, then

$$\delta_n(t) \to \begin{cases} 0 & t < 1, \\ 1 & t > 1. \end{cases}$$

Thus, when viewed on the time-scale of $n \log n$, the total variation "falls of a cliff" at $n \log n$.

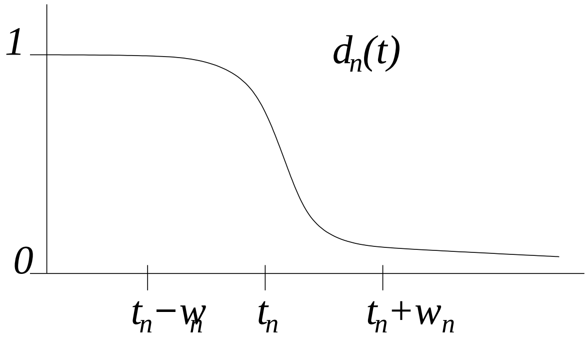


FIGURE 8.5. A graph of $d_n(t)$ against t; if the sequence of chains exhibits a cut-off, then then window where the distance drops from near 1 to near unity, is centered at t_n and shrinks (on the time scale of t_n). Fig:Cutoff

For each $n \in \{1, 2, \ldots\}$, let P_n be an irreducible and aperiodic transition matrix with stationary probability π_n on state-space Ω_n. We write $d_n(t)$ for $d(t)$ as defined in (5.22) to emphasize the dependence on n:

$$d_n(t) = \max_{x \in \Omega_n} \left\| P_n^t(x, \cdot) - \pi_n \right\|_{TV}.$$

We say that the corresponding sequence of Markov chains exhibits a *cut-off* at $\{t_n\}$ with *window* $\{w_n\}$ if $w_n = o(t_n)$, and

$$\lim_{\alpha \to \infty} \liminf_{n \to \infty} d_n(t_n - \alpha w_n) = 1,$$
$$\lim_{\alpha \to \infty} \limsup_{n \to \infty} d_n(t_n + \alpha w_n) = 0.$$

See Figure 8.5.

As a consequence, if a sequence of Markov chains has a cut-off, then there is a sequence t_n so that for some $c^\star$,

$$\lim_{n \to \infty} d_n(ct_n) = \begin{cases} 1 & c < c^\star, \\ 0 & c > c^\star. \end{cases}$$

EXAMPLE 8.10 (Random walk on the cycle). The random walk on $\mathbb{Z}_n$ does not exhibit a cut-off.

We first establish that

{Eq:RWCNCKey}
$$\min_{x \in \{0, 1, \ldots, n/8\}} \mathbf{P}_x \left\{ \tau_{\text{exit}} > n^2/100 \quad \text{and} \quad X_{n^2/100} \in \{-n/8, \ldots, -1, 0\} \right\} \geq \frac{1}{2}, \qquad (8.34)$$

where τ_{exit} is the first time the walker exits the interval $\{-n/4, \ldots, -1, 0, 1, \ldots, n/4\}$.

Iterating (8.34) shows that

$$\mathbf{P}_0\{\tau_{\text{exit}} > cn^2\} \geq \left(\frac{1}{2}\right)^{100c} = g(c).$$

Note that if $A = \{-n/2, -n/2+1, \ldots, n/2\}$, then $\pi(A) = \frac{1}{2} + o(1)$. Also,

$$\mathbf{P}\{X_t \in A \mid \tau < t\} = \frac{1}{2} + o(1),$$

by symmetry. Thus

$$\mathbf{P}\{X_{cn^2} \in A\} = [(1/2) + o(1)]\mathbf{P}\{\tau < cn^2\} \leq [(1/2) + o(1)][1 - g(c)],$$

and

$$\pi(A) - \mathbf{P}\{X_t \in A\} = (1/2)g(c) + o(1).$$

It is thus clear that for any $c > 0$,

$$\liminf_{n\to\infty} d(cn^2) \geq (1/2)g(c) > 0,$$

and there is no cut-off.

8.5.1. Random Walk on the Hypercube.

8.5.2. Cut-off for the hypercube. The Ehrenfest urn was defined in Section 4.3.2. Here we consider an upper bound on the mixing time via the *reflection coupling*. We consider the lazy version of the chain, which has transition probabilities, for $0 \leq j \leq n$,

$$P(j,k) = \begin{cases} \frac{1}{2} & k = j, \\ \frac{n-j}{2n} & k = j+1, \\ \frac{j}{2n} & k = j-1. \end{cases} \tag{8.35}$$

{Eq:LazyEhren}

As remarked in Section 4.3.2, if $(\boldsymbol{X}_t)$ is the lazy random walk on the n-dimensional hypercube $\{0,1\}^n$, then the Hamming weight chain (W_t),

$$W_t = W(\boldsymbol{X}_t) = \sum_{i=1}^{n} X_t^i,$$

has the transition probabilities in (8.35).

It will be convenient to analyze the centered process (Z_t), defined as $Z_t = W_t - n/2$ and with values in

$$\mathcal{Z} = \{-n/2, -n/2+1, \ldots, n/2-1, n/2\}.$$

The chain (Z_t) has transition matrix

$$Q(z,z') = \begin{cases} \frac{1}{2} & z' = z, \\ \frac{n/2-z}{2n} & z' = z+1, \\ \frac{n/2+z}{2n} & z' = z-1, \end{cases} \tag{8.36}$$

{Eq:ZMatrix}

If $\pi_Q(z) = \binom{n}{z+n/2}2^{-n}$, then it is easy to check that π_Q is the stationary distribution for Q and that

$$\|\mathbf{P}_w\{W_t \in \cdot\} - \pi\|_{TV} = \left\|\mathbf{P}_{w-n/2}\{Z_t \in \cdot\} - \pi_Q\right\|_{TV}. \tag{8.37}$$

{Eq:WZ}

Thus it will suffice to analyze the distance on the right-hand side of (8.37).

Define

$$\tau_0 = \min\{t \geq 0 \,:\, |Z_t| \leq 1/2\}. \tag{8.38}$$

{Eq:Tau0Defn}

Note that if n is even, then $Z_{\tau_0} = 0$, while for n odd, $Z_{\tau_0} = \pm 1/2$.

{Lem:EZ}

LEMMA 8.11. *Let (Z_t) be a Markov chain with transition matrix (8.36), and let τ_0 be the random time in (8.38). Then for $z \in \mathcal{Z}$*

{Eq:EZt}

$$\mathbf{E}_z(Z_t) = z(1 - n^{-1})^t, \tag{8.39}$$

and if $z \geq 0$, then

{Eq:EZtau}

$$\mathbf{E}_z(Z_t \mathbf{1}_{\{\tau_0 > t\}}) \leq z(1 - n^{-1})^t. \tag{8.40}$$

PROOF. Note that

{Eq:ZDrift}

$$\begin{aligned}\mathbf{E}(Z_{t+1} \mid Z_t = z) &= (z+1)\left[\frac{1}{4} - \frac{k}{2n}\right] + z\frac{1}{2} + (z-1)\left[\frac{1}{4} + \frac{k}{2n}\right] \\ &= z\left(1 - \frac{1}{n}\right).\end{aligned} \tag{8.41}$$

If $M_t = Z_t/(1 - n^{-1})^t$, then it follows from (8.41) that

{Eq:MMart}

$$\mathbf{E}(M_{t+1} \mid M_0, \ldots, M_t) = M_t. \tag{8.42}$$

Taking expectations and iterating shows that

{Eq:MConstE}

$$\mathbf{E}_z(M_t) = \mathbf{E}_z(M_0) = z, \tag{8.43}$$

which establishes (8.39).

In fact, (8.43) remains true if we replace t by the random time $\tau_0 \wedge t$, which we now show. (This is a special case of the Optional Stopping Theorem, which we prove in more generality in Chapter 19 – cf. Theorem 19.6.) We write

{Eq:MTele}

$$M_{t \wedge \tau_0} - M_0 = \sum_{s=1}^{t} (M_s - M_{s-1}) \mathbf{1}_{\{\tau_0 > s-1\}}. \tag{8.44}$$

Equation 8.42, together with the fact that the random variable $\mathbf{1}\{\tau_0 > s - 1\}$ is a function of $M_0, M_1, \ldots, M_{s-1}$, shows that

{Eq:MIncZero}

$$\begin{aligned}&\mathbf{E}_z((M_s - M_{s-1})\mathbf{1}_{\{\tau_0 > s-1\}} \mid M_0, M_1, \ldots, M_{s-1}) \\ &\qquad = \mathbf{1}_{\{\tau_0 > s-1\}} \mathbf{E}_z(M_s - M_{s-1} \mid M_0, \ldots, M_{s-1}) = 0.\end{aligned} \tag{8.45}$$

Using (8.45) in (8.44) yields the identity

$$\mathbf{E}_z(M_{t \wedge \tau_0}) = \mathbf{E}_z(M_0).$$

Since, $M_{\tau_0} \in \{0, 1/2\}$ when $z > 0$, and $\mathbf{E}_z(M_0) = z$,

$$z = \mathbf{E}_z(M_{t \wedge \tau_0}) \geq \mathbf{E}_z(M_t \mathbf{1}_{\{\tau_0 > t\}}) = \mathbf{E}_z(Z_t \mathbf{1}_{\{\tau_0 > t\}}) \left[1 - n^{-1}\right]^{-t}.$$

■

{Lem:RPforZ}

LEMMA 8.12. *For the Markov chain (Z_t), there is a constant C so that for $z \in \mathcal{Z} \cap [0, \infty)$,*

$$\mathbf{P}_z\{\tau_0 > t\} \leq \frac{Cz}{\sqrt{t}}.$$

PROOF. We will define, on the same probability space as (Z_t) and until time τ_0, a nearest-neighbor unbiased random walk (S_t), with values in $\mathcal{Z}$ and initial value $S_0 = z$, as follows: First, a fair coin is tossed; if heads, both chains move, and if tails, neither chain moves. In the case where the coin lands heads, a uniform random variable U is generated. The chains move based on U according to the following table:

U	$Z_{t+1} - Z_t$	$S_{t+1} - S_t$
$0 \leq U < \frac{1}{2}$	-1	-1
$\frac{1}{2} \leq U < \frac{1}{2} + \frac{k}{n}$	-1	$+1$
$\frac{1}{2} + \frac{k}{n} \leq U < 1$	$+1$	$+1$

Note that always, provided $\tau_0 > t$,

$$Z_{t+1} - Z_t \leq S_{t+1} - S_t,$$

so that in particular, $Z_t \leq S_t$ for $t < \tau_0$. Consequently,

$$\mathbf{P}_z\{Z_1 > 0, \ldots, Z_t > 0\} \leq \mathbf{P}_z\{S_1 > 0, \ldots, S_t > 0\}.$$

The conclusion follows then by Corollary 4.17. ■

{Lem:ZHitTime}

LEMMA 8.13. *Let (Z_t) be the Markov chain with the transition probabilities (8.36), and let τ_0 be the time in (8.38). If $z > 0$, then*

$$\mathbf{P}_z\{\tau_0 > (1/2)n \log n + \alpha n\} \leq \frac{C}{\sqrt{\alpha}}.$$

PROOF. Let $s = (1/2)n \log n$. Then by Lemma 8.12, on the event $\{\tau_0 > s\}$,

$$\mathbf{P}_z\{Z_{s+1} > 0, \ldots, Z_{s+t} > 0 \mid Z_0, \ldots, Z_s\} \leq \frac{CZ_s}{\sqrt{t}}.$$

Since $\mathbf{1}_{\{\tau_0 > s\}}$ is a function of $Z_0, Z_1, \ldots, Z_s$,

$$\mathbf{P}_z\{\tau_0 > s, Z_{s+1} > 0, \ldots, Z_{s+t} > 0 \mid Z_0, \ldots, Z_s\} \leq \frac{CZ_s \mathbf{1}_{\{\tau_0 > s\}}}{\sqrt{t}}.$$

Taking expectation and using Lemma 8.11 shows that

$$\mathbf{P}_z\{\tau_0 > s + t\} \leq \frac{C\sqrt{n}}{\sqrt{t}}.$$

■

{Prop:ZMix}

PROPOSITION 8.14. *For any $z, u \in \mathcal{Z}$, there is a coupling of two chains, each with the transition matrix defined in (8.36), one started from z and the other started from u, so that the time τ when the chains first meet satisfies*

$$\mathbf{P}_{z,u}\{\tau > (1/2)n \log n + \alpha n\} = O(\alpha^{-1/2}).$$

PROOF. We assume, without loss of generality, that $|z| \geq |u|$.

Let (Z_t) be any chain started at z with transitions (8.36). We show how to define a chain (U_t), using (Z_t) and some additional randomness.

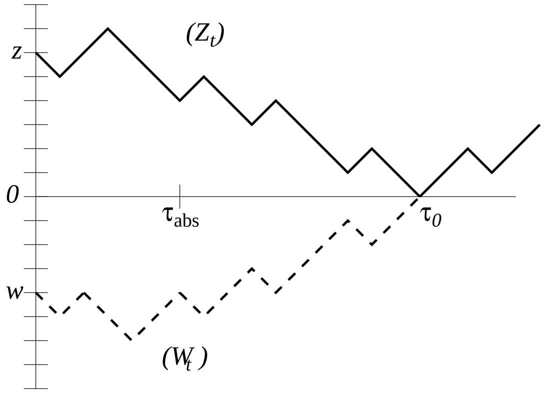


FIGURE 8.6. Run (U_t) independently of (Z_t) until the time τ_{abs} when their absolute values first agree. After this time, if the chains do not agree, run (U_t) as a reflection of (Z_t) about the t-axis. Fig:RefCoup

First, run (U_t) and (Z_t) as follows: First toss a fair coin to decide which of the two chains to move. For the chosen chain, make a "non-lazy" move using the transition probabilities:

$$P(j,k) = \begin{cases} \frac{n-k}{n} & k = j+1, \\ \frac{j}{2n} & k = j-1. \end{cases}$$

Continue this way until the time

$$\tau_{\text{abs}} = \min\{t \geq 0 \,:\, |U_t| = |Z_t|\}.$$

If $U_{\tau_{\text{abs}}} = Z_{\tau_{\text{abs}}}$, then let $U_t = Z_t$ for all $t > \tau_{\text{abs}}$.

The time τ_0 is as defined in (8.38). Since $|z| \geq |u|$, we must have $\tau_{\text{abs}} \leq \tau_0$.

If $U_{\tau_{\text{abs}}} = -Z_{\tau_{\text{abs}}}$, then for $\tau_{\text{abs}} \leq t < \tau_0$ set $U_{t+1} - U_t = -(Z_{t+1} - Z_t)$. In this case, (U_t) is a reflection of (Z_t) for $\tau_{\text{abs}} \leq t \leq \tau_0$. (See Figure 8.6.)

Case 1: n even. In this case, $Z_{\tau_0} = U_{\tau_0} = 0$. Thus, the coupling time for the two chains is simply τ_0, and the conclusion of the Proposition follows from Lemma 8.13.

Case 2: n odd. If $U_{\tau_{\text{abs}}} = Z_{\tau_{\text{abs}}}$, the chains will have coupled already by τ_0, and the Proposition again follows from Lemma 8.13.

Otherwise, suppose without loss of generality that $Z_{\tau_0} = 1/2 = -U_{\tau_0}$. Toss a fair coin; if heads move the Z-chain, and if tails move the U-chain. If the two chains do not agree, start the coupling described in this proof anew with the current states as the new starting states, wait again until the chains are at ±1/2, and again flip a coin to decide which chain to move.

■

{Thm:EhrenMix}

THEOREM 8.15. *Let* (W_t) *be the Ehrenfest chain with transition probabilities* (8.35). *Then*

$$d((1/2)n \log n + \alpha n) = O(\alpha^{-1/2}),$$

and so

$$t_{\mathrm{mix}} = [1 + o(1)](1/2)n \log n.$$

PROOF. The proof follows from Proposition 8.14, Corollary 6.3, and Equation 8.37. ■

We return now to the lazy random walk on the n-dimensional hypercube, $(\boldsymbol{X}_t)$. Conditional on $W_t = W(\boldsymbol{X}_t) = w$, the distribution of $\boldsymbol{X}_t$ is uniform over all states $\boldsymbol{x}$ with $W(\boldsymbol{x}) = w$. Using this fact, the reader should check that

$$\|\mathbf{P}_{\boldsymbol{x}}\{\boldsymbol{X}_t \in \cdot\} - \pi\|_{TV} = \left\|\mathbf{P}_{W(\boldsymbol{x})}\{W_t \in \cdot\} - \pi_W\right\|_{TV},$$

where

$$\pi_W(w) = \sum_{\substack{\boldsymbol{x} \in \{0,1\}^n \\ W(\boldsymbol{x}) = w}} \pi(\boldsymbol{x}).$$

Using this identity, Theorem 8.15 yields the following:

{Thm:RWHCMix}

THEOREM 8.16. *Let $(\boldsymbol{X}_t)$ be the lazy simple random walk on the n-dimensional hypercube. For this chain,*

$$d((1/2)n \log n + \alpha n) = O(\alpha^{-1/2}).$$

Consider again the lazy random walk on $\{0, 1\}^d$: at each move, a coordinate is selected at random and replaced by an independent random bit.

If $X(t) = (X_1(t), \ldots, X_d(t))$, let $Y_t := \sum_{i=1}^d X_i(t) - d/2$. As before, we can calculate that

$$\mathbf{E}_{\mathbf{1}}(Y_t) = \frac{d}{2}\left(1 - \frac{1}{d}\right)^t,$$

where $\mathbf{1} = (1, \ldots, 1)$.

Letting $t_0 = (1/2)d \log d$,

$$\mathbf{E}_{\mathbf{1}}(Y_{t_0}) \leq \frac{de^{-t_0/d}}{2} = \frac{\sqrt{d}}{2}.$$

{Lem:RP3}

LEMMA 8.17. *Let $(S_t)_{t=0}^{\infty}$ be a simple random walk.*

$$\mathbf{P}_k\{\tau_0 > t\} \leq \frac{ch}{\sqrt{t}}. \tag{8.46}$$

{Eq:RP}

Thus,

$$\mathbf{P}_{\mathbf{1}}\{Y_{t_0+j} \geq 0,\ 1 \leq j \leq r \mid Y_{t_0} = h\} \leq \frac{2ch}{\sqrt{r}}.$$

8.6. East Model

Let

$$\Omega := \{x \in \{0, 1\}^{n+1} \ : \ x(n + 1) = 1\}.$$

The *East model* is the Markov chain on Ω which moves from x by selecting a coordinate k from $\{1, 2, \ldots, n\}$ at random and flipping the value $x(k)$ at k if and only if $x(k + 1) = 1$. The reader should check that the uniform measure on Ω is stationary for these dynamics.

Theorem 8.18. *For the East model, $t_{\mathrm{mix}} \geq cn^2$.*

Proof. If $A = \{x : x(1) = 1\}$, then $\pi(A) = 1/2$.

On the other hand, we now show that it takes order n^2 steps until $X_t(0) = 1$ with probability near $1/2$ when starting from $x_0 = (0, 0, \ldots, 0, 1)$. Consider the motion of the left-most 1: it moves to the left by one if and only if the site immediately to its left is chosen. Thus, the waiting time for the left-most 1 to move from k to $k-1$ is bounded by a geometric random variable G_k with mean n. The sum $G = \sum_{k=1}^{n} G_k$ has mean n^2 and variance $(1 - n^{-1})n^3$. Thus if $t(n, c) = n^2 - cn^{3/2}$, then

$$\mathbf{P}\{X_{t(n,c)}(0) = 1\} \leq \mathbf{P}\{G - n^2 \leq -cn^{3/2}\} \leq \frac{1}{c^2},$$

and so

$$|P^{t(n,c)}(x_0, A) - \pi| \geq \frac{1}{2} - \frac{1}{c^2}.$$

Thus, if $t \leq n^2 - 2n^{3/2}$, then $d(t) \geq 1/4$. In other words, $t_{\mathrm{mix}} \geq n^2 - 2n^{3/2}$. ■

8.7. Problems

{Exercise:NegCor}

Exercise 8.1. Let $\boldsymbol{X}_t = (X_t^1, \ldots, X_t^n)$ be the position of the lazy random walker on the hypercube $\{0, 1\}^n$, started at $\boldsymbol{X}_0 = \mathbf{1} = (1, \ldots, 1)$. Show that the covariance between X_t^i and X_t^j is negative. Conclude that if $W(\boldsymbol{X}_t) = \sum_{i=1}^{n} X_t^i$, then $\mathrm{Var}(W(X_t)) \leq n/4$.

Hint: It may be easier to consider the variables $Y_t^i = 2X_t^i - 1$.

{Exercise:QSym}

Exercise 8.2. Show that $Q(S, S^c) = Q(S^c, S)$ for any $S \subset \Omega$. (This is easy in the reversible case, but holds generally.)

{Exercise:Diameter}

Exercise 8.3. Suppose that (X_t) is a random walk on a graph with vertex set Ω and let $\Delta = \max_{x \in \Omega} \deg(x)$. Show that for some constant c,

$$t_{\mathrm{mix}} \geq c \frac{\log(|\Omega|)}{\log(\Delta)}.$$

{Exercise:EmptyGraph}

Exercise 8.4. An *empty graph* has no edges. A proper coloring of an empty graph with vertex set V is an element of $\Omega = \{1, \ldots, q\}^V$. Each element $x \in \Omega$ can be thought of as an assignment of a *color* (an element of $\{1, 2, \ldots, q\}$) to each vertex $v \in V$. The Glauber dynamics for the uniform measure on Ω is the chain which moves by selecting at each move a vertex v from V uniformly at random, and changing the color at v to a uniform random element of $\{1, 2, \ldots, q\}$.

Show that there is a constant $c(q)$ so that

$$t_{\mathrm{mix}} \geq \frac{1}{2} n \log n - c(q)n.$$

Hint: Copy the idea of the proof of Proposition 8.8.

8.8. Notes

It is more common to relate the bottleneck ratio $\Phi_\star$ to the *spectral gap* of a Markov chain. See Chapter 12 for some of the history of this relation. The approach to the lower bound for t_{mix} presented here is more direct and avoids reversibility. Results related to Theorem 8.1 can be found in Mihail (1989), Fill (1991), and Chen, Lovász, and Pak (1998).

Hayes and Sinclair (2005) have recently shown that the Glauber dynamics for many stationary distributions, on graphs of bounded degree, have mixing time order $n \log n$.

Upper bounds on the relaxation time (see Section 12.4) for the East model are obtained in Aldous and Diaconis (2002).

CHAPTER 9

Shuffling Cards

{Ch:Shuffling}

Card shuffling is such an important example for the theory of Markov chains that we have not been able to avoid it in earlier chapters. Here we study several other natural methods of shuffling cards.

A stack of n cards can be viewed as an element of the symmetric group $\mathcal{S}_n$. A shuffling mechanism can then be specified by a probability distribution Q on $\mathcal{S}_n$. At each step, a permutation is chosen according to Q and applied to the deck. The resulting Markov chain has transition matrix

$$P(\rho_1, \rho_2) = Q(\rho_2 \rho_1^{-1}) \text{ for } \rho_1, \rho_2 \in \mathcal{S}_n.$$

As long as the support of Q generates all of $\mathcal{S}_n$, the resulting chain is irreducible. If $Q(\mathrm{id}) > 0$, then it is aperiodic. Every shuffle chain is transitive, and hence (by Exercise 7.5) has uniform stationary distribution.

A warning to the reader: in this chapter, the stationary distributions of all chains under consideration are uniform, and we often write U for the uniform distribution.

9.1. Random transpositions

Pick two cards at random; switch their locations in the deck. Repeat. It's difficult to imagine a simpler shuffle. How many shuffles are necessary before the deck has been well-randomized?

Let's be more precise about the mechanism. At each step, the shuffler chooses two cards, independently and uniformly at random. If the same card is chosen twice, nothing is done to the deck. Otherwise, the positions of the two chosen cards are switched. The possible moves have weights

$$Q(\sigma) = \begin{cases} 1/n & \rho = \mathrm{id}, \\ 2/n^2 & \rho = (ij), \\ 0 & \text{otherwise}. \end{cases} \tag{9.1}$$

{Eq:RandTransDist}

In Section 2.4, we gave a method for generating a uniform random permutation that started with the set $[n]$ sorted and used only transpositions. Thus the set of transpositions generates S_n, and the underlying Markov chain is therefore irreducible. Since $Q(id) > 0$, it is aperiodic as well.

In each round of random transposition shuffling, (almost) two cards are selected, and each is moved to an (almost) random location. In other examples, such as the hypercube, we have been able to bound convergence by tracking how many features have been randomized. If—if!—a similar analysis applies to the random

Aligning one card:

$$\begin{array}{cccc} 2 & 4 & 1 & 3 \\ 3 & 1 & 4 & 2 \end{array} \Longrightarrow \begin{array}{cccc} 1 & 4 & 2 & 3 \\ 1 & 3 & 4 & 2 \end{array}$$

Aligning two cards:

$$\begin{array}{cccc} 2 & 3 & 1 & 4 \\ 3 & 1 & 4 & 2 \end{array} \Longrightarrow \begin{array}{cccc} 1 & 3 & 2 & 4 \\ 1 & 3 & 4 & 2 \end{array}$$

Aligning three cards:

$$\begin{array}{ccc} 2 & 3 & 1 \\ 3 & 1 & 2 \end{array} \Longrightarrow \begin{array}{ccc} 1 & 3 & 2 \\ 1 & 3 & 2 \end{array}$$

Fig:RandTransCouple

FIGURE 9.1. Aligning cards using coupled random transpositions. In each example, $X_t = 1$ and $Y_t = 1$, so card 1 is transposed with the card in position 1 in both decks.

transposition shuffle, we might hope that, since each step moves (almost) two cards, half the coupon collector time of approximately $n \log n$ steps will suffice to bring the distribution close to uniform.

In fact, as Diaconis and Shahshahani (1981) proved, the random transpositions walk has a sharp cutoff of width $O(n)$ at $(1/2)n \log n$. They use Fourier analysis on the symmetric group to achieve these extremely precise results. Here, we present two upper bounds on the mixing time: a simple coupling that gives an upper bound of order n^2 for the mixing time, and a strong stationary time argument due to Broder (see Diaconis (1988)) that gives an upper bound within a constant factor of the correct answer. While the lower bound we give does not quite reach the cutoff, it does have the correct lead term constant.

9.1.1. Upper bound via coupling. For the coupling, we take a slightly different view of generating the transpositions. At each time t, the shuffler chooses a card $X_t \in [n]$ and, independently, a position $Y_t \in [n]$; she then transposes the card X_t with the card in position Y_t. Of course, if X_t already occupies Y_t, the deck is left unchanged. Hence this mechanism generates the measure described in (9.1).

To couple two decks, use the same choices (X_t) and (Y_t) to shuffle both. Let (σ_t) and (σ_t') be the two trajectories. What can happen in one step? Let a_t be the number of cards that occupy the same position in both σ_t and σ_t'.

- If X_t is in the same position in both decks, and the same card occupies position Y_t in both decks, then $a_{t+1} = a_t$.
- If X_t is in different positions in the two decks, but position Y_t is occupied by the same card, then performing the specified transposition breaks one alignment, but also forms a new one. We have $a_{t+1} = a_t$.
- If X_t is in different positions in the two decks, and if the cards at position Y_t in the two decks do not match, then at least one new alignment is made—and possibly as many as three. See Figure 9.1.

{Prop:RandTransCouple}

Proposition 9.1. *Let τ be the time required for the two decks to couple. Then, no matter the initial configurations of the two decks,* $\mathbf{E}(\tau) < \frac{\pi^2}{6}n^2$.

Proof. Decompose

$$\tau = \tau_1 + \cdots + \tau_n,$$

where τ_i is the number of transpositions between the first time that a_t is greater than or equal to $i-1$ and the first time that is a_t is greater than or equal to i. (Since a_0 can be greater than 0, and since a_t can increase by more than 1 in a single transposition, it is possible that many of the τ_i's are equal to 0.)

When t satisfies $a_t = i$, there are $n - i$ unaligned cards and the probability of increasing the number of alignments is $(n-i)^2/n^2$, since the shuffler must choose a non-aligned card and a non-aligned position. In this situation τ_{i+1} is geometric($(n-i)^2/n^2$). We may conclude that under these circumstances

$$\mathbf{E}(\tau_{i+1} | a_t = i) = n^2/(n-i)^2.$$

When no value of t satisfies $a_t = i$, then $\tau_{i+1} = 0$. Hence

$$\mathbf{E}(\tau) < n^2 \sum_{i=1}^{n} \frac{1}{(n-i)^2} < n^2 \sum_{l=1}^{\infty} \frac{1}{l^2}.$$

■

Markov's inequality and Corollary 6.3 now give an $O(n^2)$ bound on t_{mix}. However, the strong stationary time we are about to discuss does much better.

{Sec:RandTransSST}

9.1.2. Upper bound via strong stationary time.

{Prop:RandTransSST}

Proposition 9.2. *In the random transposition shuffle, let R_t and L_t be the cards chosen by the right and left hands, respectively, at time t. Assume that when $t = 0$, no cards have been marked. At time t, mark card R_t if both of the following are true:*

- *R_t is unmarked.*
- *Either L_t is a marked card, or $L_t = R_t$.*

Let τ be the time when every card has been marked. Then τ is a strong uniform time for this chain.

Here's a heuristic explanation for why the scheme described above should give a strong stationary time. One way to generate a uniform random permutation is to build a stack of cards, one at a time, inserting each card into a uniformly random position relative to the cards already in the stack. For the stopping time described above, the marked cards are carrying out such a process.

Proof. It's clear that τ is a stopping time. To show that it is a strong uniform time, we prove the following subclaim by induction on t. Let $V_t \subseteq [n]$ be the set of cards marked at or before time t, and let $U_t \subseteq [n]$ be the set of positions occupied by V_t after the t-th transposition. We claim that *given t, V_t, and U_t, all possible permutations of the cards in V_t on the positions U_t are equally likely.*

This is clearly true when $t = 1$ (and continues to clearly be true as long as at most one card has been marked).

Now, assume that the subclaim is true for t. The shuffler chooses cards L_{t+1} and R_{t+1}.

- If no new card is marked, then $V_{t+1} = V_t$. This can happen two ways:
  - If L_{t+1} and R_{t+1} were both marked at an earlier round, then $U_{t+1} = U_t$ and the shuffler applies a uniform random transposition to the cards in V_t. All permutations of V_t remain equiprobable.
  - Otherwise, L_{t+1} is unmarked and R_{t+1} was marked at an earlier round. To obtain the position set U_{t+1}, we delete the position (at time t) of R_{t+1} and add the position (at time t) of L_{t+1}. For a fixed set U_t, all choices of $R_{t+1} \in U_t$ are equally likely, as are all permutations of V_t on U_t. Hence, once the positions added and deleted are specified, all permutations of V_t on U_{t+1} are equally likely.
- If the card R_{t+1} gets marked, then L_{t+1} is equally likely to be any element of $V_{t+1} = V_t \cup \{R_{t+1}\}$, while U_{t+1} consists of U_t along with the position of L_{t+1} (at time t). Specifying the permutation of V_t on U_t and the card L_{t+1} uniquely determines the permutation of V_{t+1} on U_{t+1}. Hence all such permutations are equally likely.

In every case, the collection of all permutations of the cards V_t on a specified set U_t together make equal contributions to all possible permutations of V_{t+1} on U_{t+1}. Hence, to conclude that all possible permutations of a fixed V_{t+1} on a fixed U_{t+1} are equally likely, we simply sum over all possible preceding configurations. ■

Remark. In the preceding proof, the two subcases of the inductive step for which no new card is marked are essentially the same as checking that the uniform distribution is stationary for the random transposition shuffle and the random-to-top shuffle, respectively.

Remark. As Diaconis (1988) points out, for random transpositions some simple card-marking rules fail to give strong uniform times. See Exercise 9.5.

{Lem:RandTransSSTEst}

Lemma 9.3. *The stopping time τ defined in Proposition 9.2 satisfies*

$$\mathbf{E}(\tau) = 2n(\log n + O(1))$$

and

$$Var(\tau) = O(n^2).$$

Proof. As for the coupon collector time, we can decompose

$$\tau = \tau_0 + \cdots + \tau_{n-1},$$

where τ_k is the number of transpositions after the k-th card is marked, up to and including when the $(k+1)$-st card is marked. The rules specified in Proposition 9.2 imply that τ_k is geometric$\left(\frac{(k+1)(n-k)}{n^2}\right)$ and that the τ_i's are independent of each other. Hence

$$\mathbf{E}(\tau) = \sum_{k=0}^{n-1} \frac{n^2}{(k+1)(n-k)}.$$

Substituting the partial fraction decomposition

$$\frac{1}{(k+1)(n-k)} = \frac{1}{n+1}\left(\frac{1}{k+1} + \frac{1}{n-k}\right)$$

and recalling that

$$\sum_{j=1}^{n} \frac{1}{j} = \log n + O(1)$$

(see Exercise 4.5) completes the estimate.

Now, for the variance. We can immediately write

$$\operatorname{Var}(\tau) = \sum_{k=0}^{n-1} \frac{1 - \frac{(k+1)(n-k)}{n^2}}{\left(\frac{(k+1)(n-k)}{n^2}\right)^2} < \sum_{k=0}^{n-1} \frac{n^4}{(k+1)^2(n-k)^2}.$$

Split the sum into two pieces:

$$\begin{aligned}\operatorname{Var}(\tau) &< \sum_{0\le k<n/2} \frac{n^4}{(k+1)^2(n-k)^2} + \sum_{n/2\le k<n} \frac{n^4}{(k+1)^2(n-k)^2}\\ &< \frac{2n^4}{(n/2)^2} \sum_{0\le k\le n/2} \frac{1}{(k+1)^2} = O(n^2).\end{aligned}$$

■

{Thm:RandTrans}

Corollary 9.4. *For the random transposition chain on an n-card deck,*

$$t_{\mathrm{mix}} \le (2 + o(1))n \log n.$$

Proof. Let τ be the Broder stopping time defined in Proposition 9.2, and let $t_0 = E(\tau) + 2\sqrt{\operatorname{Var}(\tau)}$. By Chebyshev's inequality,

$$\mathbf{P}(\tau > t_0) \le \frac{1}{4}.$$

Lemma 9.3 and Proposition 7.3 now imply the desired inequality. ■

9.1.3. Lower bound.

Proposition 9.5. *Let $0 < \varepsilon < 1$. For the random transposition chain on an n-card deck,*

$$t_{\mathrm{mix}}(\varepsilon) \ge \frac{1}{2}\left(n \log n - \log\left(\frac{12}{1-\varepsilon}\right)n\right)$$

for sufficiently large n.

Proof. It is well-known (and easily proved using indicators) that the expected number of fixed points in a uniform random permutation in $\mathcal{S}_n$ is 1, regardless of the value of n.

Now let $t_n = \frac{1}{2}\left(n \log n - \log\left(\frac{12}{1-\varepsilon}\right)n\right)$, and choose σ according to $P^{t_n}(\mathrm{id}, \cdot)$. The number of fixed points of σ is at least as large as the number of cards left untouched in $2t_n$ independent uniform selections from the deck, which has a coupon collector distribution. By Lemma 8.7, the number of untouched cards has expected value

$$\mu_n = n\left(1 - \frac{1}{n}\right)^{2t_n}$$

and variance bounded by $n/4$.

Let A be the event that there are at least $\mu_n/2$ fixed points in the permutation. Let's estimate the probability of A under the two measures. First of all,

$$\mathbf{P}_U(A) \leq \frac{2}{\mu_n},$$

by Markov's inequality. On the other hand, $P^t(\mathrm{id}, A)$ is at least as large as the probability that there are more than $\mu_n/2$ cards left untouched by the first t shuffles. By Chebyshev's inequality and Lemma 8.7,

$$P^{t_n}(\mathrm{id}, A) \geq 1 - \frac{n/4}{(\mu_n/2)^2} \geq 1 - \frac{4}{\mu_n},$$

since $\mu_n > n/4$ for sufficiently large n. By the definition (5.1) of total variation distance, we have

$$\left\| P_n^t(\mathrm{id}, \cdot) - U \right\|_{\mathrm{TV}} \geq 1 - \frac{6}{\mu_n}.$$

Recall that for $0 \leq x \leq 1/2$, it's true that $\log(1-x) > -x - x^2$. It follows that for $n \geq 2$,

$$\mu_n \geq n\left(e^{-\frac{1}{n} - \frac{1}{n^2}}\right)^{n(\log n - \log \frac{12}{1-\varepsilon})} = n\left(e^{-1-\frac{1}{n}}\right)^{\log n - \log \frac{12}{1-\varepsilon}} = \frac{12}{1-\varepsilon}(1 + o(1))$$

as $n \to \infty$. In particular, for sufficiently large n, we have $\mu_n > 6/(1-\varepsilon)$ and hence $\left\| P_n^t(\mathrm{id}, \cdot) - U \right\|_{\mathrm{TV}} > \varepsilon$. ■

9.2. Random adjacent transpositions

A reasonable restriction of the random transposition shuffle to consider is to only interchange adjacent cards—see Figure 9.2. Restricting the moves in this manner will slow the shuffle down. We present a coupling (described in Aldous (1983) and also discussed in Wilson (2004)) that gives a sharp upper bound of order $n^3 \log n$, and then give a lower bound of order n^3.

Note: this shuffle is such a useful example that we discuss it in two other places. In Section 12.7 we use Wilson's method to obtain a lower bound that matches our upper bound, up to constants. In addition, in Section 13.4.2 we use Theorem 13.5 to compare the convergence to stationarity of random adjacent transpositions to that of random transpositions.

9.2.1. Upper bound via coupling. We consider a lazy version of this shuffle: at each step, with probability 1/2 do nothing, and with probability 1/2 choose uniformly between the $(n-1)$ transpositions of adjacent pairs of cards.

In order to couple two copies (σ_t) and (σ'_t) (the "left" and "right" decks) of this lazy version, 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 τ_a, the time required for a particular card a to couple. Let X_t be the (unsigned) distance between the positions of i 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 $\{0, \ldots, n-1\}$ 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 $0 \leq i \leq n-1$,

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

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

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

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 τ_a is bounded by the time τ_0^Y it takes (Y_t) to absorb at 0.

The chain (Y_t) is best viewed as a delayed version of a random walk on the path $\{0, \ldots, n-1\}$, with a hold probability of $1/2$ at $n-1$ and absorption at 0. 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 4.3 and 4.2 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

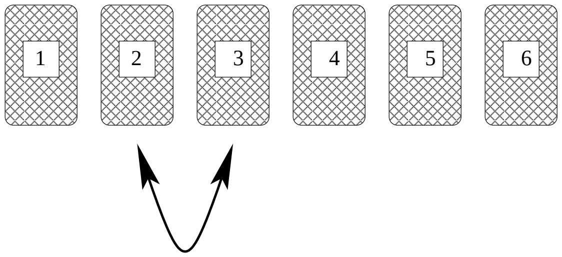


FIGURE 9.2. An adjacent transposition swaps two neighboring cards. Fig:ART

cards having not coupled, so

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

regardless of the initial states of the decks. Theorem 6.2 immediately implies that $t_{\text{mix}} < 2n^3 \log_2 n$ for sufficiently large n.

{Sec:RATransLower}

9.2.2. Lower bound for random adjacent transpositions.

Consider the set of permutations

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

Under the uniform measure we have $U(A) = \lfloor n/2 \rfloor / n \geq 1/2$, 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 doesn't get to move very often, it's 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 moves under the action of the random adjacent transposition shuffle? Each interior card (neither top nor bottom of the deck) moves with probability $2/(n-1)$, 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 $1/(n-1)$, and always moves in the one permitted direction. This means that

{Eq:DomLRE}

$$\mathbf{P}(\text{card 1 has visited position } \lfloor n/2 \rfloor \text{ by time } t) \leq \mathbf{P}\left(\max_{1 \leq s \leq t} |\tilde{S}_s| \geq \lfloor n/2 \rfloor \right), \tag{9.2}$$

where $(\tilde{S}_t)$ is a random walk on $\mathbb{Z}$ which remains in place with probability $1 - 2/(n-1)$ and increments by ± 1 with equal probability when it moves. (There is inequality in (9.2) instead of equality because the motion of card 1 at 0 is slower than $(|\tilde{S}|_t)$.)

Let (S_t) be the simple random walk on $\mathbb{Z}$: the walker moves one step right or left with equal probability. Viewed only at the times where it moves, $(\tilde{S}_t)$ has the same distribution as (S_t).

By Exercise 9.1,

$$\mathbf{P}\left(\max_{1\le s\le \alpha n^2} |S_s| > \lfloor n/2\rfloor\right) \le 2\mathbf{P}\left(|S_{\alpha n^2}| > \lfloor n/2\rfloor\right) \le \frac{8\mathbf{E}\left(S^2_{\alpha n^2}\right)}{n^2} = 8\alpha.$$

Taking $\alpha_0 = 1/(16\sqrt{2})$, and letting $\tau_{n/2}$ denote the first time (S_t) visits $\lfloor n/2\rfloor$,

$$\mathbf{P}\left(\tau_{n/2} \le \alpha_0 n^2\right) \le \frac{1}{8}. \tag{9.3}$$ {Eq:HitHalf}

Let B_t be the number of times that the delayed random walk $(\tilde{S}_t)$ has moved after t transitions. B_t is a binomial random variable with parameters t and $2/(n-1)$. If $3\beta < \alpha_0$, then

$$\begin{aligned}\mathbf{P}\{B_{\beta n^3} > \alpha_0 n^2\} &\le \mathbf{P}\{B_{\beta n^3} - \mathbf{E}(B_{\beta n^3}) > n^2(\alpha_0 - 3\beta)\}\\ &\le \frac{\operatorname{Var} B_{\beta n^3}}{n^4(\alpha_0-3\beta)^2}\\ &\le \frac{c}{n^2}.\end{aligned}$$

For n large enough, taking $\beta_0 = 1/(64\sqrt{2})$ so that $3\beta_0 < \alpha_0$,

$$\mathbf{P}\{B_{\beta_0 n^3} > \alpha_0 n^2\} \le \frac{1}{8}. \tag{9.4}$$ {Eq:MovesBound}

Putting together equation 9.2 with equation (9.4) shows that

$$\mathbf{P}\left\{\max_{1\le s\le \beta_0 n^3} |\tilde{S}_s| < \lfloor n/2\rfloor\right\} \ge \mathbf{P}\left\{B_{\beta_0 n^3} \le \alpha_0 n^3,\ \tau_{n/2} > \alpha_0 n^2\right\} \ge \frac{7}{8},$$

for n large enough. In other words,

$$\mathbf{P}\{\text{card 1 has visited position } \lfloor n/2\rfloor \text{ by time } \beta_0 n^3\} \le \frac{1}{8},$$

provided n is large enough.

Thus, $P^{\beta_0 n^3}(\mathrm{id}, A) \le 1/8$. Since $\pi(A) \ge 1/2$, it follows that $t_{\mathrm{mix}} \ge \beta_0 n^3$.

{Exercise:RP}

Exercise 9.1 (Reflection Principle). Let (S_n) be the simple random walk on $\mathbb{Z}$. Show that

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

9.3. Riffle shuffles

The method most often used to shuffle real decks of 52 cards is the following: first, the shuffler cuts the decks into two piles. Then, the piles are "riffled" together: she successively drops cards from the bottom of each pile to form a new pile. There are two undetermined aspects of this procedure. First, the numbers of cards in each pile after the initial cut can vary. Second, real shufflers drop varying numbers of cards from each stack as the deck is reassembled.

Fortunately for mathematicians, there is a tractable mathematical model for riffle shuffling. Here are three ways to shuffle a deck of n cards:

(1) Let M be a binomial$(n, 1/2)$ random variable, and split the deck into its top M cards and its bottom $n-M$ cards. There are $\binom{n}{M}$ ways to riffle these two piles together, preserving the relative order within each pile (first select the positions for the top M cards, then fill in both piles). Choose one of these arrangements uniformly at random.

(2) Let M be a binomial$(n, 1/2)$ random variable, and split the deck into its top M cards and its bottom $n-M$ cards. The two piles are then held over the table and cards are dropped one by one, forming a single pile once more, according to the following recipe: if at a particular moment, the left pile contains a cards and the right pile contains b cards, then drop the card on the bottom of the left pile with probability $a/(a+b)$, and the card on the bottom of the right pile with probability $b/(a+b)$. Repeat this procedure until all cards have been dropped.

(3) Label the n cards with n independent fairly chosen bits. Pull all the cards labeled 0 to the top of the deck, preserving their relative order.

A *rising sequence* of a permutation π is a maximal set of consecutive values that occur in the correct relative order in π. (For example, the final permutation in Figure 9.3 has 4 rising sequences: $(1,2,3,4), (5,6), (7,8,9,10)$, and $(11,12,13)$. We claim that methods 1 and 2 generate the same distribution Q on permutations, where

{Eq:RiffleTrans}
$$Q(\sigma) = \begin{cases} (n+1)/2^n & \sigma = \text{id}, \\ 1/2^n & \sigma \text{ has exactly two rising sequences}, \\ 0 & \text{otherwise}. \end{cases} \tag{9.5}$$

It should be clear that method 1 generates Q; the only tricky detail is that the identity permutation is always an option, no matter the value of M. Given M, method 2 assigns probability $M!(n-M)!/n! = \binom{n}{M}^{-1}$ to each possible interleaving, since each step drops a single card and every card must be dropped.

Recall that for a distribution R on $\mathcal{S}_n$, the *inverse distribution* $\overline{R}$ satisfies $\overline{R}(\rho) = R(\rho^{-1})$. We claim that method 3 generates $\overline{Q}$. Why? The cards labeled 0 form one increasing sequence in ρ^{-1}, and the cards labeled 1 form the other. (Again, there are $n+1$ ways to get the identity permutation; here, all strings of the form $00\dots011\dots1$.)

Thanks to Lemma 5.9 (which says that a random walk on a group and its inverse, both started from the identity, have the same distance from uniformity after the same number of steps), it will suffice to analyze method 3.

Now, consider repeatedly inverse riffle shuffling a deck, using method 3. For the first shuffle, each card is assigned a random bit, and all the 0's are pulled ahead of all the 1's. For the second shuffle, each card is again assigned a random bit, and all the 0's are pulled ahead of all the 1's. Considering both bits (and writing the second bit on the left), we see that cards labeled 00 precede those labeled 01, which precede those labeled 10, which precede those labeled 11 (see Figure 9.4). After k shuffles, each card will be labeled with a string of k bits, and cards with

First, cut the deck:

1	2	3	4	5	6		7	8	9	10	11	12	13

Then riffle together.

7	1	8	2	3	9	4	10	5	11	12	6	13

Now, cut again:

7	1	8	2	3	9	4	10		5	11	12	6	13

And riffle again.

5	7	1	8	11	12	2	6	3	13	9	4	10

Fig:riffle
FIGURE 9.3. Riffle shuffling a 13-card deck, twice.

Initial order:

card	1	2	3	4	5	6	7	8	9	10	11	12	13
round 1	1	0	0	1	1	1	0	1	0	1	1	0	0
round 2	0	1	0	1	0	1	1	1	0	0	1	0	1

After one inverse riffle shuffle:

card	2	3	7	9	12	13	1	4	5	6	8	10	11
round 1	0	0	0	0	0	0	1	1	1	1	1	1	1
round 2	1	0	1	0	0	1	0	1	0	1	1	0	1

After two inverse riffle shuffles:

card	3	9	12	1	5	10	2	7	13	4	6	8	11
round 1	0	0	0	1	1	1	0	0	0	1	1	1	1
round 2	0	0	0	0	0	0	1	1	1	1	1	1	1

Fig:RepRevRiffle
FIGURE 9.4. When inverse riffle shuffling, we first assign bits for each round, then sort bit by bit.

different labels will be in lexicographic order (cards with the same label will be in their original relative order).

{Prop:RevRiffleSUT}

PROPOSITION 9.6. *Let τ be the number of inverse riffle shuffles required for all cards to have different bitstring labels. Then τ is a strong uniform time.*

PROOF. Assume $\tau = t$. Since the bitstrings are generated by independent fair coin flips, every assignment of strings of length t to cards is equally likely. Since the labeling bitstrings are distinct, the permutation is fully determined by the labels. Hence the permutation of the cards at time τ is uniform, no matter the value of τ. ■

Now we need only estimate the tail probabilities for the strong uniform time. However, our stopping time τ is an example of the birthday problem, with the slight twist that the number of "people" is fixed, and we wish to choose an appropriate power-of-two “year length” so that all the people will, with high probability, have different birthdays.

{Prop:RiffleUpper}

Proposition 9.7. *For the riffle shuffle on an n-card deck,* $t_{\mathrm{mix}} \leq 2\log_2(4n/3)$ *for sufficiently large n.*

Proof. Consider inverse riffle shuffling an n-card deck and let τ be the stopping time defined in Proposition 9.6. If $\tau \leq t$, then different labels have been assigned to all n cards after t inverse riffle shuffles. Hence

$$\mathbf{P}(\tau \leq t) = \prod_{k=0}^{n-1}\left(1 - \frac{k}{2^t}\right),$$

since there are 2^t possible labels. Let $t = 2\log_2(n/c)$. Then $2^t = n^2/c^2$ and we have

$$\log \prod_{k=0}^{n-1}\left(1 - \frac{k}{2^t}\right) = -\sum_{k=0}^{n-1}\left(\frac{c^2k}{n^2} + O\left(\frac{k}{n^2}\right)^2\right) = -\frac{n(n-1)}{2c^2n^2} + O\left(\frac{n^3}{n^4}\right) = -\frac{c^2}{2} + O\left(\frac{1}{n}\right).$$

Hence

$$\lim_{n\to\infty} \frac{\mathbf{P}(\tau \leq t)}{e^{-c^2/2}} = 1.$$

Taking any value of c such that $c < \sqrt{2\log(4/3)} \approx 0.7585$ will give a bound on $t_{\mathrm{mix}} = t_{\mathrm{mix}}(1/4)$. A convenient value to use is $3/4$, which, combined with Proposition 7.3, gives the bound stated in the proposition.

■

To give a lower bound of logarithmic order on the mixing time for the riffle shuffle, we show that it is unlikely that a uniform random permutation will contain a long rising sequence, but that after a suitable number of riffle shuffles the deck *must* still contain a long rising sequence.

{Prop:RiffleLower}

Proposition 9.8. *Fix* $0 < \varepsilon, \delta < 1$. *Consider riffle shuffling an n-card deck. For sufficiently large n,*

$$t_{\mathrm{mix}}(\varepsilon) \geq (1-\delta)\log_2 n.$$

Proof. Let A be the event that the deck contains a rising sequence of length at least $m = \lceil \log n \rceil$. A uniform random permutation has $n - m + 1$ potential rising sequences of length m (each is a run of m consecutive values) and each has probability $1/m!$ of being increasing. By Stirling’s formula, for sufficiently large n we have

{Eq:StirEstm}

$$m! \geq \frac{(\log n)^{\log n}}{n} \geq n^2. \tag{9.6}$$

The probability of A under the uniform measure is thus bounded above by

$$\frac{1}{n^2} \cdot n = o(1)$$

as $n \to \infty$.

Now, consider riffle shuffling a sorted deck $s < (1-\delta)\log_2 n$ times. Our earlier discussion of the combinatorics of the riffle shuffle imply that the resulting deck has at most $2^s < n^{(1-\delta)}$ rising subsequences. Since the deck is partitioned into disjoint rising sequences, the pigeonhole principle implies that for sufficiently large n at least one of those sequences contains at least $n^\delta > m$ cards. Hence the event A has probability 1 after s riffle shuffles. ■

9.4. Problems

{Exer:PermDistFallacy}

Exercise 9.2. True or false: let Q be a distribution on $\mathcal{S}_n$ such that when $\sigma \in \mathcal{S}_n$ is chosen according to Q, we have

$$\mathbf{P}(\sigma(i) > \sigma(j)) = 1/2$$

for every $i, j \in [n]$. Then Q is uniform on $\mathcal{S}_n$. [Solution]

{Exer:PermDistFallacy2}

Exercise 9.3. Kolata (January 9, 1990) writes: "By saying that the deck is completely mixed after seven shuffles, Dr. Diaconis and Dr. Bayer mean that every arrangement of the 52 cards is equally likely or that any card is as likely to be in one place as in another."

True or false: let Q be a distribution on $\mathcal{S}_n$ such that when $\sigma \in \mathcal{S}_n$ is chosen according to Q, we have

$$\mathbf{P}(\sigma(i) = j) = 1/n$$

for every $i, j \in [n]$. Then Q is uniform on $\mathcal{S}_n$. [Solution]

Exercise 9.4. Let Q be a distribution on $\mathcal{S}_n$. Show that the random walk generated by Q is reversible if and only if $Q(\sigma^{-1}) = Q(\sigma)$ for all $\sigma \in \mathcal{S}_n$.

{Exer:RandTransBadTime}

Exercise 9.5. Consider the random transposition shuffle.

(a) Show that marking both cards of every transposition, and proceeding until every card is marked, does not yield a strong uniform time.

(b) Show that marking the right-hand card of every transposition, and proceeding until every card is marked, does not yield a strong uniform time.

{Exer:MaxProdPhi}

Exercise 9.6. Let $\phi : [n] \to \mathbb{R}$ be any function. Let $\sigma \in \mathcal{S}_n$. Show that the value of

$$\phi_\sigma = \sum_{k \in [n]} \phi(k)\phi(\sigma(k))$$

is maximized when $\sigma = \mathrm{id}$. [Solution]

{Exer:TrigComputation}

Exercise 9.7. Show that for any positive integer n,

$$\sum_{k \in [n]} \cos^2\left(\frac{(2k-1)\pi}{2n}\right) = \frac{n}{2}.$$

[Solution]

{Exer:ashuffle}

Exercise 9.8. Here's a way to generalize the inverse riffle shuffle. Let a be a positive integer. To perform an *inverse a-shuffle*, assign independent uniform random digits chosen from $\{0, 1, \ldots, a-1\}$ to each card. Then sort according to digit, preserving relative order for cards with the same digit. For example, if $a = 3$ and the digits assigned to cards are

1	2	3	4	5	6	7	8	9	10	11	12
2	0	2	1	2	0	1	0	1	0	0	0

then the shuffle will give

2	6	8	10	11	12	4	7	9	1	3	5

(a) Let a and b be positive integers. Show that an inverse a-shuffle followed by an inverse b-shuffle is the same as an inverse ab-shuffle.

(b) Describe (mathematically) how to perform a *forwards* a-shuffle, and show that its increment measure gives weight $\binom{a+n-r}{n}/a^n$ to every $\pi \in \mathcal{S}_n$ with exactly r rising sequences. (This is a generalization of (9.5).)

[Solution]

Remark. Exercise 9.8(b), due to Bayer and Diaconis (1992), is the key to numerically computing the total variation distance from stationarity. A permutation has r rising sequences if and only if its inverse has $r-1$ descents. The number of a permutations in $\mathcal{S}_n$ with $r-1$ descents is the *Eulerian number* $\left\langle {n \atop r-1} \right\rangle$. The Eulerian numbers satisfy a simple recursion (and are built into modern symbolic computation software, such as *Mathematica*); see Graham et al. (1994, p. 267), for details. It follows from Exercise 9.8 that the total variation distance from uniformity after t Gilbert-Shannon-Reeds shuffles of an n-card deck is

$$\sum_{r=1}^{n} \left\langle {n \atop r-1} \right\rangle \left| \frac{\binom{2^t+n-r}{n}}{2^{nt}} - \frac{1}{n!} \right| .$$

See Figure 9.5 for the values when $n = 52$ and $t \leq 12$.

9.5. Notes

9.5.1. Random transpositions. Our upper bound on the mixing time for random transpositions is off by a factor of 4. Matthews (1988b) gives an improved strong stationary time whose upper bound matches the lower bound. Here's how it works: again, let R_t and L_t be the cards chosen by the right and left hands, respectively, at time t. Assume that when $t = 0$, no cards have been marked. As long as at most $\lceil n/3 \rceil$ cards have been marked, use this rule: at time t, mark card R_t if both R_t and L_t are unmarked. When $k > \lceil n/3 \rceil$ cards have been marked, the rule is more complicated. Let $l_1 < l_2 < \cdots < l_k$ be the marked cards, and enumerate the ordered pairs of marked cards in lexicographic order:

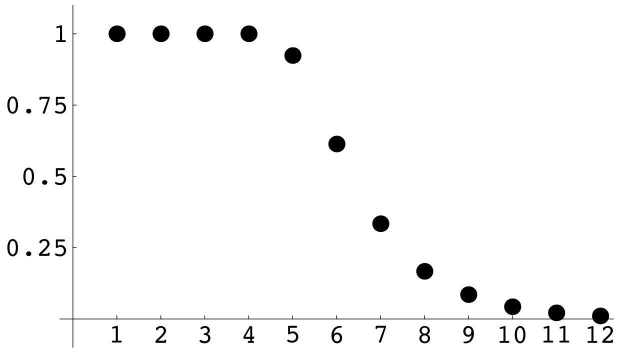
{Eq:MatthewsSST}

$$(l_1, l_1), (l_1, l_2), \ldots, (l_1, l_k), (l_2, l_1), \ldots, (l_k, l_k). \tag{9.7}$$

Also list the unmarked cards in order: $u_1 < u_n < \cdots < u_{n-k}$. At time t, if there exists an i such that $1 \leq i \leq n-k$ and one of the three conditions below is satisfied, then mark card i.

(i) $L_t = R_t = u_i$.

(ii) Either $L_t = u_i$ and R_t is marked, or $R_t = u_i$ and L_t is marked.



1	2	3	4	5	6	7	8	9	10	11	12
1.0000	1.0000	1.0000	1.0000	0.9237	0.6135	0.3341	0.1672	0.0854	0.0429	0.0215	0.0108

Fig:Riffle52TV

FIGURE 9.5. The total variation distance from stationarity (with 4 digits of precision) after t riffle shuffles of a 52-card deck, for $t = 1, \dots, 12$.

(iii) The pair (L_t, R_t) is identical to the i-th pair in the list (9.7) of pairs of marked cards.

(Note that at most one card can be marked per transposition; if case (iii) is invoked, the card marked may not be either of the selected cards.) Compared to the Broder time discussed earlier, this procedure marks cards much faster at the beginning, and essentially twice as fast at the end. The analysis is similar in spirit to, but more complex than, that presented in 9.1.2.

9.5.2. Semi-random transpositions. Consider shuffling by transposing cards. However, we allow only one hand (the right) to choose a uniform random card. The left hand picks a card according to some other rule—perhaps deterministic, perhaps randomized—and the two cards are switched. Since only one of the two cards switched is fully random, it is reasonable to call examples of this type shuffles by *semi-random transpositions*. (Note that for this type of shuffle, the distribution of allowed moves can depend on time.)

One particularly interesting variation first proposed by Thorp (1965) and mentioned as an open problem in Aldous and Diaconis (1986) is the *cyclic-to-random* shuffle: at step t, the left hand chooses card $t \pmod{n}$, the right hand chooses a uniform random card, and the two chosen cards are transposed. This chain has the property that every position is given a chance to be randomized once every n steps. Might that speed randomization? Or does the reduced randomness slow it down? (Note: Exercise 2.2 is about the state of an n-card deck after n rounds of cyclic-to-random transpositions.)

Mironov (2002) (who was interested in how many steps are needed to do a good job of initializing a standard cryptographic protocol) gives an $O(n \log n)$ upper bound, using a variation of Broder's stopping time for random transpositions. Mossel et al. (2004) prove a matching (to within a constant) lower bound. Furthermore, the same authors extend the stopping time argument to give an $O(n \log n)$ upper bound for *any* shuffle by semi-random transpositions.

9.5.3. Riffle shuffles. The most famous theorem in non-asymptotic Markov chain convergence is what is often, and perhaps unfortunately, called the "seven shuffles suffice" (for mixing a standard 52 card deck) result of Bayer and Diaconis (1992), which was featured in the New York Times (Kolata, January 9, 1990). Many elementary expositions of the riffle shuffle have been written. Our account is in debt to Aldous and Diaconis (1986), Diaconis (1988), and Mann (1994).

The model for riffle shuffling that we have discussed was developed by Gilbert and Shannon at Bell Labs in the 50's, and later independently by Reeds. It is natural to ask whether the Gilbert-Shannon-Reeds shuffle is a reasonable model for the way humans riffle cards together. Diaconis (1988) reports that when he and Reeds both shuffled repeatedly, Reeds' shuffles had packet sizes that matched the GSR model well, while Diaconis' shuffles had more small packets. The difference is not surprising, since Diaconis is an expert card magician who can perform perfect shuffles—i.e., ones in which a single card is dropped at a time.

Far more is known about the GSR shuffle than we have discussed. Bayer and Diaconis (1992) derived the exact expression for the probability of any particular permutation after t riffle shuffles discussed in Exercise 9.8. Diaconis et al. (1995) compute exact probabilities of various properties of the resulting permutations and draw beautiful connections with combinatorics and algebra. See Diaconis (2003) for a survey of mathematics that has grown out of the analysis of the riffle shuffle.

Is it in fact true that seven shuffle suffice to adequately randomize a 52 card deck? Bayer and Diaconis (1992) were the first to give explicit values for the total variation distance from stationarity after various numbers of shuffles; see Figure 9.5. After seven shuffles, the total variation distance from stationarity is approximately 0.3341. That is, after 7 riffle shuffles the probability of a given event can differ by as much as 0.3341 from its value under the uniform distribution. Indeed, Peter Doyle has described a simple solitaire game for which the probability of winning when playing with a uniform random deck is exactly $1/2$, but whose probability of winning with a deck that has been GSR shuffled 7 times from its standard order is 0.801 (as computed in van Zuylen and Schalekamp (2004)).

Ultimately the question of how many shuffles suffice for a 52-card deck is one of opinion, not mathematical fact. However, there exists at least one game playable by human beings for which 7 shuffles clearly do not suffice. A more reasonable level of total variation distance might be around 1 percent, comparable to the house advantage in casino games. This threshold would suggest 11 or 12 as an appropriate number of shuffles.

CHAPTER 10

Random Walks on Networks

{Ch:Networks}

10.1. Introduction

We have already met random walks on graphs in Section 3.4. We picture a walker moving on a network of nodes connected by line segments, such as is shown in Figure 10.1. At each move, the walker jumps to one of the nodes connected by a single segment to his current position. How long must he wander before his current location gives little clue about where he started? What is the expected time for him to reach the top-right corner starting from the lower-left corner? Is it likely that he will visit the top-right corner before he returns to his starting position? In this chapter we take up these, and many other, questions.

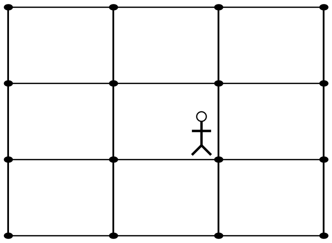


FIGURE 10.1. A random walker on a small grid. Fig:RWalker

10.2. Networks and Reversible Markov Chains

Electrical networks provide a different language for reversible Markov chains; this point of view is useful because of the insight gained from the familiar physical laws of electrical networks.

A *network* is a finite connected graph $G = (V, E)$, endowed with non-negative numbers $\{c(e)\}$, called *conductances*, that are associated to the edges of G. We often write $c(x, y)$ for $c(\{x, y\})$; clearly $c(x, y) = c(y, x)$. The reciprocal $r(e) = 1/c(e)$ is called the *resistance* of the edge e. A network will be denoted by the pair $(G, \{c(e)\})$. Vertices of G are often called *nodes*. V will denote the vertex set of G, and for $x, y \in V$, we will write $x \sim y$ to indicate that $\{x, y\}$ belongs to the edge set of G.

Consider the Markov chain on the nodes of G with transition matrix

$$P(x, y) = \frac{c(x, y)}{c(x)}, \tag{10.1}$$

{Eq:WeightedRW}

where $c(x) = \sum_{y\,:\,y\sim x} c(x,y)$. This process is called the *weighted random walk* on G with edge weights $\{c(e)\}$, or the Markov chain associated to the network $(G,\{c(e)\})$. This Markov chain is reversible with respect to the probability π defined by $\pi(x) := c(x)/c$, where $c_G = \sum_{x\in V} c(x)$:

$$\pi(x)P(x,y) = \frac{c(x)}{c_G}\frac{c(x,y)}{c(x)} = \frac{c(x,y)}{c_G} = \frac{c(y,x)}{c_G} = \frac{c(y)}{c_G}\frac{c(y,x)}{c(y)} = \pi(y)P(y,x).$$

Note that

$$c_G = \sum_{x\in V}\sum_{\substack{y\in V\\ y\sim x}} c(x,y) = 2\sum_{e\in E} c(e).$$

Simple random walk on G, defined in Section 3.4 as the Markov chain with transition probabilities

$$P(x,y) = \begin{cases} \frac{1}{\deg(x)} & \text{if } y\sim x,\\ 0 & \text{otherwise,}\end{cases} \tag{10.2}$$

is a special case of a weighted random walk: set the weights of all edges in G equal to 1.

We now show that in fact every reversible Markov chain is a weighted random walk on a network. Suppose P is a transition probability on Ω which is reversible with respect to the probability π (that is, (3.27) holds.) Define a graph with vertex set Ω by declaring $\{x,y\}$ an edge if $P(x,y) > 0$. This is a proper definition, since reversibility implies that $P(x,y) > 0$ exactly when $P(y,x) > 0$. Next, define conductances on edges by $c(x,y) = \pi(x)P(x,y)$. This is symmetric by reversibility. With this choice of weights, we have $c(x) = \pi(x)$, and thus the transition matrix associated with this network is just P. The study of reversible Markov chains is thus reduced to the study of random walks on networks.

10.3. Harmonic Functions and Voltage

Recall from Section 3.5.4 that we call a real-valued function h defined on the vertices of G *harmonic* at a vertex x if

{eq:harmonic}

$$h(x) = \sum_{y\,:\,y\sim x} P(x,y)h(y), \tag{10.3}$$

where P is the transition matrix defined in (10.1). This means that $h(x)$ is the weighted average of its neighboring values, where the weights are determined by the conductances.

We distinguish two nodes, a and z, which are called the *source* and the *sink* of the network. A function W which is harmonic on $G\setminus\{a,z\}$ will be called a *voltage*. A voltage is completely determined by its boundary values $W(a)$ and $W(z)$. In particular, the following result, whose proof should remind you of that of Lemma 3.9, is derived from the maximum principle.

{prop:6.1}

Proposition 10.1. *Let h be a function on a network $(G,\{c(e)\})$ which is harmonic on $G\setminus\{a,z\}$ and such that $h(a) = h(z) = 0$. Then h must vanish everywhere on G.*

Proof. We will first show that $h \leq 0$. Suppose this is not the case. Let $x \notin \{a, z\}$ belong to the set $A = \{x : h(x) = \max_G h\}$ and choose a neighbor y of x. By harmonicity of h on $G \setminus \{a, z\}$, if $h(y) < \max_G h$, then

$$h(x) = \sum_{z\,:\,z\sim x} h(z)P(x,z) = h(y)P(x,y) + \sum_{\substack{z\,:\,z\sim x,\\ z\neq y}} h(z)P(x,z) < \max_G h,$$

a contradiction. It follows that $h(y) = \max_G h$, that is, $y \in A$. By connectedness, $a, z \in A$, hence $h(a) = h(z) = \max_G h > 0$, contradicting our assumption. Thus $h \leq 0$. An application of this result to $-h$ also yields $h \geq 0$. ■

If h and g are two harmonic functions satisfying the boundary conditions $h(a) = g(a) = x$ and $h(z) = g(z) = y$, then the function $k = h - g$ is a harmonic function with $k(a) = k(z) = 0$. By Proposition 10.1, $k \equiv 0$, that is, $g = h$. This proves that given boundary conditions $h(a) = x$ and $h(z) = y$, if there is a function harmonic on $G \setminus \{a, z\}$ satisfying these boundary conditions, it is unique. To prove that a harmonic function with given boundary values exists, observe that the conditions (10.3) in the definition of harmonic functions form a system of linear equations with the same number of equations as unknowns, namely (number of nodes in G) $- 2$; for such a system, uniqueness of solutions implies existence.

We can also prove existence more constructively, using random walk on the underlying network. To get a voltage with boundary values 0 and 1 at z and a respectively, set

$$W^\star(x) := \mathbf{P}_x\{\tau_a < \tau_z\}, \tag{10.4}$$

where $\mathbf{P}_x$ is the probability for the walk started at node x. (You should check that W^* is actually a voltage!) To extend to arbitrary boundary values W_a and W_z for $W(a)$ and $W(z)$, respectively, define

$$W(x) = W_z + W^\star(x)\left[W_a - W_z\right]. \tag{10.5}$$ {Eq:HarmonicExists}

The reader should check that this function has all the required properties (Exercise 10.2).

Until now, we have focused on *undirected graphs*. Now we need to consider also *directed graphs*. An *oriented edge* is an *ordered* pair of nodes (x, y), which we denote by $\vec{e} = \vec{xy}$. A *directed graph* consists of a vertex set together with a collection of oriented edges. Of course, any network can be viewed as a directed graph; for each unoriented edge in the network, include both orientations in the directed graph.

A *flow* θ from a to z is a function on oriented edges which is antisymmetric, $\theta(\vec{xy}) = -\theta(\vec{yx})$, and which obeys *Kirchhoff's node law*:

$$\sum_{w\,:\,w\sim v} \theta(\vec{vw}) = 0 \text{ at all } v \notin \{a, z\}. \tag{10.6}$$

This is just the requirement "flow in equals flow out" for any node not a or z.

Observe that it is only flows that are defined on oriented edges. Conductance and resistance are defined for unoriented edges; we may of course define them on oriented edges by $c(\vec{xy}) = c(\vec{yx}) = c(x, y)$ and $r(\vec{xy}) = r(\vec{yx}) = r(x, y)$.

Given a voltage W on the network, the *current flow* associated with W is defined on oriented edges by

$$I(\vec{xy}) = \frac{W(x) - W(y)}{r(\vec{xy})} = c(x,y)\left[W(x) - W(y)\right]. \tag{10.7}$$

{Eq:CFDef}

This definition immediately implies that the current flow satisfies *Ohm's law*: if $\vec{e} = \vec{xy}$,

{Eq:OhmsLaw}

$$r(\vec{e})I(\vec{e}) = W(x) - W(y). \tag{10.8}$$

Also notice that I is antisymmetric and satisfies the node law at every $x \notin \{a, z\}$:

$$\begin{aligned}\sum_{y\,:\,y\sim x} I(\vec{xy}) &= \sum_{y\,:\,y\sim x} c(x,y)[W(x) - W(y)] \\ &= c(x)W(x) - c(x)\sum_{y\,:\,y\sim x} W(y)P(x,y) = 0.\end{aligned}$$

Thus the node law for the current is equivalent to the harmonicity of the voltage.

Finally, current flow also satisfies the *cycle law*: if the edges $\vec{e}_1, \ldots, \vec{e}_m$ form a cycle, i.e., $\vec{e}_i = (x_{i-1}, x_i)$ and $x_n = x_0$, then

$$\sum_{i=1}^{m} r(\vec{e}_i)I(\vec{e}_i) = 0\,. \tag{10.9}$$

Notice that adding a constant to all values of a voltage affects neither its harmonicity nor the current flow it determines. Hence we may, without loss of generality, fix a voltage function W on our network for which $W(z) = 0$.

We define the *strength* of an arbitrary flow θ by

$$\|\theta\| = \sum_{x\,:\,x\sim a} \theta(\vec{ax}). \tag{10.10}$$

A *unit flow* is a flow of strength 1.

{prop:6.3}

Proposition 10.2 (Node law/cycle law/strength). *If θ is a flow from a to z satisfying the cycle law*

$$\sum_{i=1}^{m} r(\vec{e}_i)\theta(\vec{e}_i) = 0 \tag{10.11}$$

for any cycle $\vec{e}_1 \ldots, \vec{e}_m$, and if $\|\theta\| = \|I\|$, then $\theta = I$.

Proof. The function $f = \theta - I$ satisfies the node law at all nodes and the cycle law. Suppose $f(e_1) > 0$ for some directed edge e_1. By the node law, e_1 must lead to some directed edge e_2 with $f(e_2) > 0$. Iterate this process to obtain a sequence of edges on which f is strictly positive. Since the underlying network is finite, this sequence must eventually revisit a node. The resulting cycle violates the cycle law. ■

10.4. Effective Resistance

Given a network, the ratio $[W(a) - W(z)]/\|I\|$, where I is the current flow corresponding to the voltage W, is independent of the voltage W applied to the network. Define the *effective resistance* between vertices a and z as

{Eq:ERDefn}
$$\mathcal{R}(a \leftrightarrow z) := \frac{W(a) - W(z)}{\|I\|}. \tag{10.12}$$

In parallel with our earlier definitions, we also define the *effective conductance* $C(a \leftrightarrow z) = 1/\mathcal{R}(a \leftrightarrow z)$. Why is $\mathcal{R}(a \leftrightarrow z)$ called the "effective resistance" of the network? Imagine replacing our entire network by a single edge joining a to z with resistance $\mathcal{R}(a \leftrightarrow z)$. If we now apply the same voltage to a and z in both networks, then the amount of current flowing from a to z in the single-edge network is the same as in the original.

Next, we discuss the probabilistic interpretation of effective resistance. By (10.5), for any vertex x

$$\mathbf{P}_x\{\tau_z < \tau_a\} = \frac{W(a) - W(x)}{W(a) - W(z)}. \tag{10.13}$$
{Eq:HitBefore}

We have

$$\mathbf{P}_a\{\tau_z < \tau_a^+\} = \sum_{x \in V} P(a,x)\mathbf{P}_x\{\tau_z < \tau_a\} = \sum_{x\,:\,x \sim a} \frac{c(a,x)}{c(a)} \frac{W(a) - W(x)}{W(a) - W(z)}. \tag{10.14}$$

Then using the definition of current flow (10.7), the above equals

$$\frac{\sum_{x\,:\,x \sim a} I(\vec{ax})}{c(a)\,[W(a) - W(z)]} = \frac{\|I\|}{c(a)\,[W(a) - W(z)]} = \frac{1}{c(a)\mathcal{R}(a \leftrightarrow z)}, \tag{10.15}$$

showing that

$$\mathbf{P}_a\{\tau_z < \tau_a^+\} = \frac{1}{c(a)\mathcal{R}(a \leftrightarrow z)} = \frac{C(a \leftrightarrow z)}{c(a)}. \tag{10.16}$$
{Eq:EscapeResistance}

The *Green's function* for the random walk stopped at a stopping time τ is defined by

$$G_\tau(a,x) = \mathbf{E}_a\,(\text{number of visits to } x \text{ before } \tau) = \mathbf{E}_a\left(\sum_{t=0}^{\infty} \mathbf{1}_{\{X_t = x, \tau > t\}}\right). \tag{10.17}$$
{Eq:GreenFunctionDefn}

{Lem:GreensFunctionResi

Lemma 10.3. *If $G_{\tau_z}(a,a)$ is the Green's function defined in (10.17), then*

$$G_{\tau_z}(a,a) = c(a)\mathcal{R}(a \leftrightarrow z). \tag{10.18}$$
{Eq:GreensFunctionResis

Proof. The number of visits to a before visiting z has a geometric distribution with parameter $\mathbf{P}_a\{\tau_z < \tau_a^+\}$. The lemma then follows from (10.16). ■

It is often possible to replace a network by a simplified one without changing quantities of interest, for example the effective resistance between a pair of nodes. The following laws are very useful.

Parallel Law. Conductances in parallel add: suppose edges e_1 and e_2, with conductances c_1 and c_2 respectively, share vertices v_1 and v_2 as endpoints. Then both edges can be replaced with a single edge of conductance $c_1 + c_2$ without affecting the rest of the network. All voltages and currents in $G \setminus \{e_1, e_2\}$ are unchanged and the current $I(\vec{e})$ equals $I(\vec{e}_1) + I(\vec{e}_2)$. For a proof, check Ohm's and Kirchhoff's laws with $I(\vec{e}) := I(\vec{e}_1) + I(\vec{e}_2)$.

Series Law. Resistances in series add: if $v \in V \setminus \{a, z\}$ is a node of degree 2 with neighbors v_1 and v_2, the edges (v_1, v) and (v, v_2) can be replaced by a single edge (v_1, v_2) of resistance $r_{v_1 v} + r_{v v_2}$. All potentials and currents in $G \setminus \{v\}$ remain the same and the current that flows from v_1 to v_2 equals $I(\vec{v_1 v}) = I(\vec{v v_2})$. For a proof, check again Ohm's and Kirchhoff's laws, with $I(\vec{v_1 v_2}) := I(\vec{v_1 v}) = I(\vec{v v_2})$.

Gluing. Another convenient operation is to identify vertices having the same voltage, while keeping all existing edges. Because current never flows between vertices with the same voltage, potentials and currents are unchanged.

Example 10.4. For a tree Γ with root ρ, let Γ_n be the vertices at distance n from ρ. Consider the case of a spherically symmetric tree, in which all vertices of Γ_n have the same degree for all $n \geq 0$. Suppose that all edges at the same distance from the root have the same resistance, that is, $r(e) = r_i$ if $|e| = i$, $i \geq 1$. Glue all the vertices in each level; this will not affect effective resistances, so we infer that

$$\mathcal{R}(\rho \leftrightarrow \Gamma_M) = \sum_{i=1}^{M} \frac{r_i}{|\Gamma_i|} \tag{10.19}$$

and

$$\mathbf{P}_\rho\{\tau_{\Gamma_M} < \tau_\rho^+\} = \frac{r_1/|\Gamma_1|}{\sum_{i=1}^M r_i/|\Gamma_i|}. \tag{10.20}$$

Therefore, $\lim_{M\to\infty} \mathbf{P}_\rho\{\tau_{\Gamma_M} < \tau_\rho^+\} > 0$ if and only if $\sum_{i=1}^{\infty} r_i/|\Gamma_i| < \infty$.

Example 10.5 (Biased nearest-neighbor random walk). Consider the network with vertices $\{0, 1, \ldots, n\}$, edges $e_k = \{k, k-1\}$ for $k = 1, 2, \ldots, n$, and weights $c(e_k) = \alpha^k$. Then

$$P(k, k+1) = \frac{\alpha}{1+\alpha},$$
$$P(k, k-1) = \frac{1}{1+\alpha}.$$

If $\alpha = p/(1-p)$, then this is the walk which when at interior vertices moves up with probability p and down with probability $1-p$.

Using the series law, we can replace the k edges to the left of k by a single edge of resistance

$$\sum_{j=1}^{k} \alpha^{-j} = \frac{\alpha^{-(k+1)} - \alpha^{-1}}{1 - \alpha^{-1}}.$$

Likewise, we can replace the $(n-k)$ edges to the right of k by a single edge of resistance

$$\sum_{j=k+1}^{n} \alpha^{-j} = \frac{\alpha^{-(n+1)} - \alpha^{-(k+1)}}{1 - \alpha^{-1}}.$$

The probability $\mathbf{P}_k\{\tau_n < \tau_0\}$ is not changed by this modification, so we can calculate simply that

$$\begin{aligned}\mathbf{P}_k\{\tau_n < \tau_0\} &= \frac{(1-\alpha^{-1})/(\alpha^{-(n+1)} - \alpha^{-(k+1)})}{(1-\alpha^{-1})/(\alpha^{-(n+1)} - \alpha^{-(k+1)}) + (1-\alpha^{-1})/(\alpha^{-(k+1)} - \alpha^{-1})} \\ &= \frac{\alpha^{-k} - 1}{\alpha^{-n} - 1}.\end{aligned}$$

In particular, for the biased random walk which moves up with probability p,

$$\mathbf{P}_k\{\tau_n < \tau_0\} = \frac{[(1-p)/p]^k - 1}{[(1-p)/p]^n - 1}. \tag{10.21}$$ {Eq:BiasedGR}

{Thm:ThompsonsPrinciple

Theorem 10.6 (Thomson's Principle). *For any finite connected graph,*

$$\mathcal{R}(a \leftrightarrow z) = \inf\left\{\mathcal{E}(\theta) :\ \theta \textit{ a unit flow from a to z}\right\}, \tag{10.22}$$

where $\mathcal{E}(\theta) := \sum_e [\theta(e)]^2 r(e)$. *The unique minimizer in the* inf *above is the unit current flow.*

Remark. The sum in $\mathcal{E}(\theta)$ is over unoriented edges, so each edge $\{x, y\}$ is only considered once in the definition of energy. Although θ is defined on oriented edges, it is antisymmetric and hence $\theta(e)^2$ is unambiguous.

Proof. By compactness, there exists a flow θ minimizing $\mathcal{E}(\theta)$ subject to $\|\theta\| = 1$. By Proposition 10.2, to prove that the unit current flow is the unique minimizer, it is enough to verify that any unit flow θ of minimal energy satisfies the cycle law.

Let the edges $\vec{e}_1, \ldots, \vec{e}_n$ form a cycle. Set $\gamma(\vec{e}_i) = 1$ for all $1 \le i \le n$ and set γ equal to zero on all other edges. Note that γ satisfies the node law, so it is a flow, but $\sum \gamma(\vec{e}_i) = n \neq 0$. For any $\varepsilon \in \mathbb{R}$, we have that

$$\begin{aligned}0 \le \mathcal{E}(\theta + \varepsilon\gamma) - \mathcal{E}(\theta) &= \sum_{i=1}^n \left[\left(\theta(\vec{e}_i) + \varepsilon\right)^2 - \theta(\vec{e}_i)^2\right] r(\vec{e}_i) \\ &= 2\varepsilon \sum_{i=1}^n r(\vec{e}_i)\theta(\vec{e}_i) + O(\varepsilon^2).\end{aligned}$$

By taking $\varepsilon \to 0$ from above and from below, we see that $\sum_{i=1}^n r(e_i)\theta(\vec{e}_i) = 0$, thus verifying that θ satisfies the cycle law.

To complete the proof, we show that the unit current flow I has $\mathcal{E}(I) = \mathcal{R}(a \leftrightarrow z)$:

$$\begin{aligned}\sum_e r(e)I(e)^2 &= \frac{1}{2}\sum_x \sum_y r(x,y)\left[\frac{W(x) - W(y)}{r(x,y)}\right]^2 \\ &= \frac{1}{2}\sum_x \sum_y c(x,y)[W(x) - W(y)]^2 \\ &= \frac{1}{2}\sum_x \sum_y [W(x) - W(y)] I(\vec{xy}).\end{aligned}$$

Since I is antisymmetric,

$$\frac{1}{2}\sum_x \sum_y [W(x) - W(y)]I(\vec{xy}) = \sum_x W(x) \sum_y I(\vec{xy}). \qquad (10.23)$$ {eq:as}

By the node law, $\sum_y I(\vec{xy}) = 0$ for any $x \notin \{a, z\}$, while $\sum_y I(\vec{ay}) = \|I\| = -\sum_y I(\vec{zy})$, so the right-hand side of (10.23) equals

$$\|I\| \left(W(a) - W(z)\right).$$

Recalling that $\|I\| = 1$, we conclude that the right-hand side of (10.23) is equal to $(W(a) - W(z))/\|I\| = \mathcal{R}(a \leftrightarrow z)$. ■

Let a, z be vertices in a network, and suppose that we add to the network an edge which is not incident to a. How does this affect the escape probability from a to z? From the point of view of probability, the answer is not obvious. In the language of electrical networks, this question is answered by:

{thm:6.5}

Theorem 10.7 (Rayleigh's Monotonicity Law). *If $\{r(e)\}$ and $\{r'(e)\}$ are sets of resistances on the edges of the same graph G, and if $r(e) \le r'(e)$ for all e, then*

$$\mathcal{R}(a \leftrightarrow z; r) \ \le \ \mathcal{R}(a \leftrightarrow z; r'). \qquad (10.24)$$

[$\mathcal{R}(a \leftrightarrow z; r)$ is the effective resistance computed with the resistances $\{r(e)\}$, while $\mathcal{R}(a \leftrightarrow z; r')$ is the effective resistance computed with the resistances $\{r'(e)\}$.]

Proof. Note that $\inf_\theta \sum_e r(e)\theta(e)^2 \ \le \ \inf_\theta \sum_e r'(e)\theta(e)^2$ and apply Thomson's Principle (Theorem 10.6). ■

{cor:6.6}

Corollary 10.8. *Adding an edge does not increase the effective resistance $\mathcal{R}(a \leftrightarrow z)$. If the added edge is not incident to a, the addition does not decrease the escape probability $\mathbf{P}_a\{\tau_z < \tau_a^+\} = [c(a)\mathcal{R}(a \leftrightarrow z)]^{-1}$.*

Proof. Before we add an edge to a network we can think of it as existing already with $c = 0$ or $r = \infty$. By adding the edge we reduce its resistance to a finite number.

Combining this with the relationship (10.16) shows that the addition of an edge not incident to a (which we regard as changing a conductance from 0 to 1) cannot decrease the escape probability $\mathbf{P}_a\{\tau_z < \tau_a^+\}$. ■

{Cor:Glue}

Corollary 10.9. *The operation of gluing vertices cannot increase effective resistance.*

Proof. When we glue vertices together, we take an infimum over a larger class of flows. ■

Moreover, if we glue together vertices with different potentials, then effective resistance will strictly decrease. A technique due to Nash-Williams (1959) often gives simple but useful lower bounds on effective resistance.

An *edge-cutset* Π *separating a from* z is a set of edges with the property that any path from a to z must include some edge in Π.

{Prop:NW}

Proposition 10.10 (Nash-Williams (1959)). *If $\{\Pi_n\}$ are disjoint edge-cutsets which separate nodes a and z, then*

{eq:nw}

$$\mathcal{R}(a \leftrightarrow z) \geq \sum_n \left(\sum_{e \in \Pi_n} c(e) \right)^{-1} . \tag{10.25}$$

The inequality (10.25) *is called the* Nash-Williams inequality.

Proof. Let θ be a unit flow from a to z. For any n, by the Cauchy-Schwarz inequality

$$\sum_{e \in \Pi_n} c(e) \cdot \sum_{e \in \Pi_n} r(e)\theta(e)^2 \geq \left(\sum_{e \in \Pi_n} \sqrt{c(e)}\sqrt{r(e)}|\theta(e)| \right)^2 = \left(\sum_{e \in \Pi_n} |\theta(e)| \right)^2$$

The right-hand side is bounded below by $\|\theta\|^2 = 1$, because Π_n is a cutset and $\|\theta\| = 1$. Therefore

$$\sum_e r(e)\theta(e)^2 \geq \sum_n \sum_{e \in \Pi_n} r(e)\theta(e)^2 \geq \sum_n \left(\sum_{e \in \Pi_n} c(e) \right)^{-1} .$$

■

10.5. Escape Probabilities on a Square

Let B_n be the $n \times n$ two-dimensional grid graph: the vertices are pairs of integers (z, w) such that $1 \leq z, w \leq n$, while the edges are pairs of points at unit (Euclidean) distance.

{Prop:ResisBn}

Proposition 10.11. *Let $a = (1, 1)$ be the lower left-hand corner of B_n, and let $z = (n, n)$ be the upper right-hand corner of B_n. The effective resistance $\mathcal{R}(a \leftrightarrow z)$ satisfies*

$$\frac{\log(n-1)}{2} \leq \mathcal{R}(a \leftrightarrow z) \leq 2 \log n. \tag{10.26}$$

{Eq:ResisBn}

We separate the proof into the lower and upper bounds.

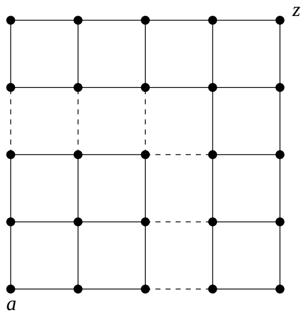


Figure 10.2. The graph B_5. The cutset Π_3 contains the edges drawn with dashed lines. Fig:SquareCutset

PROOF OF LOWER BOUND IN (10.26). Let Π_k be the edge-set

$$\Pi_k = \{(v, w) \ : \ |v| = k - 1,\ |w| = k\},$$

where $|v|$ is the length of the shortest path from v to a (see Figure 10.2). Since every path from a to z must use an edge in Π_k, the set Π_k is a cutset. Since each edge has unit conductance, $\sum_{e\in\Pi_k} c(e)$ just equals the number of edges in Π_k, namely $2k$. By Proposition 10.10,

{Eq:BnLower}

$$\mathcal{R}(a \leftrightarrow z) \geq \sum_{k=1}^{n-1} \frac{1}{2k} \geq \frac{\log(n-1)}{2}. \tag{10.27}$$

■

We now establish the upper bound:

PROOF OF UPPER BOUND IN (10.26). Thomson's Principle (Theorem 10.6) says that the effective resistance is the minimal possible energy of a unit flow from a to z. So to get an upper bound on resistance, we build a unit flow on the square.

Consider the Polya's urn process, described in Section 4.3.3. The sequence of ordered pairs listing the numbers of black and white balls is a Markov chain with state space $\{1, 2, \ldots\}^2$.

Run this process on the square—note that it necessarily starts at vertex $a = (1, 1)$—and stop when you reach the main diagonal $x + y = n + 1$. Direct all edges of the square from bottom left to top right and give each edge e on the bottom left half of the square the flow

$$f(e) = \mathbf{P}\{\text{the process went through } e\}.$$

To finish the construction, give the the upper right half of the square the symmetrical flow values.

From Lemma 4.4, it follows that for any $k \geq 0$, the Polya's urn process is equally likely to pass through each of the $k + 1$ pairs (i, j) for which $i + j = k + 2$. Consequently, when (i, j) is a vertex in the square for which $i + j = k + 2$, the sum of the flows on its incoming edges is $\frac{1}{k+1}$. Thus the energy of the flow f can be bounded by

$$\mathcal{E}(f) \leq \sum_{k=1}^{n-1} 2\left(\frac{1}{k+1}\right)^2 (k+1) \leq 2 \log n.$$

■

{Exercise:PolyaHighD}

EXERCISE 10.1. Generalize the flow in the upper bound of (10.26) to higher dimensions, using an urn with balls of d colors. Use this to show that the resistance between opposite corners of the d-dimensional box of side length n is bounded independent of n, when $d \geq 3$.

10.6. Problems

xercise:HarmonicExists}

EXERCISE 10.2. Check that the the function W defined in (10.5) has all required properties: that is, show it satisfies (10.3) at all vertices $x \notin \{a, z\}$, and show it satisfies the boundary conditions $W(a) = W_a$ and $W(z) = W_z$.

{Exercise:Umbrella}

Exercise 10.3. An Oregon professor has n umbrellas, of which initially $k \in (0, n)$ are at his office and $n - k$ are at his home. Every day, the professor walks to the office in the morning and returns home in the evening. In each trip, he takes an umbrella with him only if it is raining. Assume that in every trip between home and office or back, the chance of rain is $p \in (0, 1)$, independently of other trips. {It:Umb12}

(a) For $p = 1/2$,
   (i) How many states are needed to model this process as a Markov chain?
   (ii) Determine the stationary distribution. Asymptotically, in what fraction of his trips does the professor get wet?
   (iii) Determine the expected number of trips until all n umbrellas are at the same location.
   (iv) Determine the expected number of trips until the professor gets wet. {It:UmbBias}

(b) Same as (a) but for arbitrary p.

Part (a) can be solved using the random walker in Figure 4.1. Part (b) requires an analysis of a *biased* random walk, which moves right and left with unequal probabilities.

{Exer:DDP}

Exercise 10.4 (Discrete Dirichlet Problem). Let $(G, \{c(e)\})$ be a network, and let $A \subset V$ be a collection of vertices. Given a function $g : A \to \mathbb{R}$, show that there is a unique extension of g to V so that g is harmonic on $V \setminus A$.

{Exercise:GambRuinResis

Exercise 10.5 (Gambler's Ruin). Consider the simple random walk on $\{0, 1, 2, \dots, n\}$. Use the network reduction laws to show that $\mathbf{P}_x\{\tau_n < \tau_0\} = x/n$

Exercise 10.6. Show that $\mathcal{R}(a \leftrightarrow z)$ is a concave function of $\{r(e)\}$.

Exercise 10.7. Let B_n be the subset of $\mathbb{Z}^2$ contained in the box of side length $2n$ centered at 0. Let ∂B_n be the set of vertices along the perimeter of the box. Show that

$$\lim_{n\to\infty} \mathbf{P}_0\{\tau_{\partial B_n} < \tau_a^+\} = 0.$$

{Exercise:ResisMetric}

Exercise 10.8. Show that effective resistances form a metric on any network with conductances $\{c(e)\}$.

Hint: The only non-obvious statement is the triangle inequality

$$\mathcal{R}(x \leftrightarrow z) \le \mathcal{R}(x \leftrightarrow y) + \mathcal{R}(y \leftrightarrow z).$$

Adding the unit current flow from x to y to the unit current flow from y to z gives the unit current flow from x to z (check Kirchoff's laws!). Now use the corresponding voltage functions.

10.7. Notes

The basic reference for the connection between electrical networks and random walks on graphs is Doyle and Snell (1984), and we borrow here from Peres (1999).

CHAPTER 11

Hitting and Cover Times

11.1. Hitting Times

Global maps are often unavailable for real networks that have grown without central organization, such as the Internet. However, sometimes the structure can be queried locally, meaning that given a specific node v, for some cost all nodes connected by a single link to v can be determined. How can such local queries be used to determine whether two nodes v and w can be connected by a path in the network?

Suppose you have limited storage, but are not concerned about time. In this case, one approach to is to start a random walker at v, allow the walker to explore the graph for some time, and observe whether the node w is ever encountered. If the walker visits node w, then clearly v and w must belong to the same connected component of the network. On the other hand, if node w has not been visited by the walker by time t, it is possible that w is not accessible from v—but perhaps the the walker was simply unlucky. It is of course important to distinguish between these two possibilities! In particular, when w is connected to v, we desire an estimate of expected time until the walker visits w starting at v.

With this in mind, it is natural to define the *hitting time* τ_A of a subset A of nodes to be the first time one of the nodes in A is visited by the random walker: If (X_t) is the random walk, let

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

We will simply write τ_w for $\tau_{\{w\}}$, consistent with our notation in Section 3.5.2.

We have already seen the usefulness of hitting times. In Section 3.5.2 we used a variant

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

(called the *first return time* in the situation that $X_0 = x$) to build a candidate stationary distribution. In Section 6.3, we used the expected absorption time for the "gambler's ruin" problem (computed in Section 4.1) to bound the expected coupling time for the torus.

To connect our discussion of hitting times for random walks on networks to our leitmotif of mixing times, we mention now the problem of estimating the mixing time for two "glued" tori, the graph considered in Example 8.2.

Let V_1 be the collection of nodes in the right-hand torus, and let $v^\star$ be the node connecting the two tori.

When the walk is started at a node x in the left-hand torus, we have

$$\|P^t(x,\cdot) - \pi\|_{\text{TV}} \geq \pi(V_1) - P^t(x, V_1) = \frac{1}{2} - \mathbf{P}_x\{X_t \in V_1\} \geq \frac{1}{2} - \mathbf{P}_x\{\tau_{v^\star} \leq t\}. \tag{11.1}$$ {Eq:DistanceTwoTori}

If the walker is unlikely to have exited the left-hand torus by time t, then (11.1) shows that $d(t)$ is not much smaller 1/2. In view of this, it is not surprising that estimates for $\mathbf{E}_x(\tau_{v^\star})$ are useful for bounding t_{mix} for this chain. These ideas are developed in Section 11.7.

11.2. Hitting times and random target times

{Sec:HitTarg} {Lem:RandomTarget}

Lemma 11.1 (Random Target Lemma). *For an irreducible Markov chain with state space Ω, transition matrix P, and stationary distribution π, the quantity*

$$\sum_{x \in \Omega} \mathbf{E}_a(\tau_x)\pi(x)$$

does not depend on $a \in \Omega$.

Proof. For notational convenience, let $h_x(a) = \mathbf{E}_a(\tau_x)$. Observe that if $x \neq a$,

$$h_x(a) = \sum_{y \in \Omega} \mathbf{E}_a(\tau_x \mid X_1 = y)P(a,y) = \sum_{y \in \Omega} (1 + h_x(y))\, P(a,y) = (Ph_x)(a) + 1,$$

so that

{Eq:RT1} $$(Ph_x)(a) = h_x(a) - 1. \tag{11.2}$$

If $x = a$, then

$$\mathbf{E}_a(\tau_a^+) = \sum_{y \in \Omega} \mathbf{E}_a(\tau_a^+ \mid X_1 = y)P(a,y) = \sum_{y \in \Omega} (1 + h_a(y))\, P(a,y) = 1 + (Ph_a)(a).$$

Since $\mathbf{E}_a(\tau_a^+) = \pi(a)^{-1}$,

{Eq:RT2} $$(Ph_a)(a) = \frac{1}{\pi(a)} - 1. \tag{11.3}$$

Thus, letting $h(a) := \sum_{x \in \Omega} h_x(a)\pi(x)$, (11.2) and (11.3) show that

$$(Ph)(a) = \sum_{x \in \Omega} (Ph_x)(a)\pi(x) = \sum_{x \neq a} (h_x(a) - 1)\pi(x) + \pi(a)\left(\frac{1}{\pi(a)} - 1\right).$$

Simplifying the right-hand side and using that $h_a(a) = 0$ yields

$$(Ph)(a) = h(a).$$

That is, h is harmonic. Applying Lemma 3.9 shows that h is a constant function. ■

Consider choosing a state $y \in \Omega$ according to π. Lemma 11.1 says that the expected time to hit the "random target" state y from a specified starting state a does not depend on a. Hence we can define the *target time* of an irreducible chain by

$$t_{\text{trgt}} = \sum_{x \in \Omega} \mathbf{E}_a(\tau_x)\pi(x) = \mathbf{E}_\pi(\tau_\pi)$$

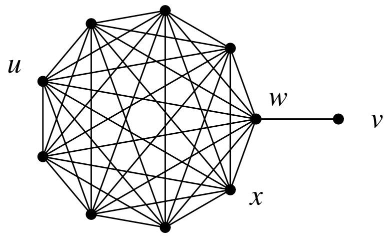


Fig:CompleteLeaf

FIGURE 11.1. For random walk on this family of graphs, $t_{\text{hit}} \gg t_{\text{trgt}}$.

(the last version is a slight abuse of our notation for hitting times). Since t_{trgt} does not depend on the state a, it is true that

{Eq:ttarg}

$$t_{\text{trgt}} = \sum_{x,y\in\Omega} \pi(x)\pi(y)\mathbf{E}_x(\tau_y) = \mathbf{E}_\pi(\tau_\pi). \tag{11.4}$$

We will often find it useful to estimate the worst-case hitting times between states in a chain. Define

$$t_{\text{hit}} = \max_{x,y\in\Omega} \mathbf{E}_x(\tau_y). \tag{11.5}$$

{Eq:ThitDef}

{Lem:HitBound}

LEMMA 11.2. *For an irreducible Markov chain with state space Ω and stationary distribution π,*

$$t_{\text{hit}} \le 2\max_w \mathbf{E}_\pi(\tau_w).$$

PROOF. For any $a, y \in \Omega$, we have

$$\mathbf{E}_a(\tau_y) \le \mathbf{E}_a(\tau_\pi) + \mathbf{E}_\pi(\tau_y), \tag{11.6}$$

{Eq:HitBound}

since we can insist that the chain go from x to y via a random state x chosen according to π. By Lemma 11.1,

$$\mathbf{E}_a(\tau_\pi) = \mathbf{E}_\pi(\tau_\pi) \le \max_w \mathbf{E}_\pi(\tau_w).$$

It is now clear that (11.6) implies the desired inequality. ■

Note that for a transitive chain,

$$t_{\text{trgt}} = \mathbf{E}_\pi(\tau_\pi) = \sum_{x\in\Omega} \mathbf{E}_a(\tau_x)\pi(x) = \sum_{x,y\in\Omega} \pi(y)\mathbf{E}_y(\tau_x)\pi(x) = \mathbf{E}_\pi(\tau_b).$$

Hence we have

{Cor:TransHitTargBound}

COROLLARY 11.3. *For an irreducible transitive Markov chain,*

$$t_{\text{hit}} \le 2t_{\text{trgt}}.$$

EXAMPLE 11.4. When the underlying chain is not transitive, it is possible for t_{hit} to be much larger than t_{trgt}. Consider the example of simple random walk on a complete graph on n vertices with a leaf attached to one vertex (see Figure 11.1). Let v be the leaf and let w be the neighbor of the leaf; call the other vertices *ordi-*

nary. Let the initial state of the walk be v. The first return time from v to v satisfies both

$$\mathbf{E}_v\tau_v^+ = \mathbf{E}_v\tau_w + \mathbf{E}_w\tau_v = 1 + \mathbf{E}_w\tau_v$$

(since the walk must take its first step to w) and

$$\mathbf{E}_v\tau_v^+ = \frac{1}{\pi(v)} = \frac{2\binom{n}{2}+1}{1} = n^2 - n + 2,$$

by Exercise 3.20 and Example 3.6. Hence $\mathbf{E}_w\tau_v = n^2 - n + 1 \leq t_{\mathrm{hit}}$.

By the random target lemma, we can use any state to estimate t_{trgt}. Let's start at v. Clearly $\mathbf{E}_v\tau_v = 0$, while $\mathbf{E}_v\tau_w = 1$ and $\mathbf{E}_v\tau_u = 1 + \mathbf{E}_w\tau_u$, where u is any ordinary vertex. How long does it take to get from w to u, on average? Let x be any *other* ordinary vertex. By conditioning on the first step of the walk, and exploiting symmetry, we have

$$\begin{aligned}\mathbf{E}_w\tau_u &= 1 + \frac{1}{n}\left(\mathbf{E}_v\tau_u + (n-2)\mathbf{E}_x\tau_u\right)\\ &= 1 + \frac{1}{n}\left(1 + \mathbf{E}_w\tau_u + (n-2)\mathbf{E}_x\tau_u\right)\end{aligned}$$

and

$$\mathbf{E}_x\tau_u = 1 + \frac{1}{n-1}\left(\mathbf{E}_w\tau_u + (n-3)\mathbf{E}_x\tau_u\right)$$

We have two equations in the two "variables" $\mathbf{E}_w\tau_u$ and $\mathbf{E}_x\tau_u$. Solving yields

$$\mathbf{E}_w\tau_u = \frac{n^2-n+4}{n} = O(n) \quad \text{and} \quad \mathbf{E}_x\tau_u = \frac{n^2-n+2}{n} = O(n)$$

(we only care about the first equation right now). Combining these results with Example 3.6 yields

$$\begin{aligned}t_{\mathrm{trgt}} = \mathbf{E}_v\tau_\pi &= \pi(v)(0) + \pi(w)(1) + (n-1)\pi(u)O(n)\\ &= \frac{1(0) + n(1) + (n-1)^2 O(n)}{2\left(\binom{n}{2}+1\right)} = O(n) \ll t_{\mathrm{hit}}.\end{aligned}$$

11.3. Commute Time

The *commute time* between nodes a and b in a network is the time to move from a to b and then back to a:

{Eq:CommuteTimeDefn}
$$\tau_{a,b} = \min\{t \geq \tau_b \ : \ X_t = a\}, \tag{11.7}$$

where we assume that $X_0 = a$. The commute time is of intrinsic interest and can be computed or estimated using resistance (the *commute time identity*, Proposition 11.6). In graphs for which $\mathbf{E}_a(\tau_b) = \mathbf{E}_b(\tau_a)$, the expected hitting time is half the commute time, so estimates for the commute time yield estimates for hitting times. *Transitive* networks enjoy this property (Proposition 11.7).

The following lemma will be used in the proof of the commute time identity:

{Lem:Renewal}

Lemma 11.5 (Aldous, Fill). *If τ is a stopping time for a finite and irreducible Markov chain satisfying $\mathbf{P}_a\{X_\tau = a\} = 1$, and $G_\tau(a,x)$ is the Green's function (as defined in (10.17)) then*

$$\frac{G_\tau(a,x)}{\mathbf{E}_a(\tau)} = \pi(x) \qquad \text{for all } x.$$

Exercise 11.1. Prove Lemma 11.5 by copying the proof in Proposition 3.8 that $\tilde{\pi}$ as defined in (3.18) satisfies $\tilde{\pi} = \tilde{\pi}P$, substituting $G_\tau(a,x)/\mathbf{E}_a(\tau)$ in place of $\tilde{\pi}$.

{Prop:CommuteTimeIdenti

Proposition 11.6 (Commute Time Identity). *Let $(G, \{c(e)\})$ be a network, and let (X_t) be the random walk on this network. For any nodes a and b in V, let $\tau_{a,b}$ be the commute time defined in (11.7) between a and b. Then*

$$\mathbf{E}_a(\tau_{a,b}) = \mathbf{E}_a(\tau_b) + \mathbf{E}_b(\tau_a) = c_G \mathcal{R}(a \leftrightarrow b). \tag{11.8}$$

{Eq:CommuteTimeIdentity

(Recall that $c(x) = \sum_{x \in V} c(x)$ and that $c_G = \sum_{x \in V} c(x) = 2\sum_{e \in E} c(e)$.)

Proof. By Lemma 11.5,

$$\frac{G_{\tau_{a,b}}(a,a)}{\mathbf{E}_a(\tau)} = \pi(a) = \frac{c(a)}{c_G}.$$

By definition, after visiting b the chain does not visit a until time $\tau_{a,b}$, so $G_{\tau_{a,b}}(a,a) = G_{\tau_b}(a,a)$. The conclusion follows from Lemma 10.3. ■

Note that $\mathbf{E}_a(\tau_b)$ and $\mathbf{E}_b(\tau_a)$ can be very different for general Markov chains, and even for reversible chains (see Exercise 11.6). However, for certain types of random walks on networks they are equal. A network $\langle G, \{c(e)\}\rangle$ is *transitive* if for any pair of vertices $x, y \in V$ there exists a permutation $\psi_{x,y} : V \to V$ with

$$\psi_{x,y}(x) = y, \qquad \text{and} \qquad c(\psi_{x,y}(u), \psi_{x,y}(v)) = c(u,v) \text{ for all } u, v \in V. \tag{11.9}$$

{Eq:psidef}

Remark. In Section 7.5 we defined transitive Markov chains. The reader should check that a random walk on a transitive graph is a transitive Markov chain.

{Prop:Eabba}

Proposition 11.7. *For a simple random walk on a transitive connected graph G, for any vertices $a,\ b \in V$,*

$$\mathbf{E}_a(\tau_b) = \mathbf{E}_b(\tau_a) \tag{11.10}$$

{Eq:Eabba}

Before proving this, it is helpful to establish the following identity:

{Lem:CycleIdentity}

Lemma 11.8. *For any three states a, b, c of a reversible Markov chain,*

$$\mathbf{E}_a(\tau_b) + \mathbf{E}_b(\tau_c) + \mathbf{E}_c(\tau_a) = \mathbf{E}_a(\tau_c) + \mathbf{E}_c(\tau_b) + \mathbf{E}_b(\tau_a)$$

Proof. We can reword this lemma as

$$\mathbf{E}_a(\tau_{bca}) = \mathbf{E}_a(\tau_{cba}), \tag{11.11}$$

{reword}

where τ_{bca} is the time to visit b, then visit c, and then hit a. It turns out that it is much easier to start at stationarity, since it allows us to use reversibility easily. Recall that we use $\mathbf{E}_\pi(\cdot)$ to denote the expectation operator for the chain started with initial distribution π.

Adding $\mathbf{E}_\pi(\tau_a)$ to both sides of (11.11), we find it is enough to show that

$$\mathbf{E}_\pi(\tau_{abca}) = \mathbf{E}_\pi(\tau_{acba}).$$

In fact, we will show equality in distribution, not just expectation. Suppose ξ and $\xi^\star$ are finite strings with letters in V, meaning $\xi \in V^m$ and $\xi^\star \in V^n$ with $m \leq n$. We say that $\xi \preceq \xi^\star$ if and only if ξ is a subsequence of $\xi^\star$, that is, there exist indices $1 \leq i_1 < \cdots < i_m \leq n$ with $\xi(k) = \xi^\star(i_k)$ for all $1 \leq k \leq m$. Using the identity (3.29) for reversed chains,

{Eq:abca1}

$$\mathbf{P}_\pi\{\tau_{abca} > k\} = \mathbf{P}_\pi\{abca \npreceq X_0 \dots X_k\} = \mathbf{P}_\pi\{abca \npreceq X_k \dots X_0\}. \tag{11.12}$$

Clearly, $abca \preceq X_k \dots X_0$ is equivalent to $acba \preceq X_0 \dots X_k$ (just read from right-to-left!), so the right-hand side of (11.12) equals

$$\mathbf{P}_\pi\{acba \npreceq X_0 \dots X_k\} = \mathbf{P}_\pi\{\tau_{acba} > k\}.$$

■

Proof of Proposition 11.7. Let ψ be a map satisfying the conditions (11.9) with $u = a$ and $v = b$. Let $a_0 = a$, and $a_j = \psi^{(j)}(a_0)$ for $j \geq 1$, where $\psi^{(j)}$ denotes the j-th iterate of ψ. The sequence $a_0, a_1, \dots$ will return to a_0 eventually; say $a_m = a_0$, where $m > 0$. The function $\psi^{(j)}$ takes a, b to a_j, a_{j+1}, so for any j,

$$\mathbf{E}_{a_j}(\tau_{a_{j+1}}) = \mathbf{E}_a(\tau_b), \tag{11.13}$$

Summing over j from 0 to $m-1$ we obtain

{Eq:cycle}

$$\mathbf{E}_{a_0}(\tau_{a_1 a_2 \dots a_{m-1} a_0}) = m\mathbf{E}_a(\tau_b). \tag{11.14}$$

For the same reason,

{Eq:cyclereverse}

$$\mathbf{E}_{a_0}(\tau_{a_{m-1} a_{m-2} \dots a_1 a_0}) = m\mathbf{E}_b(\tau_a) \tag{11.15}$$

By the same argument as we used for (11.11), we see that the left hand sides of equation (11.14) and (11.15) are the same. This proves (11.10). ■

11.4. Hitting Times for the Torus

{Sec:HitTimeTorus}

Putting together Exercise 11.8, Proposition 11.7, and the Commute Time Identity (Proposition 11.6), it follows that for random walk on the d-dimensional torus,

HitTimeTorusResistance}

$$\mathbf{E}_a(\tau_b) = 2n^d \mathcal{R}(a \leftrightarrow b). \tag{11.16}$$

(For an unweighted graph, $c = 2 \times |\text{edges}|$.) Thus, to get estimates on the hitting time $\mathbf{E}_a(\tau_b)$, it is enough to get estimates on the effective resistance.

{Prop:HitForTorus}

Proposition 11.9. *Let x and y be two points at distance k in the torus $\mathbb{Z}_n^d$, and let τ_y be the time of the first visit to y. There exist constants $0 < c_d \leq C_d < \infty$ such that*

{Eq:TorHit3d}

$$c_d n^d \leq \mathbf{E}_x(\tau_y) \leq C_d n^d \qquad \textit{uniformly in } k \textit{ if } d \geq 3, \tag{11.17}$$

{Eq:TorHit2d}

$$c_2 n^2 \log(k) \leq \mathbf{E}_x(\tau_y) \leq C_2 n^2 \log(k) \qquad \textit{if } d = 2. \tag{11.18}$$

PROOF. First, the lower bounds. Choose Π_j to be the box centered around x of side-length $2j$. There is a constant κ_1 so that for $j \leq \kappa_1 k$, the box Π_j is a cutset separating x from y. Note Π_j has order j^{d-1} edges. By Proposition 10.10,

$$\mathcal{R}(a \leftrightarrow z) \geq \sum_{j=1}^{\kappa_1 k} \kappa_2 j^{1-d} \geq \begin{cases} \kappa_3 \log(k) & \text{if } d = 2, \\ \kappa_3 & \text{if } d \geq 3. \end{cases}$$

The lower bounds in (11.17) and (11.18) are then immediate from (11.16).

If the points x and y are the diagonally opposite corners of a square, the upper bound in (11.18) follows using the flow constructed from Polya's urn in Proposition 10.11.

Now consider the case where x and y are in the corners of a non-square rectangle. Examine Figure 11.2. Connect x and y via a third point z, where z is on a vertical line segment going through x and on a horizontal line segment through y. Suppose that the path connecting x to z has $2i$ edges, and the path connecting z to y has $2j$ segments. (Note $2i + 2j = k$, since x and y are at distance k.) Now let u be the point diagonal to x in a $2i \times 2i$ square on one side of the path from x to z (see again Figure 11.2.) Define v similarly. We construct 4 flows and concatenate them: flow from x to u, from u to z, from z to v, and from v to y. Each of these flows is constructed via Polya's urn, as in Proposition 10.11. Note that the edges in these four flows are disjoint, so we find the energy $\mathcal{E}$ by adding the energies of the four individual flows. Each has energy bounded by $c \log(k)$. Using Thomson's Principle, the resistance is then bounded above by $c \log(k)$. If the path lengths are not even, just direct the flow all along the last edge in the path. This establishes the upper bound (11.18).

The upper bound in (11.17) uses the resistance bound in Exercise 10.1.

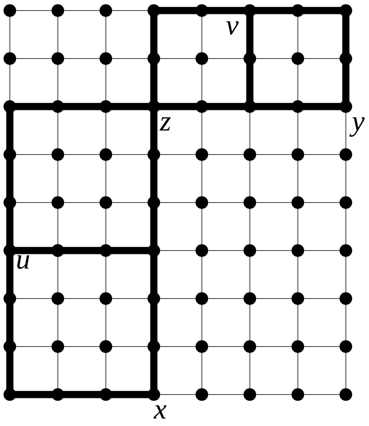


FIGURE 11.2. Constructing a flow from a to z. Fig:Flow

■

11.5. Hitting Times for Birth-and-Death Chains

A *birth-and-death* chain has state-space $\Omega = \{0, 1, 2, \dots, n\}$, and moves only to neighboring integers (or remains in place.) The transition probabilities are specified by $\{(p_k, r_k, q_k)\}_{k=0}^n$, where $p_k + r_k + q_k = 1$ and

- p_k is the probability of moving from k to $k+1$ when $0 \le k < n$,
- q_k is the probability of moving from k to $k-1$ when $0 < k \le n$,
- r_k is the probability of remaining at k when $0 < l < n$,
- At 0, the chain remains at 0 with probability $r_0 + q_0$,
- At n, the chain remains at n with probability $r_n + p_n$.

To find the stationary distribution of the chain, we need to solve the equations

$$\begin{aligned} \pi(k) &= \pi(k)r_k + \pi(k-1)p_{k-1} + \pi(k+1)q_{k+1} \qquad \text{for } 1 < k < n, \\ \pi(0) &= \pi(0)\left[r_0 + q_0\right] + \pi(1)q_1 \\ \pi(n) &= \pi(n)\left[r_n + q_n\right] + \pi(n-1)p_{n-1}. \end{aligned}$$

Solving,

$$\begin{aligned} \pi(1) &= \frac{(1 - r_0 - q_0)\pi(0)}{q_1} = \frac{p_0}{q_1}\pi(0) \\ \pi(2) &= \frac{\pi(1)\left[1 - r_k\right] - \pi(0)p_0}{q_2} = \frac{p_0 p_1}{q_1 q_2}\pi(0) \\ &\vdots \\ \pi(n) &= \frac{\pi(n-1)p_{n-1}}{1 - r_n - p_n} = \frac{p_0 p_1 \cdots p_{n-2} p_{n-1}}{q_1 \cdots q_{n-1} q_n}\pi(0) \end{aligned}$$

That is,

$$\pi(k) = c_{p,r,q} \prod_{j=1}^{k} \frac{p_j}{q_j},$$

where $c_{p,r,q} := \left[\sum_{k=0}^n \prod_{j=1}^k \frac{p_j}{q_j}\right]^{-1}$ is a normalizing constant.

Fix $\ell \in \{0, 1, \dots, n\}$ and consider the restriction of 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 ℓ, the chain either moves down or remains in place, with probabilities q_ℓ and $r_\ell + p_\ell$, respectively.

We write $\tilde{\mathbf{E}}$ for expectations for this new chain. The stationary probability $\tilde{\pi}$ is given by

$$\tilde{\pi}(k) = \frac{\pi(k)}{\pi(\{0, 1, \dots, \ell\})}.$$

Thus,

{Eq:BDHit1}
$$\frac{\pi(\{0, 1, \dots, \ell\})}{\pi(\ell)} = \frac{1}{\tilde{\pi}(\ell)} = \tilde{\mathbf{E}}_\ell(\tau_\ell^+) = 1 + q_\ell \tilde{\mathbf{E}}_{\ell-1}(\tau_\ell) \tag{11.19}$$

Note that $\tilde{\mathbf{E}}_{\ell-1}(\tau_\ell) = \mathbf{E}_{\ell-1}(\tau_\ell)$, and rearranging (11.19) shows that

$$\begin{aligned}\mathbf{E}_{\ell-1}(\tau_\ell) &= \frac{\pi(\{0,1,\dots,\ell\})/\pi(\ell) - 1}{q_\ell} = \frac{\pi(\{0,1,\dots,\ell-1\})}{\pi(\ell) q_\ell} \\ &= \frac{\sum_{k=0}^{\ell-1} \prod_{j=1}^{k} \left(\frac{p_j}{q_j}\right)}{\prod_{j=1}^{\ell} \left(\frac{p_j}{q_j}\right) q_\ell}.\end{aligned}$$

In the special case that (p_k, r_k, q_k) does not depend on k and $p \neq q$, we get

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

When $p_k = q_k$, we get

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

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

$$\begin{aligned}\mathbf{E}_a(\tau_b) &= \sum_{\ell=a+1}^{b} \mathbf{E}_{\ell-1}(\tau_\ell) \\ &= \sum_{\ell=a+1}^{b} \frac{\sum_{k=0}^{\ell-1} \prod_{j=1}^{k} \left(\frac{p_j}{q_j}\right)}{\prod_{j=1}^{\ell} \left(\frac{p_j}{q_j}\right) q_\ell}.\end{aligned}$$

{Xmpl:EhrUrnHit}

Example 11.10 (Ehrenfest Urn). Suppose d balls are split between two urns, labelled A and B. At each move, a ball is selected at random and moved from its current urn to the other urn. If the location of each ball is recorded, the chain has state-space $\{0,1\}^d$ and is the familiar random walk on the hypercube. We consider instead the chain which just tracks the number of balls in urn A. The transition probabilities are, for $0 \leq k \leq d$,

$$\begin{aligned}P(k, k+1) &= \frac{d-k}{d} \\ P(k, k-1) &= \frac{k}{d}.\end{aligned}$$

This is a birth-and-death chain with $p_k = (d-k)/d$ and $q_k = k/d$.

11.6. Bounding Mixing Times via Hitting Times

The goal of this section is to prove the following:

{thm:mixhit}

Theorem 11.11. *Consider a finite reversible chain with transition matrix P and stationary distribution π on Ω.*

{It:HitMix1}

(i) *For all $m \geq 0$ and $x \in \Omega$, we have*

$$\|P^m(x,\cdot) - \pi\|_{\mathrm{TV}}^2 \leq \frac{1}{4}\left[\frac{P^{2m}(x,x)}{\pi(x)} - 1\right]. \tag{11.20}$$

{cauchy}

(ii) *If the chain satisfies* $P(x, x) \geq 1/2$ *for all* x*, then* {It:HitMix2}

$$t_{\mathrm{mix}}(1/4) \leq 2 \max_{x \in \Omega} \mathbf{E}_{\pi}(\tau_x) + 1. \qquad (11.21)$$

{eq:mixhit}

Remark 11.1. (i) says that the total variation distance to stationarity starting from x, for reversible chains, can be made small just by making the return time to x close to its stationary probability.

Remark 11.2. Note that by conditioning on X_0,

$$\mathbf{E}_{\pi}(\tau_x) = \sum_{y \in \Omega} \mathbf{E}_y(\tau_x)\pi(y) \leq \max_{y \in \Omega} \mathbf{E}_y(\tau_x) = t_{\mathrm{hit}}.$$

Thus the bound (11.21) implies

{eq:mixhitmax}

$$t_{\mathrm{mix}}(1/4) \leq 2t_{\mathrm{hit}} + 1. \qquad (11.22)$$

To prove this, we will need a few preliminary results.

{Prop:Mono}

Proposition 11.12. *Let* P *be the transition matrix for a finite reversible chain on state-space* Ω *with stationary distribution* π*.*

{It:PmMono}
{It:PmMonoL}

(i) *For all* $t \geq 0$ *and* $x \in \Omega$ *we have* $P^{2t+2}(x, x) \leq P^{2t}(x, x)$*.*

(ii) *If the chain* P_L *is lazy, that is* $P_L(x, x) \geq 1/2$ *for all* x*, then for all* $t \geq 0$ *and* $x \in \Omega$ *we have* $P_L^{t+1}(x, x) \leq P_L^t(x, x)$*.*

See Exercise 12.4 in Chapter 12 for a proof using eigenvalues. Here, we give a direct proof using the Cauchy-Schwarz inequality.

Proof. (i) Since $P^{2t+2}(x, x) = \sum_{y,z \in \Omega} P^t(x, y)P^2(y, z)P^t(z, x)$, we have

{decomp}

$$\pi(x)P^{2t+2}(x, x) = \sum_{y,z \in \Omega} P^t(y, x)\pi(y)P^2(y, z)P^t(z, x) = \sum_{y,z \in \Omega} \psi(y, z)\psi(z, y), \qquad (11.23)$$

where $\psi(y, z) = P^t(y, x)\sqrt{\pi(y)P^2(y, z)}$. (By Exercise 3.14, the matrix P^2 is reversible with respect to π.)

By Cauchy-Schwarz, the right-hand side of (11.23) is at most

$$\sum_{y,z \in \Omega} \psi(y, z)^2 = \sum_{y \in \Omega} [P^t(y, x)]^2 \pi(y) = \pi(x)P^{2t}(x, x).$$

(ii) Given a lazy chain $P_L = (P + I)/2$, enlarge the state space by adding a new state $m_{xy} = m_{yx}$ for each pair of states $x, y \in \Omega$. (See Figure 11.3.)

On the larger state space Ω_K define a transition matrix K by

$$\begin{aligned} K(x, m_{xy}) &= P(x, y) && \text{for } x, y \in \Omega, \\ K(m_{xy}, x) = K(m_{xy}, y) &= 1/2 && \text{for } x \neq y, \\ K(m_{xx}, x) &= 1 && \text{for all } x, \end{aligned}$$

other transitions having K-probability 0. Then K is reversible with stationary measure π_K given by $\pi_K(x) = \pi(x)/2$ for $x \in \Omega$ and $\pi_K(m_{xy}) = \pi(x)P(x, y)$. Clearly $K^2(x, y) = P_L(x, y)$ for $x, y \in \Omega$, so $K^{2t}(x, y) = P_L^t(x, y)$, and the claimed monotonicity follows. ■

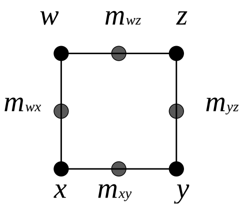


FIGURE 11.3. Adding states m_{xy} for each pair $x, y \in \Omega$. Fig:mxy

The following proposition, that does not require reversibility, relates the mean hitting time of a state x to return probabilities.

{Prop:Warm}

PROPOSITION 11.13 (Hitting time from stationarity). *Consider a finite irreducible aperiodic chain with transition matrix P with stationary distribution π on Ω. Then for any $x \in \Omega$,*

$$\pi(x)\mathbf{E}_\pi(\tau_x) = \sum_{t=0}^{\infty}[P^t(x,x) - \pi(x)]. \tag{11.24}$$ {Eq:Warm}

We give two proofs, one using generating functions and one using stopping times, following (Aldous and Fill, in progress, Lemma 11, Chapter 2).

PROOF OF PROPOSITION 11.13 VIA GENERATING FUNCTIONS. Define

$$f_k := \mathbf{P}_\pi\{\tau_x = k\} \quad \text{and} \quad u_k := P^k(x,x) - \pi(x).$$

Since $\mathbf{P}_\pi\{\tau_x = k\} \le \mathbf{P}_\pi\{\tau_x \ge k\} \le C\alpha^k$ for some $\alpha < 1$ (see (3.17)), the power series $F(s) := \sum_{k=0}^\infty f_k s^k$ converges in an interval $[0, 1+\delta_1]$ for some $\delta_1 > 0$.

Also, since $|P^k(x,x) - \pi(x)| \le d(k)$, and $d(k)$ decays at least geometrically fast (Theorem 5.6), $U(s) := \sum_{k=0}^\infty u_k s^k$ converge in an interval $[0, 1+\delta_2]$ for some $\delta_2 > 0$. Note that $F'(1) = \sum_{k=0}^\infty k f_k = \mathbf{E}_\pi(\tau_x)$ and $U(1)$ equals the right-hand side of (11.24).

For every $m \ge 0$,

$$\pi(x) = \mathbf{P}_\pi\{X_m = x\} = \sum_{k=0}^{m} f_k P^{m-k}(x,x) = \sum_{k=0}^{m} f_k \left[\left(P^{m-k}(x,x) - \pi(x)\right) + \pi(x)\right]$$
$$= \sum_{k=0}^{m} f_k[u_{m-k} + \pi(x)].$$

Thus, the constant sequence with every element equal to $\pi(x)$ is the convolution of the sequence $\{f_k\}_{k=0}^\infty$ with the sequence $\{u_k - \pi(x)\}_{k=0}^\infty$, so its generating function $\sum_{m=0}^\infty \pi(x)s^m = \pi(x)(1-s)^{-1}$ equals the product of the generating function F with the generating function

$$\sum_{m=0}^{\infty}[u_m - \pi(x)]s^m = U(s) - \pi(x)\sum_{m=0}^{\infty} s^m = U(S) - \frac{\pi(x)}{1-s}.$$

(See Exercise 11.15.) That is, for $0 < s < 1$,

$$\frac{\pi(x)}{1-s} = \sum_{m=0}^{\infty} \pi(x)s^m = F(s)\left[U(s) + \frac{\pi(x)}{1-s}\right],$$

and multiplying by $1 - s$ gives $\pi(x) = F(s)[(1-s)U(s) + \pi(x)]$, which clearly holds also for $s = 1$. Differentiating the last equation at $s = 1$, we obtain that $0 = F'(1)\pi(x) - U(1)$, and this is equivalent to (11.24). ■

Proof of Proposition 11.13 via stopping times. Define

$$\tau_x^{(m)} := \min\{t \geq m \, : \, X_t = x\},$$

and write $\mu_m := P^m(x, \cdot)$. By the Convergence Theorem (Theorem 5.6), μ_m tends to π as $m \to \infty$. By Lemma (11.5), we can represent the expected number of visits to x before time $\tau_x^{(m)}$ as

$$\sum_{k=0}^{m-1} P^k(x,x) = \pi(x)\mathbf{E}_x\left(\tau_x^{(m)}\right) = \pi(x)[m + \mathbf{E}_{\mu_m}(\tau_x)].$$

Thus $\sum_{k=0}^{m-1}[P^k(x,x) - \pi(x)] = \pi(x)\mathbf{E}_{\mu_m}(\tau_x)$.

Taking $m \to \infty$ completes the proof. ■

We are now able to prove Theorem 11.11.

Proof of Theorem 11.11. (i) By Cauchy-Schwarz,

$$\left(\frac{1}{2}\sum_{y\in\Omega}\pi(y)\left|\frac{P^m(x,y)}{\pi(y)} - 1\right|\right)^2 \leq \sum_{y\in\Omega}\pi(y)\left[\frac{P^m(x,y)}{\pi(y)} - 1\right]^2 .$$

Therefore

$$\|P^m(x,\cdot) - \pi\|_{\mathrm{TV}}^2 \leq \frac{1}{4}\sum_{y\in\Omega}\left[\frac{P^m(x,y)P^m(y,x)}{\pi(x)} - 2P^m(x,y) + 1\right] = \frac{1}{4}\left[\frac{P^{2m}(x,x)}{\pi(x)} - 1\right].$$

(ii) By the identity (11.24) in Proposition 11.13 and the monotonicity in Proposition 11.12(ii), for any $m > 0$ we have

$$\pi(x)\mathbf{E}_\pi(\tau_x) \geq \sum_{k=1}^{2m}[P^k(x,x) - \pi(x)] \geq 2m[P^{2m}(x,x) - \pi(x)].$$

Dividing by $8m\,\pi(x)$ and invoking (11.20) gives

$$\frac{\mathbf{E}_\pi(\tau_x)}{8m} \geq \|P^m(x,\cdot) - \pi\|_{\mathrm{TV}}^2,$$

and the left-hand side is less than $1/16$ for $m \geq 2\mathbf{E}_\pi(\tau_x)$. ■

{Xmpl:CycleMixHit}

Example 11.14 (Random walks on cycles). We have already derived an $O(n^2)$ bound for the mixing time of the lazy random walk on the cycle C_n, using coupling—it is the dimension 1 case of Theorem 6.4. However, Theorem 11.11 can also be used, and gives a result for the simple (non-lazy) random walk on odd cycles. (Simple random walk on even cycles is periodic; see Example 3.4.)

Label the states of C_n with $\{0, 1, \ldots, n-1\}$. By identifying the states 0 and n, we can see that $\mathbf{E}_k\tau_0$ for the simple random walk on the cycle must be the same as the expected time to ruin or success in a gambler's ruin on the path $\{0, 1, \ldots, n\}$. Hence, for simple random walk on the cycle, Exercise 4.1 implies

$$t_{\text{hit}} = \max_{x,y} \mathbf{E}_x\tau_y = \max_{0\leq k\leq n} k(n-k) = \frac{\lfloor n^2 \rfloor}{4}.$$

For odd n, (11.22) gives

$$t_{\text{mix}} \leq \frac{n^2-1}{2} + 1 = \frac{n^2+1}{2}.$$

For lazy random walk on any cycle, whether even or odd, we have $t_{\text{hit}} = \lfloor n^2 \rfloor/2$, so

$$t_{\text{mix}} \leq n^2 + 1.$$

Example 11.15 (Random walk on binary trees). In Example 8.4 the lazy random walk on the binary tree of depth k was defined, and a lower bound on t_{mix} was obtained via the bottleneck ratio. Here we obtain an upper bound of the same order.

The maximal hitting time between two vertices is obtained for ℓ_1 and ℓ_2 two leaves whose most recent common ancestor is the root v_0. This hitting time is equal to the commute time from the root to one of the leaves, say ℓ_1. For convenience, we first consider the simple random walk without holding. Using the Commute Time Identity in Equation 11.8, c_G is the number of edges and equals $2(n-1)$, and the effective resistance equals the depth k. Thus,

$$\max_{x,y\in\Omega} \mathbf{E}_y(\tau_x) = \mathbf{E}_{\ell_1}(\tau_{\ell_1}) = \mathbf{E}_{v_0}(\tau_{\ell_1}) + \mathbf{E}_{\ell_1}(\tau_{v_0}) = 2(n-1)k.$$

For the lazy walk, this expected time is doubled, since at each move the chain remains in place with probability $1/2$.

Using Theorem 11.11(ii), this shows that $t_{\text{mix}} = O(n \log n)$. (The number of vertices n and the depth k are related by $n = 2^{k+1} - 1$.) The lower bound obtained in Example 8.4 was of order n – which is indeed the correct order for t_{mix}.

11.6.1. Cesaro mixing time. Let the Markov chain $(X_t)_{t\geq 0}$ have stationary distribution π. The stopping time τ is a *stationary time* for the chain if $\mathbf{P}_x\{X_\tau = y\} = \pi(y)$ for arbitrary states x, y.

The simplest stationary time is the first hitting time of a state chosen independently according to π.

Consider a finite chain (X_t) with transition matrix P and stationary distribution π on Ω. Given $t \geq 1$, suppose that we choose uniformly a time $\sigma \in \{0, 1, \ldots, t-1\}$, and run the Markov chain for σ steps. Then the state X_σ has distribution

$$\nu_x^t := \frac{1}{t}\sum_{s=0}^{t-1} P^s(x, \cdot). \qquad (11.25) \quad \{\texttt{cesaro}\}$$

The *Cesaro mixing time* $t_{\text{mix}}^{\star}(\varepsilon)$ is defined as the first t such that for all $x \in \Omega$,

$$\|\nu_x^t - \pi\|_{\text{TV}} \leq \varepsilon\,.$$

See Exercises 11.17 through 11.19 for some properties of the Cesaro mixing time.

The following general result due to Lovász and Winkler (1998) shows that the expectation of any stationary time yields an upper bound for $t^{\star}_{\text{mix}}(1/4)$. Remarkably, this does not need reversibility or laziness. Lovász and Winkler also prove a converse of this result.

{thm:lovwink}

Theorem 11.16. *Consider a finite chain with transition matrix P and stationary distribution π on Ω. If τ is a stationary time for the chain, then $t^{\star}_{\text{mix}}(1/4) \leq 4 \max_{x \in \Omega} \mathbf{E}_x(\tau) + 1$.*

Proof. Denote by ν^t_x the Cesaro average (11.25). Since τ is a stationary time, so is $\tau + s$ for every $s \geq 1$. Therefore, for every $y \in \Omega$,

$$t\pi(y) = \sum_{s=0}^{t-1} \mathbf{P}_x\{X_{\tau+s} = y\} = \sum_{\ell=0}^{\infty} \mathbf{P}_x\{X_\ell = y,\ \tau \leq \ell < \tau + t\}.$$

Consequently,

$$t\nu^t_x(y) - t\pi(y) \leq \sum_{\ell=0}^{t-1} \mathbf{P}_x\{X_\ell = y,\ \tau > \ell\}.$$

Summing the last inequality over all $y \in \Omega$ such that the right-hand side is positive,

$$t\|\nu^t_x - \pi\|_{\text{TV}} = \sum_{\ell=0}^{t-1} \mathbf{P}_x\{\tau > \ell\} \leq \mathbf{E}_x(\tau).$$

Thus for $t \geq 4\mathbf{E}_x(\tau)$ we have $\|\nu^t_x - \pi\|_{\text{TV}} \leq 1/4$. ■

{Sec:TwoGraphMix}

11.7. Mixing for the Walker on Two Glued Graphs

We state the main result of this section:

{Prop:ConvergeTwo}

Proposition 11.17. *Suppose that the graph H is obtained by taking two disjoint copies of a graph G and identifying two corresponding vertices, one from each graph. Let τ^G_{couple} be the time for a coupling of two walkers on G to meet. Then there is a coupling of two walkers on H which has a coupling time τ^H_{couple} satisfying*

$$\max_{u,v \in H} \mathbf{E}_{u,v}(\tau^H_{\text{couple}}) \leq 16 \left[\max_{x,y \in G} \mathbf{E}_x(\tau^G_y) + \max_{x,y \in G} \mathbf{E}(\tau^G_{\text{couple}}) \right]. \tag{11.26}$$

(Here τ^G_y is the hitting time of y in the graph G.)

Proof. Let $v^\star$ be the one vertex shared by the two copies of G. We couple two walkers, labeled A and B, started at vertices u and v in H. Initially, let the two walks move independently.

Denote the hitting time of y by walker A by τ^A_y and define the event $N_1 := \{\tau^A_{v^\star} > 2t_1\}$, where $t_1 := \max_{x,y \in G} \mathbf{E}_x(\tau_y)$. By Markov's inequality, $\mathbf{P}_{u,v}(N_1) \leq 1/2$.

At time $\tau^A_{v^\star}$, couple together A and B in the *projected space* identifying the two graphs, according to the original coupling in the graph G. Let τ^P_{couple} be the time until the particles couple in the projected space. The distribution of τ^P_{couple}

is the same as the distribution of τ^G_{couple}. Letting $t_2 = \max_{x,y \in G} \mathbf{E}_{x,y}(\tau^G_{\text{couple}})$ and $N_2 = \{\tau^P_{\text{couple}} > 2t_1\}$, we have

$$\max_{v \in H} \mathbf{P}_{v^\star, v}(N_2) \le \frac{1}{2}.$$

Finally, let N_3 be the event that when the particles couple in the projected space, the actual particles are in different copies of G. By symmetry, $\max_{u,v \in H} \mathbf{P}_{u,v}(N_3) \le 1/2$. We conclude that $\max_{u,v \in H} \mathbf{P}\{N_1^c \cap N_2^c \cap N_3^c\} \ge 1/8$. In other words, with probability at least $1/8$, the two particles couple by time $2(t_1 + t_2)$. To finish, apply Exercise 6.2. ■

{Exercise:TwoSST}

Exercise 11.2. Suppose that τ is a strong stationary time for simple random walk started at the vertex v on the graph G. Let H consist of two copies G_1 and G_2 of G, glued at v. Note that $\deg_H(v) = 2\deg_G(v)$. Let τ_v be the hitting time of v:

$$\tau_v = \min\{t \ge 0 \ : \ X_t = v\}.$$

Show that starting from any vertex x in H, the random time $\tau_v + \tau$ is a strong stationary time for H (where τ is applied to the walk after it hits v.)

Remark 11.3. It is also instructive to give a general direct argument controlling mixing time in the graph H described in Exercise 11.2:

Let $h_{\max}$ be the maximum expected hitting time of v in G, maximized over starting vertices. For $t > 2kh_{\max} + t_{\text{mix}G}(\varepsilon)$ we have in H that

$$|P^t(x, A) - \pi(A)| < 2^{-k} + \varepsilon. \qquad (11.27)$$ {Eq:DirectSt1}

Indeed for all x in H, we have $\mathbf{P}_x\{\tau_v > 2h_{\max}\} < 1/2$ and iterating, $\mathbf{P}_x\{\tau_v > 2kh_{\max}\} < (1/2)^k$. On the other hand, conditioning on $\tau_v < 2kh_{\max}$, the bound (11.27) follows from considering the projected walk.

We can now return to the example mentioned in this chapter's introduction:

{Cor:ConvergenceTwoTori

Corollary 11.18. *Consider the lazy random walker on two tori glued at a single vertex. (See Example 8.2 and in particular Figure 8.2.) There are constants c_1, c_2 such that*

$$c_1 n^2 \log n \le t_{\text{mix}} \le c_2 n^2 \log n, \qquad (11.28)$$ {Eq:TwoToriMix}

where t_{mix} is the mixing time defined in (5.33).

Proof of upper bound in (11.28). Applying Proposition 11.17, using the bounds in Proposition 11.9 and the bound (6.11) for the coupling on the torus used in Theorem 6.4 shows that there is a coupling with

$$\max_{x,y \in G} \mathbf{E}_{x,y}(\tau_{\text{couple}}) \le C_1 n^2 \log n. \qquad (11.29)$$

Applying Theorem 6.2 shows that

$$\bar{d}(t) \le \frac{C_1 n^2 \log n}{t},$$

proving the right-hand inequality in (11.28). ■

11.8. Cover Times

Herb Wilf, in the *American Mathematical Monthly*, offers the following account of waiting for a random walk to visit every pixel of his first personal computer's screen:

> *For a while, near the start of such a program, the walk is almost always quickly visiting pixels that it hasn't visited before, so one sees an irregular pattern that grows in the center of the screen. After a while, though, the walk will more often visit pixels that have previously been visited. Since they have already been lit up, and once they are lit up they are never turned off, the viewer sees no change on the screen.*
>
> *Hence there are periods when the screen seems frozen, and then suddenly the walk will visit some new pixel in another corner of the pattern, and more of them will be lit up.*
>
> *After quite a long while, when the screen is perhaps 95% illuminated, the growth process will have slowed down tremendously, and the viewer can safely go read* War and Peace *without missing any action. After a minor eternity, every cell will have been visited, the screen will be white, and the game will be over. Any mathematician who watched this will want to know how long, on average, it will take before, for the first time, all pixels have been visited.* (Wilf, 1989).

Let (X_t) be a finite Markov chain with state space Ω. The *cover time* C of (X_t) is the first time at which all the states have been visited. More formally, C is the minimal value such that, for every state $x \in \Omega$, there exists $t \le C$ with $X_t = x$.

The cover time of a Markov chain is a natural concept. As Wilf (1989) observed (quoted above), it can be large enough for relatively small chains to arouse mathematical curiosity. Of course, there are also "practical" interpretations of the cover time. For instance, we might view the progress of a web crawler as a random walk on the graph of World Wide Web pages: at each step, the crawler chooses a linked page at random and goes there. The time taken to scan the entire collection of pages is the cover time of the underlying graph.

{Xmpl:covercycle}

Example 11.19. Lovász (1993) gives an elegant computation of the expected cover time of simple random walk on the n-cycle. This walk is simply the remainder modulo n of a simple random walk on $\mathbb{Z}$. The walk on the remainders has covered all n states exactly when the walk on $\mathbb{Z}$ has first visited n distinct states.

Let c_n be the expected value of the time when a simple random walk on $\mathbb{Z}$ has first visited n states, and consider a walk which has just reached its $(n-1)$-st new state. The visited states form a subsegment of the number line and the walk must be at one end of that segment. Reaching the n-th new state is now a gambler's ruin situation: the walker's position corresponds to a fortune of 1 (or $n-1$), and we are waiting for her to reach either 0 or n. Either way, the expected time is

$(1)(n-1) = n-1$, as shown in Exercise 4.1. It follows that

$$c_n = c_{n-1} + (n-1) \quad \text{for} \quad n \geq 1.$$

Since $c_1 = 0$ (the first state visited is $X_0 = 0$), we may conclude that $c_n = n(n-1)/2$.

11.9. The Matthews method

It is not surprising that there is an essentially monotone relationship between hitting times and cover times: the longer it takes to travel between states, the longer it should take to visit all of them. Of course, a walk covering all the states can visit them in many different orders. This indeterminacy can be exploited: randomizing the order in which we check whether states have been visited (which, following Aldous and Fill (in progress), we will call the Matthews method—see Matthews (1988a) for the original version) allows us to prove both upper and lower bounds on expected cover times. Despite the simplicity of the arguments, these bounds are often remarkably good.

{th:covertime}

Theorem 11.20 (Matthews (1988a)). *Let (X_t) be an irreducible finite Markov chain on n states. Then, for any initial state x,*

$$\mathbf{E}_x(C) \leq \left[\max_{a,b} \mathbf{E}_a(\tau_b)\right]\left[1 + \frac{1}{2} + \cdots + \frac{1}{n}\right].$$

Proof. Without loss of generality, we may assume that our state space is $\{1, \ldots, n\}$. Let $\sigma \in S_n$ be a uniform random permutation, chosen independently of the chain; we will look for states in order σ. Let T_k be the first time that the states $\sigma(1), \ldots, \sigma(k)$ have all been visited, and let $L_k = X_{T_k}$ be the last state among $\sigma(1), \ldots, \sigma(k)$ to be visited.

Of course, when $\sigma(1) = x$, we have $T_1 = 0$. We will not usually be so lucky. In general,

$$\begin{aligned} \mathbf{E}_x(T_1 \mid \sigma(1) = s_1) &= \mathbf{E}_x(\tau_{s_1}) \\ &\leq \max_{a,b} \mathbf{E}_a(\tau_b). \end{aligned}$$

By Exercise 11.3, immediately below, $\mathbf{E}_x(T_1) \leq \max_{a,b} \mathbf{E}_a(\tau_b)$.

How much further along is T_2 than T_1?

- When the chain visits $\sigma(1)$ before $\sigma(2)$, then $T_2 - T_1$ is the time required to travel from $\sigma(1)$ to $\sigma(2)$, and $L_2 = \sigma(2)$.
- When the chain visits $\sigma(2)$ before $\sigma(1)$, we have $T_2 - T_1 = 0$ and $L_2 = \sigma(1)$.

Let's analyze the first case a little more closely. For any two distinct states $r, s \in \Omega$, define the event

$$A_2(r, s) = \{\sigma(1) = r, \sigma(2) = L_2 = s\}.$$

Clearly

$$\begin{aligned} \mathbf{E}_x(T_2 - T_1 \mid A_2(r, s)) &= \mathbf{E}_r(\tau_s) \\ &\leq \max_{a,b} \mathbf{E}_a(\tau_b). \end{aligned}$$

Conveniently,

$$A_2 = \bigcup_{r \neq s} A_2(r, s)$$

is simply the event that $\sigma(2)$ is visited after $\sigma(1)$, that is, $L_2 = \sigma(2)$. By Exercise 11.3,

$$\mathbf{E}_x(T_2 - T_1 \mid A_2) \leq \max_{a,b} \mathbf{E}_a(\tau_b).$$

Just as conveniently, A_2^c is the event that $\sigma(2)$ is visited before $\sigma(1)$. It immediately follows that

$$\mathbf{E}_x(T_2 - T_1 \mid A_2^c) = 0.$$

Since σ was chosen uniformly and independently of the chain trajectory, it is equally likely for the chain to visit $\sigma(2)$ before $\sigma(1)$, or after $\sigma(1)$. Thus

$$\begin{aligned}\mathbf{E}_x(T_2 - T_1) &= \mathbf{P}_x(A_2)\mathbf{E}_x(T_2 - T_1 \mid A_2) + \mathbf{P}_x(A_2^c)\mathbf{E}_x(T_2 - T_1 \mid A_2^c)\\ &\leq \frac{1}{2} \max_{a,b} \mathbf{E}_a(\tau_b).\end{aligned}$$

We can estimate $T_k - T_{k-1}$ for $3 \leq k \leq n$ in the same fashion; here, we carefully track whether $L_k = \sigma(k)$ or not. For any distinct $r, s \in \Omega$, define

$$A_k(r, s) = \{\sigma(k-1) = r, \sigma(k) = L_k = s\},$$

so that

$$\mathbf{E}_x(T_k - T_{k-1} \mid A_k(r, s)) = \mathbf{E}_r(\tau_s) \leq \max_{a,b} \mathbf{E}_a(\tau_b)$$

and

$$A_k = \bigcup_{r \neq s} A_k(r, s)$$

is the event that $L_k = \sigma(k)$. Just as above, exercise 11.3 implies that

$$\mathbf{E}_x(T_k - T_{k-1} \mid A_k) \leq \max_{a,b} \mathbf{E}_a(\tau_b),$$

while

$$\mathbf{E}_x(T_k - T_{k-1} \mid A_k^c) = 0.$$

As in the $k = 2$ case, independence and symmetry ensure that each of $\sigma(1), \dots, \sigma(k)$ is equally likely to be the last visited. Thus $\mathbf{P}_x(A_k) = 1/k$ and

$$\begin{aligned}\mathbf{E}_x(T_k - T_{k-1}) &= \mathbf{P}_x(A_k)\mathbf{E}_x(T_k - T_{k-1} \mid A_k) + \mathbf{P}_x(A_k^c)\mathbf{E}_x(T_k - T_{k-1} \mid A_k^c)\\ &\leq \frac{1}{k} \max_{a,b} \mathbf{E}_a(\tau_b).\end{aligned}$$

Finally, summing all these estimates yields

$$\begin{aligned}\mathbf{E}_x(C) &= \mathbf{E}_x(T_n)\\ &= \mathbf{E}_x(T_1) + \mathbf{E}_x(T_2 - T_1) + \cdots + \mathbf{E}_x(T_n - T_{n-1})\\ &\leq \max_{a,b} \mathbf{E}_a(\tau_b)\left(1 + \frac{1}{2} + \cdots + \frac{1}{n}\right).\end{aligned}$$

■

condprobpartitionbound}

Exercise 11.3. Let Y be a random variable on some probability space, and let $B = \cup_j B_j$ be a partition of an event B into (finitely or countably many) disjoint subevents B_j.

(a) Prove that when $\mathbf{E}(Y \mid B_j) \leq M$ for every j, then $\mathbf{E}(Y \mid B) \leq M$.

(b) Give an example to show that the conclusion of part (a) can fail when the events B_j are not disjoint.

[Solution]

Example 11.21. The proof above strongly parallels the standard argument for the coupon collecting problem, which we discussed in Section 4.2 and have applied several times: for instance, coupon collector bounds were used to lower bound mixing times for both random walk on the hypercube (Proposition 8.8) and Glauber dynamics on the graph with no edges (Exercise 8.4). For random walk on a complete graph with self-loops, the cover time coincides with the time to "collect" all coupons. In this case $\mathbf{E}_\alpha(\tau_\beta) = n$ is constant for $\alpha \neq \beta$, so the upper bound is tight.

A slight modification of this technique can be used to prove lower bounds: instead of looking for every state along the way to the cover time, we look for the elements of some subset of Ω. As long as the elements of the subset are far away from each other, in the sense that the hitting time between any two of them is large, we can get a good lower bound on the cover time.

{ex:Matthewslowerbound}

Exercise 11.4. For $A \subset X$ let C_A denote the first time such that every state of A has been visited. Let $\tau^A_{\min} = \min_{a,b \in A, a \neq b} \mathbf{E}_a(\tau_b)$.

(a) Show that for any state $x \in A$,

$$\mathbf{E}_x(C_A) \geq \tau^A_{\min}\left(1 + \frac{1}{2} + \cdots + \frac{1}{|A-1|}\right).$$

(Hint: begin by considering a uniform random permutation σ of the elements of A, and be careful when estimating the time to get to its first state.)

(b) Conclude that

$$\mathbf{E}_x(C) \geq \max_{x \in A \subseteq \Omega} \tau^A_{\min}\left(1 + \frac{1}{2} + \cdots + \frac{1}{|A-1|}\right).$$

[Solution]

Remark. While any subset A yields a lower bound, some choices for A are uninformative. For example, when the underlying graph of (Y_t) contains a leaf, $\tau^A_{\min} = 1$ for any set A containing both the leaf and its (unique) neighbor.

{Xmpl:TorusCover}

Example 11.22. In Section 11.4 we derived fairly sharp (up to constants) estimates for the hitting times of simple random walks on finite tori of various dimensions. Let's use these bounds and the Matthews method to determine equally sharp bounds on the expected cover times of tori. Since Wilf (1989) (quoted at the beginning of this chapter) allowed his random walker to wrap around the edges of his slowly-whitening computer screen, the resulting random walk took place on a discrete 2-torus. Below we provide a fairly precise answer to his question. However, we discuss the case of dimension at least 3 first, since the details are a bit simpler.

When the dimension $d > 3$, Proposition 11.9 tells us that there exist constants c_d and C_d such that for any distinct vertices x, y of $\mathbb{Z}_n^d$,

$$c_d n^d \leq \mathbf{E}_x(\tau_y) \leq C_d n^d$$

Remarkably, this bound does not depend on the distance between x and y! By Theorem 11.20, the average cover time satisfies

{eq:bigtoruscoverupper}

$$\begin{aligned} \mathbf{E}C &\leq C_d n^d \left(1 + \frac{1}{2} + \cdots + \frac{1}{n^d}\right) \qquad (11.30) \\ &= C_d d n^d \log n(1 + o(1)). \qquad (11.31) \end{aligned}$$

To derive an almost-matching lower bound out from Exercise 11.4, we must choose a set A large enough that the sum of reciprocals in the second factor is substantial. Let's take A to be $\mathbb{Z}_n^d$ itself (any set containing a fixed positive fraction of the points of the torus would work as well). Then

$$\begin{aligned} \mathbf{E}C &\geq \tau_{\min}^A \left(1 + \frac{1}{2} + \cdots + \frac{1}{|A-1|}\right) \\ &\geq c_d d n^d \log n(1 + o(1)), \end{aligned}$$

which is only a constant factor away from our upper bound.

In dimension 2, Proposition 11.9 says that when x and y are vertices of $\mathbb{Z}_n^2$ at distance k,

$$c_2 n^2 \log(k) \leq \mathbf{E}_x(\tau_y) \leq C_2 n^2 \log(k).$$

In this case the Matthews upper bound gives

{Eq:2dtoruscoverupper}

$$\mathbf{E}(C) \leq 2C_2 n^2 (\log n)^2 (1 + o(1)), \tag{11.32}$$

since the furthest apart two points can be is n.

To get a good lower bound, we must choose a set A which is as large as possible, but for which the minimum distance between points is also large. Assume for simplicity that n is a perfect square, and let A be the set of all points in Z_d^2 both of whose coordinates are multiples of $\sqrt{n}$. Then Exercise 11.4 and Proposition 11.9 imply

$$\begin{aligned} \mathbf{E}(C) &\geq c_2 n^2 \log(\sqrt{n}) \left(1 + \frac{1}{2} + \dots \frac{1}{n-1}\right) \\ &= \frac{c_2}{2} n^2 (\log n)^2 (1 + o(1)). \end{aligned}$$

Exercises 11.23 and 11.24 use improved estimates on the hitting times to get our upper and lower bounds for cover times on tori even closer together. The exact asymptotics of the expected cover time on Z_n^2 have only recently been determined.

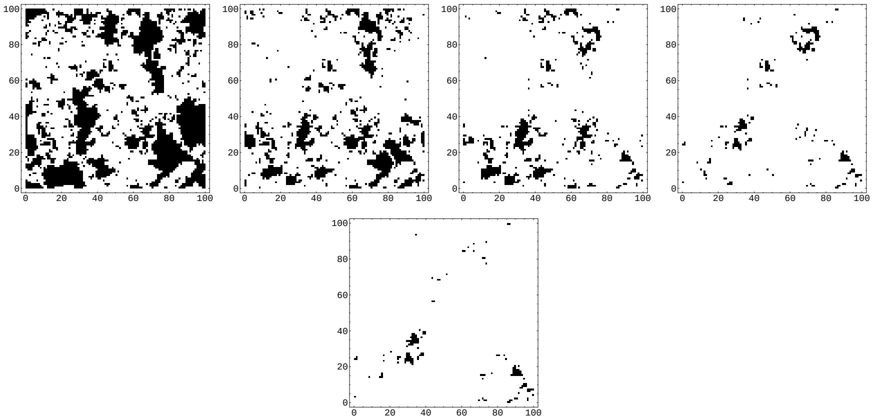


Fig:CoverTorus

Figure 11.4. Black squares show the unvisited states in a single trajectory of simple random walk on a 100×100 torus, after 54004, 108008, 162012, 216016, and 270020 steps, respectively.

Zuckerman (1992) was the first to estimate the expected cover time to within a constant, while Dembo et al. (2004) show that

$$\mathbf{E}(C) \sim \frac{4}{\pi} n^2 (\log n)^2.$$

Figure 11.4 shows the points of a 100×100 torus left uncovered by a single random walk trajectory at approximately 20%, 40%, 60%, 80%, and 100% of this time.

11.10. Problems

{Exercise:Patterns}

Exercise 11.5. Consider the problem of waiting for sequence TTT to appear in a sequence of fair coin tosses. Is this the same as the waiting time for the sequence HTH?

These waiting times are hitting times for a Markov chain: let X_t be the triplet consisting of the outcomes of tosses $(t, t+1, t+2)$. Then (X_t) is a Markov chain, and the waiting time for TTT is a hitting time. Find $\mathbf{E}(\tau_{TTT})$ and $\mathbf{E}(\tau_{HTH})$.

{Exer:UnequalHit}

Exercise 11.6. Let G be a connected graph on at least 3 vertices in which the vertex v has only one neighbor, namely w. Show that in for the simple random walk on G, $\mathbf{E}_v \tau_w \neq \mathbf{E}_w \tau_v$.

{Exer:CycleMixHit}

Exercise 11.7. Compute $\mathbf{E}_\pi \tau_0$ for random walk (lazy or not) on the cycle C_n, and apply Theorem 11.11 directly to bound $t_{\rm mix}$ for this walk. How much does this improve on the results of Example 11.14 (which relied upon (11.22))?

{Exercise:TorusIsTransi

Exercise 11.8. Check that the torus $\mathbb{Z}_n^d$ is transitive.

{Exercise:cubecount}

Exercise 11.9.

(a) Show that in the m-dimensional hypercube there are exactly $m2^{m-1}$ edges.

(b) Show that there are $k\binom{m}{k}$ edges that connect a node with Hamming weight $k-1$ to a node with Hamming weight k. (The Hamming weight is the sum of the coordinates.)

[SOLUTION]

{Exercise:cubehit}

EXERCISE 11.10. Let $\mathbf{0} = (0, 0, \ldots, 0)$ be the all zero vector in the m-dimensional hypercube $\{0, 1\}^m$, and let v be a vertex with Hamming weight k. Write $h_m(k)$ for the expected hitting time from v to $\mathbf{0}$ for simple (that is, not lazy) random walk on the hypercube. Determine $h_m(1)$ and $h_m(m)$. Deduce that both $\min_{k>0} h_m(k)$ and $\max_{k>0} h_m(k)$ are asymptotic to 2^m as m tends to infinity. (We say that $f(m)$ is asymptotic to $g(m)$ if their ratio tends to 1.)

Hint: Consider the multigraph G_m obtained by gluing together all vertices of Hamming weight k for each k between 1 and $m-1$. This is a graph on the vertex set $\{0, 1, 2, \ldots, m\}$ with $k\binom{m}{k}$ edges from $k-1$ to k. [SOLUTION]

Exercise:TwoHypercubes}

EXERCISE 11.11. Use Proposition 11.17 to bound the mixing time for two hypercubes identified at a single vertex.

{Exercise:taubca}

EXERCISE 11.12. Let (X_t) be a random walk on a network with conductances $\{c_e\}$. Show that

$$\mathbf{E}_a(\tau_{bca}) = [\mathcal{R}(a \leftrightarrow b) + \mathcal{R}(b \leftrightarrow c) + \mathcal{R}(c \leftrightarrow a)] \sum_{e \in E} c_e,$$

where τ_{bca} is the first time that the sequence (b, c, a) appears as a subsequence of $(X_1, X_2, \ldots)$. [SOLUTION]

{Exercise:HitStatesEx}

EXERCISE 11.13. Show that for a random walk (X_t) on a network, for every three vertices a, x, z,

$$\mathbf{P}_x\{\tau_z < \tau_a\} = \frac{\mathcal{R}(a \leftrightarrow x) - \mathcal{R}(x \leftrightarrow z) + \mathcal{R}(a \leftrightarrow z)}{2\mathcal{R}(a \leftrightarrow z)}.$$

Hint: Run the chain from x until it first visits a and then z. This will also be the first visit to z from x, unless $\tau_z < \tau_a$. In the latter case the path from x to a to z involves an extra commute from z to a beyond time τ_z. Thus, starting from x we have

{Eq:HTI1}

$$\tau_{az} = \tau_z + \mathbf{1}_{\{\tau_z < \tau_a\}} \tau'_{az}, \tag{11.33}$$

where the variable τ'_{az} refers to the chain starting from its first visit to z. Now take expectations and use the cycle identity (Lemma 11.8). [SOLUTION]

Exer:NodeCycleAltProof}

EXERCISE 11.14. Let θ be a flow from a to z which satisfies both the cycle law and $\|\theta\| = \|I\|$. Define a function h on nodes by

$$h(x) = \sum_{i=1}^{m} [\theta(\vec{e}_i) - I(\vec{e}_i)]\, r(\vec{e}_i), \tag{11.34}$$

where $\vec{e}_i, \ldots, \vec{e}_m$ is an arbitrary path from a to x.

(a) Show that h is well-defined and harmonic at all nodes.

(b) Use part (a) to give an alternate proof of Proposition 10.2.

{Exercise:ConvGF}

Exercise 11.15. Suppose that $\{a_k\}$ is sequence with generating function $A(s) := \sum_{k=0}^{\infty} a_k s^k$, and $\{b_k\}$ is a sequence with generating function $B(s) := \sum_{k=0}^{\infty} b_k s^l$. Let $\{c_k\}$ be the sequence defined as $c_k := \sum_{j=0}^{k} a_j b_{k-j}$, called the *convolution* of $\{a_k\}$ and $\{b_k\}$. Show that the generating function of $\{c_k\}$ equals $A(s)B(s)$. [Solution]

{Exercise:Even}

Exercise 11.16.

(i) Let $\tau_x^{\sharp}$ denote the first even time that the Markov chain visits x. Prove that the inequality

$$t_{\text{mix}}(1/4) \le 8 \max_{x \in \Omega} \mathbf{E}_\pi \left(\tau_x^{\sharp} \right) + 1$$

holds without assuming the chain is lazy (cf. Theorem 11.11).

(ii) Prove an analog of (11.21) for continuous time chains without assuming laziness.

{Exercise:CM1}

Exercise 11.17. Show that $t_{\text{mix}}^{\star}(1/4) \le 6 t_{\text{mix}}(1/8)$.

{Exercise:CM2}

Exercise 11.18. Show that $t_{\text{mix}}^{\star}(2^{-k}) \le k t_{\text{mix}}^{\star}(1/4)$ for all $k \ge 1$.

{Exercise:CM3}

Exercise 11.19. Consider a lazy biased random walk on the n-cycle. That is, at each time $t \ge 1$, the particle walks one step clockwise with probability $p \in (1/4, 1/2)$, stays put with probability $1/2$, and walks one step counter-clockwise with probability $1/2 - p$.

Show that $t_{\text{mix}}(1/4)$ is bounded above and below by constant multiples of n^2, but $t_{\text{mix}}^{\star}(1/4)$ is bounded above and below by constant multiples of n.

{Exer:LastVertexCycle}

Exercise 11.20. For a graph G, let W be the (random) vertex visited at the cover time for the simple random walker on G. That is, W is the last new vertex to be visited by the random walk. Prove the following remarkable fact: for random walk on an n-cycle, W is uniformly distributed over all vertices different from the starting vertex.

Hint: Exercise 10.5, on further aspects of the gambler's ruin problem, may be helpful.

Remark 11.4. Let W be the random vertex defined in Exercise 11.20. Lovász and Winkler (1993) demonstrate that cycles and complete graphs are the only graphs for which W is this close to uniformly distributed. More precisely, these families are the only ones for which W is equally likely to be any vertex other than the starting state.

{Exer:CycleCover}

Exercise 11.21. What upper and lower bounds does the Matthews method give for cycle $\mathbb{Z}_n$? Compare to the actual value, computed in Example 11.19, and explain why the Matthews method gives a poor result for this family of chains.

{Exer:HypercubeCover}

Exercise 11.22. Show that the cover time of the m-dimensional hypercube is asymptotic to $m 2^m \log(2)$ as $m \to \infty$.

{Exer:covertorussharp}

Exercise 11.23. In this exercise, we demonstrate that for tori of dimension $d \ge 3$, just a little more information on the hitting times suffices to prove a matching lower bound.

(a) Show that when a sequence of pairs of points $x_n, y_n \in \mathbb{Z}_n^d$ has the property that the distance between them tends to infinity with n, then the upper-bound constant C_d of (11.17) can be chosen so that $\mathbf{E}_{x_n}(\tau_{y_n})/n^d \to C_d$.

(b) Give a lower bound on $\mathbf{E}C$ that has the same initial constant as the upper bound of (11.30).

Exer:cover2Dtorussharp}

Exercise 11.24. Following the example of Exercise 11.23, derive a lower bound for $\mathbf{E}(C)$ on the two-dimensional torus that is within a factor of 4 of the upper bound (11.32).

Notes

[compare results for $t_{\text{mix}} = t_{\text{mix}}(1/4)$ on the cycle from Example 11.14 and Exercise 11.7 to actual asymptotic constant?]

For much more on waiting times for patterns in coin tossing, see Li (1980).

The mean commute identity appears in Chandra, Raghavan, Ruzzo, Smolensky, and Tiwari (1996/97).

A graph similar to our glued tori was analyzed in Saloff-Coste (1997, Section 3.2) using other methods. This graph originated in Diaconis and Saloff-Coste (1996, Remark 6.1).

CHAPTER 12

Eigenvalues

{Ch:Eigenvalues}

In this chapter we assume, unless stated otherwise, that the transition matrix P is reversible with respect to the stationary measure π (recall the definition (3.27)), aperiodic, and irreducible.

12.1. The Spectral Representation of a Transition Matrix

We begin by collecting some facts about the eigenvalues of transition matrices:

{Exercise:Spectrum}

Exercise 12.1.

(a) Show that for *any* transition matrix P (not necessarily reversible, irreducible, or aperiodic), all eigenvalues λ satisfy $|\lambda| \leq 1$.
Hint: Letting $\|f\|_\infty := \max_{x \in \Omega} |f(x)|$, show that $\|Pf\|_\infty \leq \|f\|_\infty$. Apply this with the eigenfunction φ corresponding to the eigenvalue λ.

(b) Suppose P is irreducible and aperiodic. Show that -1 is not an eigenvalue, and that the vector space of eigenfunctions corresponding to the eigenvalue 1 is all scalar multiples of the vector $\mathbf{1} := (1, 1, \ldots, 1)$.
Hint: Check directly or use the Convergence Theorem.

Exercise 12.1 shows that 1 is always an eigenvalue and the remaining $n-1$ eigenvalues lie in the interval $(1, -1)$. We label the eigenvalues in decreasing order:

$$1 = \lambda_1 > \lambda_2 \geq \cdots \geq \lambda_{|\Omega|} > -1. \qquad (12.1)$$ {Eq:OrderEigen}

Define

$$\lambda_\star := \max\{|\lambda| : \lambda \text{ is an eigenvalue of } P,\ \lambda \neq 1\}. \qquad (12.2)$$ {Eq:Lambda2Defn}

The difference $\gamma_\star := 1 - \lambda_\star$ is called the *absolute spectral gap*; Exercise 12.1 shows that $\gamma_\star$ is strictly positive.

If at each move, the chain holds its current position with probability at least $1/2$, then $\gamma_\star = 1 - \lambda_2$:

{Exercise:Lazy}

Exercise 12.2. Show that if $\tilde{P} = (1/2)P + (1/2)I$, where I is the identity matrix, then all eigenvalues of $\tilde{P}$ are non-negative. This is the *lazy* version of P: at each move, depending on the outcome of a fair coin toss, the chain either transitions according to P or remains in its current state.

Denote by $\langle \cdot, \cdot \rangle$ the usual inner product on $\mathbb{R}^{|\Omega|}$, given by $\langle f, g \rangle = \sum_{x \in \Omega} f(x) g(x)$. We will need a different inner product, denoted by $\langle \cdot, \cdot \rangle_\pi$ and defined as

$$\langle f, g \rangle_\pi := \sum_{x \in \Omega} f(x) g(x) \pi(x). \qquad (12.3)$$ {Eq:DefInnerProd}

The reason for introducing this new inner product is:

{Lem:SpectralDecomp}

Lemma 12.1. *The inner-product space* $(\mathbb{R}^{|\Omega|}, \langle \cdot, \cdot \rangle_\pi)$ *has an orthonormal basis* $\{f_j\}$ *of eigenfunctions of* P *so that*

{Eq:SpecDec}

$$\frac{P^t(x,y)}{\pi(y)} = \sum_{j=1}^{|\Omega|} f_j(x) f_j(y) \lambda_j^t, \tag{12.4}$$

where $1 = \lambda_1 > \lambda_2 \geq \cdots \geq \lambda_{|\Omega|} > -1$ *are the eigenvalues of* P*. The eigenfunction* f_1 *is taken to be the constant vector* $\mathbf{1}$*.*

Proof. Define $A(x,y) := \pi(x)^{1/2}\pi(y)^{-1/2}P(x,y)$. Reversibility of P implies that A is symmetric. The Spectral Theorem (Theorem B.6) guarantees that the inner product space $(\mathbb{R}^{|\Omega|}, \langle \cdot, \cdot \rangle)$ has an orthonormal basis $\{\varphi_j\}_{j=1}^{|\Omega|}$ of eigenfunctions of A. We write $\{\lambda_j\}$ for the eigenvalues of A.

The reader should directly check that $\sqrt{\pi}$ is an eigenfunction of A with corresponding eigenvalue 1; we set $\varphi_1 := \sqrt{\pi}$ and $\lambda_1 := 1$.

Letting Π be the diagonal matrix with diagonal entries $\Pi(x,x) = \pi(x)$, by definition $A = \Pi^{\frac{1}{2}} P \Pi^{-\frac{1}{2}}$. If $f_j := \Pi^{-\frac{1}{2}}\varphi_j$, then f_j is an eigenfunction of P with eigenvalue λ_j:

$$Pf_j = P\Pi^{-\frac{1}{2}}\varphi_j = \Pi^{-\frac{1}{2}}(\Pi^{\frac{1}{2}}P\Pi^{-\frac{1}{2}})\varphi_j = \Pi^{-\frac{1}{2}}A\varphi_j = \Pi^{-\frac{1}{2}}\lambda_j\varphi_j = \lambda_j f_j.$$

Although these eigenfunctions are not necessarily orthonormal with respect to the usual inner product, they are orthonormal with respect to $\langle \cdot, \cdot \rangle_\pi$ defined in (12.3):

{Eq:NewOrth}

$$\delta_{ij} = \langle \varphi_i, \varphi_j \rangle = \langle \Pi^{\frac{1}{2}} f_i, \Pi^{\frac{1}{2}} f_j \rangle = \langle f_i, f_j \rangle_\pi. \tag{12.5}$$

(The first equality follows since $\{\varphi_j\}$ is orthonormal with respect to the usual inner product.)

Let δ_y be the function

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

Considering $(\mathbb{R}^{|\Omega|}, \langle \cdot, \cdot \rangle_\pi)$ with its orthonormal basis of eigenfunctions $\{f_j\}_{j=1}^{|\Omega|}$, the function δ_y can be written via basis decomposition as

{Eq:DeltaDecomp}

$$\delta_y = \sum_{j=1}^{|\Omega|} \langle \delta_y, f_j \rangle_\pi f_j = \sum_{j=1}^{|\Omega|} f_j(y)\pi(y) f_j. \tag{12.6}$$

Since $P^t f_j = \lambda_j^t f_j$ and $P^t(x,y) = (P^t\delta_y)(x)$,

$$P^t(x,y) = \sum_{j=1}^{|\Omega|} f_j(y)\pi(y)\lambda_j^t f_j(x).$$

Dividing by $\pi(y)$ completes the Lemma. ■

12.2. Spectral Representation of Simple Random Walks

The simple random walk on the n-cycle was introduced in Example 3.2. We discuss here the eigenfunctions and eigenvalues for this chain, along with the random walk on the interval.

The *nth roots of unity* are the complex numbers z which solve the equation $z^n = 1$. There are n such solutions, given by $\omega_k = \exp(i2\pi k/n)$ for $k = 0, 1, 2, \ldots, n-1$. Geometrically, these are the points in the complex plane which lie on the unit circle with angles $2\pi k/n$.

Observe that

$$\omega_k \omega_j = \exp(i2\pi (j \oplus k)/n) = \omega_{k \oplus j},$$

where $j \oplus k := (j+k) \mod n$. Thus, the set $\{\omega_0, \ldots, \omega_{n-1}\}$ together with complex multiplication is a group isomorphic to the group $\mathbb{Z}_n$ of integers $\{0, 1, 2, \ldots, n-1\}$ with the operation of addition modulo n.

{Sec:RWCSpectral}

12.2.1. The cycle.

The simple random walk on the cycle can be realized as a random walk on the nth roots of unity, where at each step the current position is multiplied by a random choice from $\{\omega_1, \omega_1^{-1}\}$.

A (possibly complex-valued) eigenfunction f satisfies

$$\lambda f(\omega_k) = Pf(\omega_k) = \frac{f(\omega_{k \ominus 1}) + f(\omega_{k \oplus 1})}{2}$$

for all ω_k.

For $j = 0, 1, 2, \ldots, n-1$, define $f_j(\omega_k) := \omega_k^j = \omega_{jk}$, where the multiplication jk is modulo n. Then

$$Pf_j(\omega_k) = \frac{f_j(\omega_{k \oplus 1}) + f_j(\omega_{k \ominus 1})}{2} = \frac{\omega_{jk \oplus j} + \omega_{jk \ominus j}}{2} \qquad (12.7)$$ {Eq:CircAve}

For any ℓ and j, the average of the vectors $\omega_{\ell \ominus j}$ and $\omega_{\ell \oplus j}$ is a scalar multiple of ω_ℓ; this is illustrated on the left-hand side of Figure 12.1 for $j = 1$. Note that the cord connecting $\omega_{\ell \oplus j}$ with $\omega_{\ell \ominus j}$ intersects ω_ℓ at a right angle, so the projection of $\omega_{\ell \oplus j}$ onto ω_ℓ has length $\cos(2\pi j/n)$. In view of this,

$$Pf_j(\omega_k) = \cos(2\pi j/n)\omega_{jk} = \cos(2\pi j/n) f_j(\omega_k).$$

In other words, f_j is an eigenfunction with eigenvalue $\lambda_j = \cos(2\pi j/n)$.

Because f_j is an eigenvector with a real eigenvalue λ_j, both its real part and its imaginary parts are (real-valued) eigenfunctions. In particular,

$$\mathrm{Re}(f_j(\omega_k)) = \mathrm{Re}(e^{i2\pi jk/n}) = \cos(2\pi jk/n)$$

is an eigenfunction.

{Sec:PathEigs}

12.2.2. Lumped chains and the path.

Consider the random walk on the $(4n)$th roots of unity $\{\omega_k\}_{k=1}^{4n}$, where at each move the current position is multiplied by a random element of $\{\omega_2, \omega_2^{-1}\}$. Suppose the walker is started at ω_{2k_0+1} for some k_0. The state-space for this chain is $\{\omega_{2k+1}\}_{k=0}^{2n-1}$, of size $2n$.

Denote by $\bar{z}$ the complex conjugate of z: if $z = x + iy$, then $\bar{z} := x - iy$. If the states ω_{2k+1} and $\bar{\omega}_{2k+1}$ are identified with each other for $k = 0, 1, \ldots, n-1$, then resulting chain is a random walk on the interval $\{0, 1, \ldots, n-1\}$ with holding at the

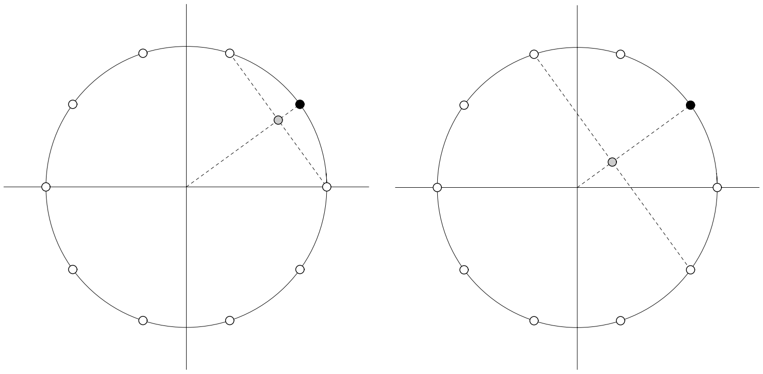


Fig:RWCycEig

FIGURE 12.1. The eigenvalues must be the cosines.

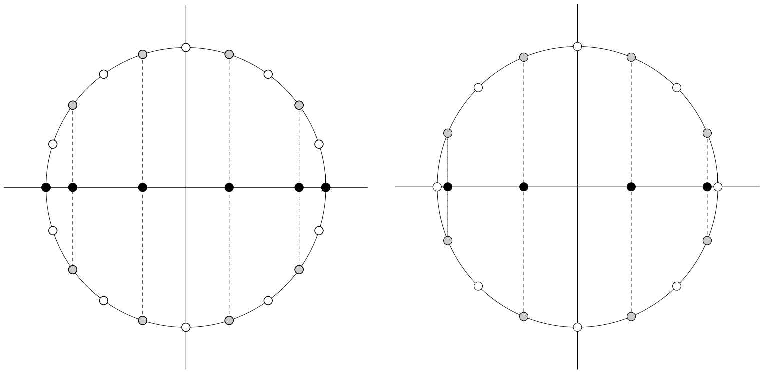


Fig:RWPathEig

FIGURE 12.2. Random walks on cycles project to random walks on paths. On the left, the walk reflects at the end points. On the right, it holds with probability 1/2.

end points. That is, when the walk is at 0, it moves to 1 with probability $1/2$ and stays at 0 with probability $1/2$, and when the walk is at $n-1$, it moves to $n-2$ with probability $1/2$ and stays at $n-1$ with probability $1/2$.

Consider for $j = 0, 1, \ldots, 2n-1$ the function ϕ_j defined by

$$\phi_j(\omega_k) := \omega_k^j = \exp\left(i\frac{\pi}{2n}jk\right), \quad k = 1, 3, \ldots, 4n-1.$$

Now let $\oplus$ and $\ominus$ denote addition and subtraction modulo $4n$. Then

$$\begin{aligned} P\phi_j(\omega_k) &= \frac{\phi_j(\omega_{k\oplus 2}) + \phi_j(\omega_{k\ominus 2})}{2} \\ &= \frac{\exp\left[i\frac{\pi}{2n}\left(jk \oplus 2j\right)\right] + \exp\left[i\frac{\pi}{2n}\left(jk \ominus 2j\right)\right]}{2} \\ &= \lambda_j\phi_j(\omega_k), \end{aligned}$$

where λ_j is the projection of the unit vector with angle $\frac{\pi}{2n}(jk + 2j)$ onto the unit vector with angle $\frac{\pi}{2n}jk$. Indeed, for $j = 0, 1, \dots, 2n - 1$,

$$\lambda_j = \cos\left(\frac{\pi j}{n}\right).$$

Since λ_j is real, the real part f_j of ϕ_j is a real eigenfunction. Using the identities $\mathrm{Re}(z) = (z + \bar{z})/2$ and $\overline{z^j} = \bar{z}^j$,

$$f_j(\omega_k) = \frac{1}{2}\left[\omega_k^j + \bar{\omega}_k^j\right] = \cos\left(\frac{\pi jk}{2n}\right). \tag{12.8}$$ {Eq:RePartEig}

We return now to the random walk on the path, obtained by identifying the two states ω_{2k+1} and $\bar{\omega}_{2k+1}$ with $k + 1$ for $1 \leq k \leq n$. We first give a general lemma on projecting a Markov chain onto equivalence classes.

{Lem:EquivChain}

Lemma 12.2. *Let Ω be a the state-space of a Markov chain (X_t) with transition matrix P. Let $\sim$ on Ω be an equivalence relation on X with equivalence classes $\Omega' = \{[x] : x \in \Omega\}$, and assume that the measures $P(x, \cdot)$ and $P(x', \cdot)$ satisfy*

$$P(x, [y]) = P(x', [y]) \tag{12.9}$$ {Eq:Lump}

whenever $x \sim x'$. Then:

(i) *$[X_t]$ is a Markov chain with transition matrix $P'([x], [y]) = P(x, [y])$.*

(ii) *Let $f : \Omega \to \mathbb{R}$ be an eigenfunction of P with eigenvalue λ which is constant on equivalence classes. Then the natural projection $f' : \Omega' \to R$ of f,defined by $f'([x]) = f(x)$, is an eigenfunction of P' with eigenvalue λ.*

(iii) *Conversely, if $f' : \Omega' \to \mathbb{R}$ is an eigenfunction of P' with eigenvalue λ, then its lift $f : \Omega \to \mathbb{R}$, defined by $f(x) = f'([x])$, is an eigenvector of P with eigenvalue λ.*

Proof. The first assertion is an immediate consequence of (12.9). For the second, we can simply compute:

$$\begin{aligned}(P'f')([x]) &= \sum_{[y]\in\Omega'} P'([x],[y])f'([y]) = \sum_{[y]\in\Omega'} P(x,[y])f(y) \\ &= \sum_{[y]\in\Omega'}\sum_{z\in[y]} P(x,z)f(z) = \sum_{z\in\Omega} P(x,z)f(z) = (Pf)(x) = \lambda f(x) = \lambda f([x]).\end{aligned}$$

To prove the third assertion, just run the same computations in reverse:

$$\begin{aligned}(Pf)(x) &= \sum_{z\in\Omega} P(x,z)f(z) = \sum_{[y]\in\Omega'}\sum_{z\in[y]} P(x,z)f(z) = \sum_{[y]\in\Omega'} P(x,[y])f(y) \\ &= \sum_{[y]\in\Omega'} P'([x],[y])f'([y]) = (P'f')([x]) = \lambda f'([x]) = \lambda f(x).\end{aligned}$$

■

Remark. The process of constructing a new chain by taking equivalence classes for an equivalence relation compatible with the transition matrix is sometimes called *lumping*.

Returning to the example of the path: it is clear from (12.8) that the eigenfunction f_j is constant on equivalence classes when ω_k and $\bar{\omega}_k$ are identified. By part (ii) of Lemma 12.2 it thus becomes an eigenfunction for the lumped chain. If we identify the pair $\{\omega_{2k+1}, \omega_{\overline{2k}+1}\}$ with the integer $k+1$, we get a chain with state space $[n] = \{1, 2, \dots, n\}$. Its eigenfunctions are given by

{Eq:PathEvecs}
$$k \mapsto \cos\left(\frac{\pi j(2k-1)}{2n}\right), \tag{12.10}$$

which has eigenvalue $2\pi j/2n = \pi j/n$, for $j = 0, \dots, n-1$.

{Sec:ProductChains}

12.3. Product chains

For each $j = 1, 2, \dots, d$, let P_j be a transition matrix on the state-space Ω_j. Consider the chain on $\Omega_1 \times \Omega_2 \cdots \times \Omega_d$ which moves by selecting a coordinate at each step and moving only in the chosen coordinate according to the corresponding transition matrix. The transition matrix $\tilde{P}$ for this chain is

{Eq:ProdMatrix}
$$\tilde{P}((x_1, \dots, x_j, \dots x_d), (x_1, \dots, y_j, \dots, x_d)) = \frac{P_j(x_j, y_j)}{d}. \tag{12.11}$$

See Exercise 12.6 for a different product chain.

If f_j is a function on Ω_j for each $j = 1, 2, \dots, d$, the *tensor product* of $\{f_j\}_{j=1}^d$ is the function on $\Omega_1 \times \cdots \times \Omega_d$ defined by

$$(f_1 \otimes f_2 \otimes \cdots \otimes f_d)(x_1, \dots, x_d) := f_1(x_1) f_2(x_2) \cdots f_d(x_d).$$

{Lem:ProdChain}

LEMMA 12.3. *Suppose that for each $j = 1, 2, \dots, d$, the transition matrix P_j on state-space Ω_j has eigenfunction ϕ_j with eigenvalue λ_j. Then $\tilde{\phi} := \phi_1 \otimes \cdots \otimes \phi_d$ is an eigenfunction of the transition matrix $\tilde{P}$ defined in (12.11), with eigenvalue $d^{-1} \sum_{j=1}^d \lambda_j$.*

PROOF. Lift P_j from Ω_j to $\Omega_1 \times \cdots \times \Omega_d$ by defining $\tilde{P}_j$ by

$$\tilde{P}_j((x_1, \dots, x_j, \dots, x_d), (x_1, \dots, y_j, \dots, x_d)) = P_j(x_j, y_j).$$

This corresponds to the chain on $\Omega_1 \times \cdots \times \Omega_d$ which makes moves in the jth coordinate according to P_j.

Letting $\boldsymbol{x} = (x_1, \dots, x_d)$, it is simple to check that

$$\tilde{P}_j \tilde{\phi}(\boldsymbol{x}) = \lambda_j \tilde{\phi}(\boldsymbol{x}).$$

From this and noting that $\tilde{P} = d^{-1} \sum_{j=1}^d \tilde{P}_j$ it follows that

$$\tilde{P}\tilde{\phi}(\boldsymbol{x}) = d^{-1} \sum_{j=1}^d \tilde{P}_j \tilde{\phi}(\boldsymbol{x}) = \left[d^{-1} \sum_{j=1}^d \lambda_j \right] \tilde{\phi}(\boldsymbol{x}).$$

■

{Example:EigenHC}

EXAMPLE 12.4 (Random walk on n-dimensional hypercube). Consider the chain (X_t) on $\Omega := \{-1, 1\}$ which is an i.i.d. sequence of random signs. That is, the transition matrix is

{Eq:TwoState}
$$P(x, y) = \frac{1}{2} \quad \text{for all } x, y \in \{-1, 1\}. \tag{12.12}$$

Let $I_1(x) = x$, and note that

$$PI_1(x) = \frac{1}{2} + \frac{-1}{2} = 0.$$

Thus there are two eigenfunction: I_1 (with eigenvalue 0), and **1**, the constant function (with eigenvalue 1).

Consider the lazy random walker on the n-dimensional hypercube, but for convenience write the state-space as $\{-1, 1\}^n$. In this state-space, the chain moves by selecting a coordinate uniformly at random and refreshing the chosen coordinate with a new random sign, independent of everything else. The transition matrix is exactly (12.11), where each P_j is the two-state transition matrix in (12.12).

By Lemma 12.3, the eigenfunctions are of the form

$$f(x_1, \ldots, x_k) = \prod_{j=1}^{k} f_j(x_j)$$

where f_j is either I_1 or $\mathbf{1}$. In other words, for each subset of coordinates $J \subset \{1, 2, \ldots, k\}$,

$$f_J(x_1, \ldots, x_k) := \prod_{j \in J} x_j$$

is an eigenfunction. The corresponding eigenvalue is

$$\lambda_J = \frac{\sum_{i=1}^{k}(1 - \mathbf{1}_{\{i \in J\}})}{k} = \frac{k - |J|}{k}.$$

We take $f_\varnothing(\boldsymbol{x}) := 1$, which is the eigenfunction corresponding to the eigenvalue 1.

12.4. The Relaxation Time

{Sec:Gap}

The *relaxation time* t_{rel} is defined as $\gamma_\star^{-1}$, where $\gamma_\star$ is the absolute spectral gap $1 - \max_{j \geq 1} |\lambda_j|$. The connection between the relaxation time and mixing times is the following:

{Thm:MixRelax}

THEOREM 12.5. *Let t_{rel} be the relaxation time $1/\gamma_\star$ for a reversible, irreducible Markov chain, and let $\pi_{\min} := \min_{x \in \Omega} \pi(x)$. Then*

$$t_{\mathrm{mix}}(\varepsilon) \leq -\log(\varepsilon \pi_{\min}) t_{\mathrm{rel}}. \tag{12.13}$$

{Eq:MixingLambdaTwo}

PROOF OF THEOREM 12.5. By Lemma 12.1, since $f_1 = \mathbf{1}$,

$$\frac{P^t(x, y)}{\pi(y)} = 1 + \sum_{j=2}^{|\Omega|} f_j(x) f_j(y) \lambda_j^t.$$

By the Cauchy-Schwarz inequality,

$$\left| \frac{P^t(x, y)}{\pi(y)} - 1 \right| \leq \sum_{j=2}^{|\Omega|} |f_j(x) f_j(y)| \lambda_\star^t \leq \lambda_\star^t \left[\sum_{j=2}^{|\Omega|} f_j^2(x) \sum_{j=2}^{|\Omega|} f_j^2(y) \right]^{1/2}. \tag{12.14}$$

{Eq:WeighedAver1}

Using (12.6) and the orthonormality of $\{f_j\}$ shows that

$$\pi(x) = \langle \delta_x, \delta_x \rangle_\pi = \left\langle \sum_{j=1}^{|\Omega|} f_j(x)\pi(x) f_j, \sum_{j=1}^{|\Omega|} f_j(x)\pi(x) f_j \right\rangle_\pi = \pi(x)^2 \sum_{j=1}^{|\Omega|} f_j(x)^2.$$

Consequently, $\sum_{j=2}^{|\Omega|} f_j(x)^2 \leq \pi(x)^{-1}$. This together with (12.14) proves that

$$\left| \frac{P^t(x,y)}{\pi(y)} - 1 \right| \leq \frac{\lambda_\star^t}{\sqrt{\pi(x)\pi(y)}} \leq \frac{\lambda_\star^t}{\pi_{\min}} = \frac{(1-\gamma_\star)^t}{\pi_{\min}} \leq \frac{e^{-\gamma_\star t}}{\pi_{\min}}.$$

Applying Lemma 7.5 shows that $d(t) \leq \pi_{\min}^{-1} \exp(-\gamma_\star t)$. The conclusion now follows from the definition of $t_{\mathrm{mix}}(\varepsilon)$. ∎

Example 12.6 (Random walk on n-dimensional hypercube). The eigenvalues for the lazy random walk on the hypercube $\{0,1\}^n$ were computed in Example 12.4. (We used the more convenient state-space $\{-1,1\}^n$, but the eigenvalues are the same.)

In particular, the eigenfunction $f_{\{1,\dots,n\}}$ has eigenvalue 0 and the eigenfunction $f_\varnothing$ has $\lambda_1 = 1$. Each f_J with $|J| = 1$ has $\lambda_2 = 1 - 1/n$, and consequently $\gamma_\star = 1/n$.

Theorem 12.5 gives

$$t_{\mathrm{mix}}(\varepsilon) \leq n\left(-\log\varepsilon + \log(2^n)\right) = n^2\left(\log 2 - n^{-1}\log\varepsilon\right) = O(n^2).$$

Note that this bound is not as good as the bound obtained previously in Section 7.4.2.

{Thm:LBSGMix}

Theorem 12.7. *For a reversible and irreducible Markov chain*

$$t_{\mathrm{mix}}(\varepsilon) \geq (t_{\mathrm{rel}} - 1)\log\left(\frac{1}{2\varepsilon}\right).$$

In particular,

$$t_{\mathrm{mix}} \geq \frac{\log 2}{2} t_{\mathrm{rel}}.$$

{Rmk:SG}

Remark 12.1. If $\gamma_\star$ is small because the smallest eigenvalue $\lambda_{|\Omega|}$ is near -1, the slow mixing suggested by this lower bound can be rectified by passing to a lazy chain to make the eigenvalues positive. For such lazy chains, $\gamma_\star = \gamma$, where $\gamma := 1 - \lambda_2$. (See Exercise 12.2.) If γ is near 0, then the mixing may be very slow indeed. Therefore, we are mainly concerned with γ, not $\gamma_\star$.

Proof. Suppose that f is an eigenfunction of P with eigenvalue $\lambda \neq 1$, so that $Pf = \lambda f$. Note that since the eigenfunctions are orthogonal with respect to $\langle \cdot, \cdot \rangle_\pi$, and $\mathbf{1}$ is an eigenfunction,

$$\sum_{y \in \Omega} \pi(y) f(y) = \langle \mathbf{1}, f \rangle_\pi = 0.$$

Then

$$|\lambda^t f(x)| = |P^t f(x)| = \left| \sum_{y\in\Omega} \left[P^t(x,y) f(y) - \pi(y) f(y) \right] \right| \leq \|f\|_\infty 2d(t).$$

With this inequality, we can obtain a lower bound on the mixing time. Taking x with $|f(x)| = \|f\|_\infty$ yields $|\lambda|^{t_{\mathrm{mix}}(\varepsilon)} \leq 2\varepsilon$, and so

$$t_{\mathrm{mix}}(\varepsilon)\left(\frac{1}{|\lambda|} - 1\right) \geq t_{\mathrm{mix}}(\varepsilon)\log\left(\frac{1}{|\lambda|}\right) \geq \log\left(\frac{1}{2\varepsilon}\right).$$

Minimizing the left-hand side over eigenvalues different from 1 and rearranging finishes the proof. ■

12.5. Bounds on Spectral Gap via Contractions

Suppose that Ω is a metric space with distance ρ.

{Thm:Contraction}

THEOREM 12.8 (M.F. Chen (1998)). *Let P be a transition matrix for a Markov chain, not necessarily reversible. Suppose there exists a constant $\theta < 1$ and for each $x, y \in \Omega$ there is a coupling (X_1, Y_1) of $P(x, \cdot)$ and $P(y, \cdot)$ satisfying*

$$\mathbf{E}_{x,y}\left(\rho(X_1, Y_1)\right) \leq \theta\rho(x, y). \tag{12.15}$$ {Eq:ContrHyp}

If λ is an eigenvalue of P different from 1, then $|\lambda| \leq \theta$. In particular, the absolute spectral gap satisfies

$$\gamma_\star \geq 1 - \theta.$$

The *lipschitz constant* of a function f on a metric space (Ω, ρ) is defined as

$$\mathrm{lip}(f) := \max_{\substack{x,y\in\Omega \\ x\neq y}} \frac{|f(x) - f(y)|}{\rho(x, y)}.$$

PROOF. For any function f,

$$|Pf(x) - Pf(y)| = \left|\mathbf{E}_{x,y}\left(f(X_1) - f(Y_1)\right)\right| \leq \mathbf{E}_{x,y}\left(|f(X_1) - f(Y_1)|\right).$$

By the definition of lip(f) and the hypothesis (12.15),

$$|Pf(x) - Pf(y)| \leq \mathrm{lip}(f)\mathbf{E}_{x,y}\left(d(X_1, Y_1)\right) \leq \theta\,\mathrm{lip}(f)d(x, y).$$

This proves that

$$\mathrm{lip}(Pf) \leq \theta\,\mathrm{lip}(f).$$

Taking ϕ to be a non-constant eigenfunction with eigenvalue λ,

$$|\lambda|\,\mathrm{lip}(\phi) = \mathrm{lip}(\lambda\phi) = \mathrm{lip}(P\phi) \leq \theta\,\mathrm{lip}(\phi).$$

■

EXAMPLE 12.9. Consider again the lazy random walker on the hypercube $\{0, 1\}^n$, taking the metric to be the Hamming distance $\rho(x, y) = \sum_{i=1}^d |x_i - y_i|$.

Let (X_1, Y_1) be the coupling which updates the same coordinate in both chains. The distance decreases by one if one among the $\rho(x, y)$ disagreeing coordinates is selected, and otherwise remains the same. Thus,

$$\mathbf{E}_{x,y}\left(\rho(X_1, Y_1)\right) \leq \left(1 - \frac{\rho(x, y)}{n}\right)\rho(x, y) + \frac{\rho(x, y)}{n}(\rho(x, y) - 1) = \left(1 - \frac{1}{n}\right)\rho(x, y).$$

Applying Theorem 12.8 yields the bound $\gamma_\star \geq n^{-1}$. In Example 12.4 it was shown that $\gamma_\star = n^{-1}$, so the bound of Theorem 12.8 is sharp in this case.

12.6. An ℓ^2 Bound and Cut-Off for the Hypercube

{Sec:L2CutOff}

For each $p \geq 0$, the $\ell^p(\pi)$ norm is defined as

$$\|f\|_p := \left[\sum_{x\in\Omega} |f(x)|^p \pi(x)\right]^{1/p}.$$

An important case is $\ell = 2$, as $\ell^2(\pi)$ is the inner-product space with $\|f\|_2 = \sqrt{\langle f, f\rangle_\pi}$.

{Exercise:ellp} Exercise 12.3. Show that the function $p \mapsto \|f\|_p$ is non-decreasing for $p \geq 1$.

{Lem:L2Bound} Lemma 12.10. *Let P be a reversible transition matrix, with eigenvalues*

$$1 = \lambda_1 > \lambda_2 \geq \cdots \geq \lambda_{|\Omega|} \geq -1,$$

{It:L2Bound} *and associated eigenfunctions $\{f_j\}$, orthonormal with respect to $\langle\cdot,\cdot\rangle_\pi$. Then*

(i)

$$4\|P^t(x,\cdot) - \pi\|_{TV}^2 \leq \left\| \frac{P^t(x,\cdot)}{\pi(\cdot)} - 1 \right\|_2^2 = \sum_{j=2}^{|\Omega|} f_j(x)^2 \lambda_j^{2t}.$$

{It:L2Trans} (ii) *If the chain is transitive, then*

$$4\|P^t(x,\cdot) - \pi\|_{TV}^2 \leq \left\| \frac{P^t(x,\cdot)}{\pi(\cdot)} - 1 \right\|_2^2 = \sum_{j=2}^{|\Omega|} \lambda_j^{2t}.$$

Proof. By Lemma 12.1,

{Eq:UsingSpDe}
$$\left\| \frac{P^t(x,\cdot)}{\pi(\cdot)} - 1 \right\|_2^2 = \left\| \sum_{j=2}^{|\Omega|} \lambda_j^t f_j(x) f_j \right\|_2^2 = \sum_{j=2}^{|\Omega|} f_j(x)^2 \lambda_j^{2t}. \tag{12.16}$$

By Exercise 12.3,

{Eq:TVandL2}
$$4\|P^t(x,\cdot) - \pi\|_{TV}^2 = \left\| \frac{P^t(x,\cdot)}{\pi(\cdot)} - 1 \right\|_1^2 \leq \left\| \frac{P^t(x,\cdot)}{\pi(\cdot)} - 1 \right\|_2^2, \tag{12.17}$$

which with (12.16) establishes (i).

Suppose the Markov chain is transitive. Then π is uniform, and the left-hand side of (12.16) does not depend on x. Summing over x,

$$|\Omega| \left\| \frac{P^t(x,\cdot)}{\pi(\cdot)} - 1 \right\|_2^2 = |\Omega| \sum_{j=2}^{|\Omega|} \left[\sum_{x\in\Omega} f_j(x)^2 \pi(x) \right] \lambda_j^{2t},$$

where we have multiplied and divided by $\pi(x) = 1/|\Omega|$ on the right-hand side. Since $\|f_j\|_2 = 1$, the inner sum on the right-hand side equals 1, and so

$$\left\| \frac{P^t(x,\cdot)}{\pi(\cdot)} - 1 \right\|_2^2 = \sum_{j=2} \lambda_j^{2t}.$$

Combining with (12.17) establishes (ii). ■

{Xmpl:HCcutoff}

EXAMPLE 12.11. For lazy simple random walk on the hypercube $\{0, 1\}^n$, the eigenvalues and eigenfunctions were found in Example 12.4. This chain is transitive, so applying Lemma 12.10 shows that

$$4\|P^t(x,\cdot) - \pi\|_{TV}^2 \leq \sum_{k=1}^{n} \left(1 - \frac{k}{m}\right)^{2t} \binom{n}{k} \leq \sum_{k=1}^{n} e^{-2tk/n}\binom{n}{k} = \left(1 + e^{-2t/n}\right)^n - 1.$$

Taking $t = (1/2)n\log n + cn$,

$$4\|P^t(x,\cdot) - \pi\|_{TV}^2 \leq \left(1 + \frac{1}{n}e^{-2c}\right)^n - 1 \leq e^{e^{-2c}} - 1.$$

On the other hand, the argument in Proposition 8.8 shows that

$$d((1/2)n\log n - cn) \geq 1 - \frac{8}{e^{2c}}\left[1 + o(1)\right].$$

Thus we see that $d(t)$ exhibits a sharp *cut-off* at $(1/2)n\log n$.

Suppose that for each $n \in \mathbb{Z}^+$, there is a transition matrix P_n on state-space Ω_n with stationary distribution π_n. Define

$$d_n(t) := \max_{x\in\Omega_n} \left\|P_n^t(x,\cdot) - \pi_n\right\|_{\mathrm{TV}}$$

We say this family of Markov chains has a *cut-off* at $\{t_n\}$ with *window* w_n if $w_n = o(t_n)$ and for any sequence $\{\tilde{w}_n\}$ with $\tilde{w}_n/w_n \to \infty$,

$$\lim_{n\to\infty} d_n(t_n - \tilde{w}_n) = 1,$$

and

$$\lim_{n\to\infty} d_n(t_n + \tilde{w}_n) = 0.$$

LEMMA 12.12. *A family of Markov chains has a cut-off if and only if*

$$\lim_{n\to\infty} \frac{t_{\mathrm{mix}}^{(n)}(\varepsilon)}{t_{\mathrm{mix}}^{(n)}(1-\varepsilon)} = 1.$$

PROOF. ■

12.7. Wilson's method and random adjacent transpositions

{Sec:WilsonRAT}

The lower bound we present is due to David Wilson (see Wilson (2004)). Theorem 12.13 is the key. In its proof an eigenfunction Φ of a chain is used to construct a distinguishing statistic; Proposition 8.5 then bounds the distance from stationarity.

{Thm:WilsonLower}

THEOREM 12.13. *Let (X_t) be an irreducible aperiodic Markov chain with state space Ω and transition matrix P. Let Φ be an eigenvector of P with eigenvalue λ satisfying $1/2 < \lambda < 1$. Fix $0 < \varepsilon < 1$ and let $R > 0$ satisfy*

$$\mathbf{E}_y\left|\Phi(X_1) - \Phi(y)\right|^2 \leq R \qquad (12.18)$$

{Eq:DefR}

for all $y \in \Omega$. Then for any $x \in \Omega$

$$t < \frac{\log \frac{(1-\varepsilon)(1-\lambda)\Phi(x)^2}{2\varepsilon R}}{2\log(1/\lambda)} \tag{12.19}$$

{Eq:WilsonLower}

implies

$$\left\|P^t(x,\cdot) - \pi\right\|_{TV} \geq \varepsilon.$$

At first glance, Theorem 12.13 appears daunting! Yet it gives sharp lower bounds in many important examples. Let's take a closer look, and work through an example, before proceeding with the proof.

Remark. In applications, ε may not be tiny. For instance, when proving a family of chains has a cutoff, we will need to consider all values $0 < \varepsilon < 1$.

Remark. Generally λ will be taken to be the second largest eigenvalue in situations where $\gamma_\star = \gamma = 1-\lambda$ is small. Under these circumstances a one-term Taylor expansion yields

{Eq:WilsonDiscussion}

$$\frac{1}{\log(1/\lambda)} = \frac{1}{\gamma_\star + O(\gamma_\star)^2} = t_{\mathrm{rel}}(1 + O(\gamma_\star)). \tag{12.20}$$

According to Theorems 12.5 and 12.7,

$$\log\left(\frac{1}{2\varepsilon}\right)(t_{\mathrm{rel}} - 1) \leq t_{\mathrm{mix}}(\varepsilon) \leq -\log(\varepsilon\pi_{\mathrm{min}})t_{\mathrm{rel}},$$

where $\pi_{\mathrm{min}} = \min_{x\in\Omega}\pi(x)$. One way to interpret (12.20) is that the denominator of (12.19) gets us up to the relaxation time (ignoring constants, for the moment). The numerator, which depends on the geometry of Φ, determines how much larger a lower bound we can get.

Remark. Note that multiplying Φ by a scalar c multiplies the minimum possible value of the bound R by a factor of c^2. Hence the numerator of (12.19) is invariant under multiplication of Φ by a scalar.

{Xmpl:HypercubeWilson}

Example 12.14. Recall from Example 12.4 that the second-largest eigenvalue of the lazy random walk on the n-dimensional hypercube $\{0,1\}^n$ is $1 - \frac{1}{n}$. The corresponding eigenspace has dimension n, but a convenient representative to take is

$$\Phi(\boldsymbol{x}) = W(\boldsymbol{x}) - \frac{n}{2},$$

where $W(\boldsymbol{x})$ is the Hamming weight (i.e. the number of 1's) in the bitstring $\boldsymbol{x}$. For any bitstring $\boldsymbol{y}$, we have

$$\mathbf{E}_{\boldsymbol{y}}((\Phi(X_1) - \Phi(\boldsymbol{y}))^2) = \frac{1}{2}(1) + \frac{1}{2}(0) = \frac{1}{2},$$

since the value changes by exactly 1 whenever the walker actually moves. Now apply Theorem 12.13, taking the initial state to be the all-ones vector $\mathbf{1}$ and $R = 1/2$. We get

$$t_{\mathrm{mix}}(\varepsilon) \geq \frac{\log \frac{(1-\varepsilon)(1/n)(n/2)^2}{2\varepsilon(1/2)}}{2\log(1/(1-1/n))} = \frac{n\log n}{2} - \log\left(\frac{4\varepsilon}{1-\varepsilon}\right)n + O(\log n).$$

In Example 12.11, we showed that this family of chains has a sharp cutoff at $(1/2)n \log n$. The argument, given in Proposition 8.8 and using the Hamming weight directly as a distinguishing statistic, was actually quite similar; the major difference is that we used the structure of the hypercube walk to bound the variances. Wilson's method can be seen as a natural (in hindsight!) extension of that argument. What makes Theorem 12.13 widely applicable is the simple form of its implicit bound on the variance.

Proof of Theorem 12.13. Since

$$\mathbf{E}(\Phi(X_{t+1})|X_t = z) = \lambda\Phi(z) \qquad (12.21) \quad \texttt{\{Eq:Evec\}}$$

for all $t \geq 0$ and $z \in \Omega$, we have

$$\mathbf{E}_x\Phi(X_t) = \lambda^t\Phi(x) \quad \text{for } t \geq 0 \qquad (12.22) \quad \texttt{\{Eq:ExpectedPhi\}}$$

by induction. Fix a value t, let $z = X_t$, and define $\Delta = \Phi(X_{t+1}) - \Phi(z)$. By (12.21) and (12.18), respectively, we have

$$\mathbf{E}_x(\Delta|X_t = z) = (\lambda - 1)\Phi(z)$$

and

$$\mathbf{E}_x(\Delta^2|X_t = z) \leq R.$$

Hence

$$\begin{aligned}\mathbf{E}_x(\Phi(X_{t+1})^2|X_t = z) &= \mathbf{E}_x((\Phi(z) + \Delta)^2|X_t = z)\\ &= \Phi(z)^2 + 2\mathbf{E}_x(\Delta\Phi(z)|X_t = z) + \mathbf{E}_x(\Delta^2|X_t = z)\\ &\leq (2\lambda - 1)\Phi(z)^2 + R.\end{aligned}$$

Averaging over the possible values of $z \in \Omega$ with weights $P^t(x, z) = \mathbf{P}_x(X_t = z)$ gives

$$\mathbf{E}_x\Phi(X_{t+1})^2 \leq (2\lambda - 1)\mathbf{E}_x\Phi(X_t)^2 + R.$$

At this point, we could apply this estimate inductively, then sum the resulting geometric series. It is equivalent (and neater) to subtract $R/(2(1 - \lambda))$ from both sides, obtaining

$$\mathbf{E}_x\Phi(X_{t+1})^2 - \frac{R}{2(1-\lambda)} \leq (2\lambda - 1)\left(\mathbf{E}_x\Phi(X_t)^2 - \frac{R}{2(1-\lambda)}\right),$$

from which it is clear that

$$\mathbf{E}_x\Phi(X_t)^2 \leq (2\lambda - 1)^t\Phi(x) + \frac{R}{2(1-\lambda)}. \qquad (12.23) \quad \texttt{\{Eq:ExpectedPhiSquared\}}$$

Combining (12.22) and (12.23) gives

$$\mathrm{Var}_x\Phi(X_t) \leq \left[(2\lambda - 1)^t - \lambda^{2t}\right]\Phi(x)^2 + \frac{R}{2(1-\lambda)} < \frac{R}{2(1-\lambda)}, \qquad (12.24) \quad \texttt{\{Eq:WilsonVar\}}$$

since $2\lambda - 1 < \lambda^2$ ensures the the first term is negative.

Now, let $X_\infty \in \Omega$ have distribution π and let $t \to \infty$ in (12.22). Then Theorem 5.6 implies that $\mathbf{E}(\Phi(X_\infty)) = 0$ (as does the orthogonality of eigenvectors). Similarly, letting $t \to \infty$ in (12.24) gives

$$\mathrm{Var}_x \Phi(X_\infty) \leq \frac{R}{2(1-\lambda)}.$$

Applying Proposition 8.5 with $r^2 = \frac{2(1-\lambda)\lambda^{2t}\Phi(x)^2}{R}$ gives

$$\left\|P^t(x,\cdot) - \pi\right\|_{\mathrm{TV}} \geq \frac{r^2}{4+r^2} = \frac{(1-\lambda)\lambda^{2t}\Phi(x)^2}{2R + (1-\lambda)\lambda^{2t}\Phi(x)^2}.$$

When t satisfies (12.19), we have

$$(1-\lambda)\lambda^{2t}\Phi(x)^2 > \frac{\varepsilon}{1-\varepsilon}(2R)$$

and hence

$$\left\|P^t(x,\cdot) - \pi\right\|_{\mathrm{TV}} \geq \varepsilon.$$

■

Remark. The variance estimate of 12.24 may look crude, but only $O(\lambda^{2t})$ is being discarded. In applications this is generally quite small.

In order to apply Wilson's method to the random adjacent transpositions shuffle, we must specify an eigenvector and initial state.

First, some generalities on eigenvalues and eigenfunctions of shuffle chains. Let (σ_t) be a shuffle chain with state space $\mathcal{S}_n$ and shuffle distribution Q (that is, at each step a permutation is chosen according to Q and composed with σ_t to generate σ_{t+1}). Fix $k \in [n]$. Then Lemma 12.2(i) 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. In addition, Lemma 12.2(iii) tells us that when $\Phi' : [n] \to \mathbb{R}$ is an eigenfunction of the single-card chain with eigenvalue λ, then $\Phi : \mathcal{S}_n \to \mathbb{R}$ defined by $\Phi(\sigma) = \Phi'(\sigma(k))$ is an eigenfunction of the shuffle chain with eigenvalue λ.

For the random adjacent transpositions chain, the single-card chain is an extremely lazy version of a random walk on the path whose eigenvectors and eigenvalues were determined in Section 12.2.2. 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 (12.10) that

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

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 $\sigma \mapsto \phi(\sigma(k))$ is an eigenfunction of the random transposition walk with eigenvalue λ. Since these eigenfunctions all lie in the same eigenspace, so will any linear combination of them. We set

$$\Phi(\sigma) = \sum_{k\in[n]} \phi(k)\phi(\sigma(k)). \tag{12.25}$$ {Eq:EvecDef}

Remark. See Exercise 9.6 for some motivation on our choice of Φ. By making sure that $\Phi(\mathrm{id})$ is as large as possible, we ensure that when $\Phi(\sigma_t)$ is small, then σ_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 Φ. Only two terms in (12.25) change, and we compute:

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

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

$$|\Phi(\sigma(k-1\ k)) - \Phi(\sigma)| \leq \frac{\pi}{n}(2) = \frac{2\pi}{n}. \tag{12.26}$$ {Eq:ComputeR}

Combining (12.26) with Theorem 12.13 and the fact that $\Phi(\mathrm{id}) = n/2$ (see Exercise 9.7) tells us that when the random adjacent transposition shuffle is started with a sorted deck, after

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

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

12.8. Time Averages

{Lem:AveSecMom}

Lemma 12.15. *Let (X_t) be a reversible Markov chain, and f an eigenfunction of the transition matrix P with eigenvalue λ and with $\langle f, f\rangle_\pi = 1$. Then*

$$\mathbf{E}_\pi\left[\left(\sum_{s=0}^{t-1} f(X_s)\right)^2\right] \leq \frac{2t}{1-\lambda}. \tag{12.27}$$ {Eq:EigAve}

If f is any real-valued function defined on Ω, then

$$\mathbf{E}_\pi\left[\left(\sum_{s=0}^{t-1} f(X_s)\right)^2\right] \leq \frac{2tE_\pi(f^2)}{\gamma}. \tag{12.28}$$ {Eq:TimeAve}

Proof. For $r < s$,

$$\begin{aligned}\mathbf{E}_\pi\left[f(X_r)f(X_s)\right] &= \mathbf{E}_\pi\left[\,\mathbf{E}_\pi\left(f(X_r)f(X_s) \mid X_r\right)\right]\\ &= \mathbf{E}_\pi\left[f(X_r)\,\mathbf{E}_\pi\left(f(X_s)\mid X_r\right)\right] = \mathbf{E}_\pi\left[f(X_r)\,(P^{s-r}f)(X_r)\right].\end{aligned}$$

Since f is an eigenfunction and $E_\pi(f^2) = \langle f, f\rangle_\pi = 1$,

$$\mathbf{E}_\pi\left[f(X_r)f(X_s)\right] = \lambda^{s-r}\mathbf{E}_\pi\left[f(X_r)^2\right] = \lambda^{s-r}\mathbf{E}_\pi(f^2) = \lambda^{s-r}.$$

Then by considering separately the diagonal and cross terms when expanding the square,

$$\mathbf{E}_\pi\left[\left(\sum_{s=0}^{t-1} f(X_s)\right)^2\right] = t + 2\sum_{r=0}^{t-1}\sum_{s=1}^{t-1-r} \lambda^s.$$

Summing the geometric sum,

$$\mathbf{E}_\pi\left[\left(\sum_{s=0}^{t-1} f(X_s)\right)^2\right] = t + \frac{2t\lambda - (\lambda - \lambda^t)/(1-\lambda)}{1-\lambda}.$$

Equation (12.27) follows from the inequality $|(\lambda - \lambda^m)/(1-\lambda)| \leq 1$.

Let $\{f_j\}_{j=1}^{\Omega}$ be the orthonormal eigenfunctions of P of Lemma 12.1. Decompose a general f as $f = \sum_{j=1}^{|\Omega|} a_j f_j$. By Parseval's Identity, $E_\pi(f^2) = \sum_{j=1}^{n} a_j^2$.

Defining $G_j := \sum_{s=0}^{t-1} a_j f_j(X_s)$, we can write

$$\sum_{s=0}^{t-1} f(X_s) = \sum_{j=1}^{|\Omega|} a_j G_j$$

If $r \leq s$ then

$$\begin{aligned}
\mathbf{E}_\pi\left[f_j(X_s)f_k(X_r)\right] &= \mathbf{E}_\pi\left[f_k(X_r)\,\mathbf{E}_\pi(f_j(X_s) \mid X_r)\right] \\
&= \mathbf{E}_\pi\left[f_k(X_r)(P^{s-r}f_j)(X_r)\right] \\
&= \lambda_j^{s-r}\mathbf{E}_\pi\left[f_k(X_r)f_j(X_r)\right] \\
&= \lambda_j^{s-r}E_\pi(f_k f_j) \\
&= 0.
\end{aligned}$$

Consequently, $\mathbf{E}_\pi\left(G_j G_k\right) = 0$ for $j \neq k$. It follows that

$$\mathbf{E}_\pi\left[\left(\sum_{s=0}^{t-1} f(X_s)\right)^2\right] = \sum_{i=1}^{|\Omega|} a_i^2 \mathbf{E}_\pi\left(G_i^2\right). \tag{12.29}$$

By (12.27), the right-hand side is bounded by

$$\sum_{j=1}^{|\Omega|} \frac{2ta_j^2}{1-\lambda_j} \leq \frac{2tE_\pi(f^2)}{\gamma}.$$

■

Theorem 12.16. *Let (X_t) be an reversible Markov chain. If $r \geq t_{\mathrm{mix}}(\varepsilon/2)$ and $t \geq [4\operatorname{Var}_\pi(f)/(\eta^2\varepsilon)]t_{\mathrm{rel}}$, then for any starting state $x \in \Omega$,*

$$\mathbf{P}_x\left\{\left|\frac{1}{t}\sum_{s=0}^{t-1} f(X_{r+s}) - E_\pi(f)\right| \geq \eta\right\} \leq \varepsilon.$$

Proof. Assume without loss of generality that $E_\pi(f) = 0$; if not, replace f by $f - E_\pi(f)$.

Let p_r be the optimal coupling of $P^r(x, \cdot)$ with π so that

$$\sum_{x \neq y} p_r(x, y) = \left\| P^r(x, \cdot) - \pi \right\|_{\mathrm{TV}} .$$

Define a Markov chain $(Y_s, Z_s)_{s \geq 0}$ by starting (Y_0, Z_0) with p_r and using the transition matrix

$$Q((x, y), (z, w)) = \begin{cases} P(x, z) & \text{if } x = y \text{ and } z = w, \\ P(x, z)P(y, w) & \text{if } x \neq y, \\ 0 & \text{otherwise.} \end{cases}$$

The sequences (Y_s) and (Z_s) are each Markov chains with transition matrix P, started in state x and with π, respectively. The chains (Y_s) and (Z_s) move independently until they meet, after which they move together. Because the distribution of (Y_0, Z_0) is p_r,

$$\mathbf{P}\{Y_0 \neq Z_0\} = \left\| P^r(x, \cdot) - \pi \right\|_{\mathrm{TV}} .$$

Since $(Y_s)_{s \geq 0}$ and $(X_{r+s})_{r \geq 0}$ have the same distribution, we rewrite the probability in the statement as

$$\mathbf{P}_x \left\{ \left| \frac{1}{t} \sum_{s=0}^{t-1} f(X_{r+s}) - E_\pi(f) \right| > \eta \right\} = \mathbf{P} \left\{ \left| \frac{1}{t} \sum_{s=0}^{t-1} f(Y_s) - E_\pi(f) \right| > \eta \right\} .$$

By considering whether or not $Y_0 = Z_0$, this probability is bounded above by

$$\mathbf{P}\left\{X_0 \neq Z_0\right\} + \mathbf{P} \left\{ \left| \frac{1}{t} \sum_{s=0}^{t-1} f(Z_s) - E_\pi(f) \right| > \eta \right\} .$$

By definition of $t_{\mathrm{mix}}(\varepsilon)$, if $r \geq t_{\mathrm{mix}}(\varepsilon/2)$, then the first term is bounded by $\varepsilon/2$. We use Chebyshev on the second term along with Lemma 12.15 to show that if $t \geq 4\,\mathrm{Var}_\pi(f)/(\eta^2\varepsilon)t_{\mathrm{rel}}$ then the second term is bounded by $\varepsilon/2$. ■

12.9. Problems

{Exer:MonoSelfTran}

Exercise 12.4. Let P be a reversible transition matrix with stationary distribution π. Use Lemma 12.1 to prove that $P^{2t+2}(x, x) \leq P^{2t}(x, x)$.

{Exercise:RWInt}

Exercise 12.5. Consider the random walk on the interval $\{0, 1, \dots, n-1\}$ which moves with equal probability left and right when at the interior points, and has "inelastic" boundary behavior:

$$P(0, 1) = 1 \quad \text{and} \quad P(n-1, n-2) = 1.$$

By considering the simple random walk on the $(2n-2)$th roots of unity, find the eigenvalues and eigenfunctions for this chain.

Solution to 12.5. The simple random walk on the $(2n-2)$ roots of unity at each move multiplies by a random choice from $\{\omega_1, \omega_1^{-1}\}$. As shown in Section 12.2.1, the eigenvalues for this walk are

$$\lambda_j = \cos\left(\frac{\pi j}{n-1}\right).$$

When ω_k and $\bar{\omega}_k$ are identified, the walk on the interval with inelastic boundary conditions is obtained. ■

{Exercise:ProdChain}

Exercise 12.6. Let P_1 and P_2 by transition matrices on state-spaces Ω_1 and Ω_2 respectively. Consider the chain on $\Omega_1 \times \Omega_2$ which moves independently in the first and second coordinates according to P_1 and P_2 respectively. Its transition matrix is the *tensor product* $P_1 \otimes P_2$, defined as

$$P_1 \otimes P_2((x,y),(z,w)) = P_1(x,z)P_2(y,w).$$

The tensor product of a function ϕ on Ω_1 and a function ψ on Ω_2 is the function on $\Omega_1 \times X_2$ defined by $(\phi \otimes \psi)(x,y) = \phi(x)\psi(y)$.

Let ϕ and ψ be eigenfunctions of P_1 and P_2 respectively, with eigenvalues λ and μ. Show that $\phi \otimes \psi$ is an eigenfunction of $P_1 \otimes P_2$ with eigenvalue $\lambda\mu$.

12.10. Notes

The connection between the spectral gap of the Laplace-Beltrami operator on Riemannian manifolds and an isoperimetric constant is due to Cheeger (1970), hence the bottleneck ratio is often called the *Cheeger constant*. The relationship between the bottleneck ratio and the spectral gap for random walks on graphs was observed by Alon and Milman (1985) and further developed in Alon (1986). For general Markov chains this was independently exploited by Sinclair and Jerrum (1989) and Lawler and Sokal (1988).

Theorem 12.8 can be combined with Theorem 12.5 to get a bound on mixing time when there is a coupling which contracts, in the reversible case: If for each pair of states x, y, there exists a coupling (X_1, Y_1) of $P(x,\cdot)$ and $P(y,\cdot)$ satisfying

$$\mathbf{E}_{x,y}(\rho(X_1,Y_1)) \leq \theta\rho(x,y),$$

then

{Eq:BadBound}

$$t_{\text{mix}}(\varepsilon) \leq \frac{-\log(\varepsilon) - \log(\pi_{\min})}{1-\theta} \tag{12.30}$$

Compare with Corollary 14.3, which bounds mixing time directly from a contractive coupling. Since $\pi_{\min}\text{diam} \leq \pi_{\min}|\Omega| \leq 1$, it follows that $-\log(\pi_{\min}) \geq \log(\text{diam})$ and the bound in (12.30) is never better than the bound given by Corollary 14.3. In fact, (12.30) can be much worse. For example, for the hypercube, $\pi_{\min}^{-1} = 2^d$, while the diameter is d.

CHAPTER 13

The Variational Principle and Comparison of Chains

{Ch:VPC}

In this chapter, we will always assume that P is a reversible transition matrix with stationary distribution π.

13.1. The Dirichlet Form

The *Dirichlet form* associated to the pair (P, π) is defined for functions f and g on Ω by

$$\mathcal{E}(f, h) := \langle (I - P)f, h \rangle_\pi .$$

We write simply $\mathcal{E}(f)$ for $\mathcal{E}(f, f)$.

{Lem:DFAlt}

Lemma 13.1. *For a reversible transition matrix P with stationary distribution π,*

$$\mathcal{E}(f) = \frac{1}{2} \sum_{x,y \in \Omega} [f(x) - f(y)]^2 \, \pi(x) P(x, y). \tag{13.1}$$

{Eq:DirForm}

Proof. First write

$$\langle (I - P)f, f \rangle_\pi = \sum_{x \in \Omega} [f(x) - Pf(x)] \, f(x) \pi(x)$$

$$= \sum_{x \in \Omega} \left[f(x) - \sum_{y \in \Omega} f(y) P(x, y) \right] f(x) \pi(x).$$

Since $\sum_{y \in \Omega} P(x, y) = 1$, the right-hand side above equals

$$\sum_{x \in \Omega} \left[\sum_{y \in \Omega} P(x, y) f(x) - \sum_{y \in \Omega} f(y) P(x, y) \right] f(x) \pi(x).$$

Simplifying,

$$\mathcal{E}(f) = \sum_{x \in \Omega} \sum_{y \in \Omega} [f(x) - f(y)] f(x) \pi(x) P(x, y) \tag{13.2}$$

{Eq:OnePartDF}

By reversibility, the right-hand side of (13.2) equals

$$\sum_{x \in \Omega} \sum_{y \in \Omega} [f(x) - f(y)] f(x) \pi(y) P(y, x).$$

Reindexing shows that

$$\mathcal{E}(f) = \sum_{x \in \Omega} \sum_{y \in \Omega} [f(y) - f(x)] f(y) \pi(x) P(x, y). \tag{13.3}$$

{Eq:OtherPartDF}

Adding together (13.2) and (13.3) establishes (13.1). ■

We write $v \perp_\pi w$ to mean $\langle v, w\rangle_\pi = 0$.

{Lem:GapVar}

LEMMA 13.2. *The spectral gap* $\gamma = 1 - \lambda_2$ *satisfies*

{Eq:RayleighGap}

$$\gamma = \min_{\substack{f\,:\,E_\pi(f)=0,\\ \mathrm{Var}_\pi(f)=1}} \langle (I-P)f, f\rangle_\pi = \min_{\substack{f\,:\,E_\pi(f)=0,\\ f\not\equiv 0}} \frac{\langle (I-P)f, f\rangle_\pi}{\langle f, f\rangle_\pi}. \tag{13.4}$$

PROOF. As noted in the proof of Lemma 12.1, if $f_1, f_2, \ldots, f_n$ are the eigenfunctions of P associated to the eigenvalues ordered as in (12.1), then $\{f_k\}$ is an orthonormal basis for the inner-product space $(\mathbb{R}^n, \langle\cdot,\cdot\rangle_\pi)$. Therefore, any function f can be written as $f = \sum_{j=1}^n \langle f, f_j\rangle_\pi f_j$. Recall that Parseval's identity is the equality

$$\sum_{j=1}^{|\Omega|} |\langle f, f_j\rangle_\pi|^2 = \sum_{x\in\Omega} |f(x)|^2 \pi(x).$$

Accordingly, if $\sum_{x\in\Omega} f(x)^2\pi(x) = 1$ and $E_\pi(f) = 0$, then $f = \sum_{j=2}^{|\Omega|} a_j f_j$ where $\sum_{j=2}^{|\Omega|} a_j^2 = 1$. Thus,

$$\langle (I-P)f, f\rangle_\pi = \sum_{j=2}^{|\Omega|} a_j^2 (1-\lambda_j) \geq 1 - \lambda_2,$$

from which follows (13.4). ■

The Dirichlet form appears in the variational characterization of $g = 1 - \lambda_2$; The statement of Lemma 13.2 can be rewritten as

{Eq:GapDF}

$$g = \min_{\substack{f\,:\,E_\pi(f)=0,\\ f\not\equiv 0}} \frac{\mathcal{E}(f)}{\langle f, f\rangle_\pi}. \tag{13.5}$$

13.2. The Bottleneck Ratio Revisited

We have already met the bottleneck ratio $\Phi_\star$ in Section 8.2, where we established a lower bound on t_{mix} directly in terms of $\Phi_\star$.

We define the *spectral gap* as $\gamma = 1 - \lambda_2$. The reader should note the distinction with the absolute spectral gap $\gamma_\star$ defined earlier. As mentioned previously, for lazy chains, $\gamma = \gamma_\star$.

The following theorem bounds γ in terms of the bottleneck ratio:

{t.cheeger}

THEOREM 13.3 (Alon (1986), Jerrum and Sinclair (1989), and Lawler and Sokal (1988)). *Let* λ_2 *be the second largest eigenvalue of a reversible transition matrix* P*, and let* $\gamma = 1 - \lambda_2$. *Then*

{Eq:Cheeger}

$$\frac{\Phi_\star^2}{2} \leq \gamma \leq 2\Phi_\star. \tag{13.6}$$

PROOF OF UPPER BOUND IN EQUATION 13.6. By Lemma 13.2 and the identity in Exercise **??**,

{Eq:GapRatio}

$$\gamma = \min_{\substack{f\not\equiv 0\\ E_\pi(f)=0}} \frac{\sum_{x,y\in\Omega} \pi(x)P(x,y)\,[f(x)-f(y)]^2}{\sum_{x,y\in\Omega} \pi(x)\pi(y)\,[f(x)-f(y)]^2}. \tag{13.7}$$

For any S with $\pi(S) \leq 1/2$ define the function f_S by

$$f_S(x) = \begin{cases} -\pi(S^c) & \text{for } x \in S, \\ \pi(S) & \text{for } x \notin S. \end{cases}$$

Since $E_\pi(f_s) = 0$, it follows from (13.7) that

$$\gamma \leq \frac{2Q(S, S^c)}{2\pi(S)\pi(S^c)} \leq \frac{2Q(S, S^c)}{\pi(S)} \leq 2\Phi_S.$$

Since this holds for all S, the upper bound is proved. ■

13.3. Proof of Lower Bound in Theorem 13.3*

We need the following lemma:

{l.helplemma}

Lemma 13.4. *Given a non-negative function ψ defined on Ω, order Ω so that ψ is non-increasing. If $\pi\{\psi > 0\} \leq 1/2$, then*

$$E_\pi(\psi) \leq \Phi_\star^{-1} \sum_{\substack{x,y \in \Omega \\ x<y}} [\psi(x) - \psi(y)]\, Q(x, y).$$

Proof. Recalling that $\Phi_\star$ is defined as a minimum in (8.5), letting $S = \{x : \psi(x) > t\}$ with $t > 0$ shows that

$$\Phi_\star \leq \frac{Q(S, S^c)}{\pi(S)} = \frac{\sum_{x,y\in\Omega} Q(x, y)\mathbf{1}_{\{\psi(x)>t\geq\psi(y)\}}}{\pi\{\psi > t\}}.$$

Rearranging, and noting that $\psi(x) > \psi(y)$ only for $x < y$,

$$\pi\{\psi > t\} \leq \Phi_\star^{-1} \sum_{x<y} Q(x, y)\mathbf{1}_{\{\psi(x)>t\geq\psi(y)\}}.$$

Integrating over t, noting that $\int_0^\infty \mathbf{1}_{\{\psi(x)>t\geq\psi(y)\}} dt = \psi(x) - \psi(y)$, and using Exercise 13.1 shows that

$$E_\pi(\psi) \leq \Phi_\star^{-1} \sum_{x<y} [\psi(x) - \psi(y)]\, Q(x, y).$$

■

Let f_2 be an eigenfunction corresponding to the eigenvalue λ_2, so that $Pf_2 = \lambda_2 f_2$. Assume that $\pi\{f_2 > 0\} \leq 1/2$. (If not, use $-f_2$ instead.) Defining $f := \max\{f_2, 0\}$,

$$(I - P)f(x) \leq \gamma f(x) \quad \text{for all } x. \tag{13.8}$$

{Eq:IPf}

This is verified separately in the two cases $f(x) = 0$ and $f(x) > 0$. In the former case, (13.8) reduces to $-Pf(x) \leq 0$, which holds because $f(x) \geq 0$. In the case $f(x) > 0$, note that since $f \geq f_2$,

$$(I - P)f(x) \leq (I - P)f_2(x) = (1 - \lambda_2)f_2(x) = \gamma f(x).$$

Because $f \geq 0$,

$$\langle (I - P)f, f\rangle_\pi \leq \gamma \langle f, f\rangle_\pi.$$

Equivalently,

$$\gamma \geq \frac{\langle (I-P)f, f\rangle_\pi}{\langle f, f\rangle_\pi}.$$

Note there is no contradiction to (13.4) because $E_\pi(f) \neq 0$. Applying Lemma 13.4 with $\psi = f^2$ shows that

$$\langle f, f\rangle_\pi^2 \leq \Phi_\star^{-2} \left[\sum_{x<y} \left[f^2(x) - f^2(y) \right] Q(x,y) \right]^2 .$$

By the Cauchy-Schwarz inequality,

$$\langle f, f\rangle_\pi^2 \leq \Phi_\star^{-2} \left[\sum_{x<y} [f(x) - f(y)]^2 \, Q(x,y) \right] \left[\sum_{x<y} [f(x) + f(y)]^2 \, Q(x,y) \right].$$

Using the identity (13.1) of Exercise **??** and

$$[f(x) + f(y)]^2 = 2f^2(x) + 2f^2(y) - [f(x) - f(y)]^2 ,$$

we find that

$$\langle f, f\rangle_\pi^2 \leq \Phi_\star^{-2} \langle (I-P)f, f\rangle_\pi \left[2\langle f, f\rangle_\pi - \langle (I-P)f, f\rangle_\pi \right].$$

Let $R := \langle (I-P)f, f\rangle_\pi / \langle f, f\rangle_\pi$ and divide by $\langle f, f\rangle_\pi^2$ to show that

$$\Phi_\star^2 \leq R(2-R)$$

and

$$1 - \Phi_\star^2 \geq 1 - 2R + R^2 = (1-R)^2 \geq (1-\gamma)^2.$$

Finally,

$$\left(1 - \frac{\Phi_\star^2}{2}\right)^2 \geq 1 - \Phi_\star^2 \geq (1-\gamma)^2,$$

proving that $\gamma \geq \Phi_\star^2/2$, as required.

13.4. Comparison of Markov Chains

Recall that for lazy simple random walk on the d-dimensional torus $\mathbb{Z}_n^d$, we used coupling to show that $t_{\text{mix}} \leq C_d n^2$ and $g^{-1} \leq K_d n^2$ for constants C_d and K_d. If some edges are removed from the graph (e.g. some subset of the horizontal edges at even heights), then coupling cannot be applied due to the irregular pattern. In this chapter, such perturbations of "nice" chains can be studied via comparison. The technique will be exploited later when we study site Glauber dynamics via comparison with block dynamics.

Throughout this section, we will assume that the transition matrix P is reversible.

13.4.1. The Comparison Theorem. The following theorem—proved in various forms by Jerrum and Sinclair (1989), Diaconis and Stroock (1991), and Quastel (1992), and in the form presented here by Diaconis and Saloff-Coste, allows one to compare the behavior of similar chains to achieve bounds on the mixing time in general.

Define $E = \{(x,y) \;:\; P(x,y) > 0\}$. An *E-path* from x to y is a sequence $\gamma = (e_1, e_2, \ldots, e_m)$ of pairs from E so that $e_1 = (x,z)$ and $e_m = (w,y)$ for some z and w. The length of an E-path γ is denoted by $|\gamma|$. As usual, $Q(x,y)$ denotes $\pi(x)P(x,y)$.

Let P and $\tilde{P}$ be two transition matrices with stationary distributions π and $\tilde{\pi}$, respectively. Supposing that for each $(x,y) \in \tilde{E}$ there is an E-path from x to y, choose one and denote it by γ_{xy}. Given such a choice of paths, define the *congestion ratio* as

$$B := \max_{e \in E}\left(\frac{1}{Q(e)} \sum_{\substack{x,y \\ \gamma_{xy} \ni e}} \tilde{Q}(x,y)|\gamma_{xy}|\right). \tag{13.9}$$

{Eq:ConRat}

{Thm:Comparison}

Theorem 13.5 (The Comparison Theorem). *If B is the congestion ratio between transition matrices P and $\tilde{P}$ for a choice of E-paths, as defined in (13.9), then $\tilde{\mathcal{E}}(f) \le B\mathcal{E}(f)$. Consequently, if P and $\tilde{P}$ are reversible, then $\tilde{g} \le Bg$.*

{Cor:GapConRat}

Corollary 13.6. *Let P be a reversible and irreducible transition matrix with stationary distribution π. Suppose γ_{xy} is a choice of E-path for each x and y, and let*

$$B = \max_{e \in E} \sum_{\substack{x,y \\ \gamma_{xy} \ni e}} \pi(x)\pi(y)|\gamma_{xy}|.$$

Then the difference $g = 1 - \lambda_2$ satisfies $g \le B^{-1}$.

Proof. Let $\tilde{P}(x,y) = \pi(y)$ and $\tilde{\pi}$, and observe that

$$\tilde{\mathcal{E}}(f) = \frac{1}{2} \sum_{x,y \in \Omega} [f(x) - f(y)]^2\, \pi(x)\pi(y) = \mathrm{Var}_\pi(f).$$

Applying Theorem 13.5 shows that $\mathcal{E}(f) \ge B^{-1}\,\mathrm{Var}_\pi(f)$, from which follows the conclusion. ■

Proof of Theorem 13.5. Observe that

$$2\tilde{\mathcal{E}}(f) = \sum_{(x,y) \in \tilde{E}} \tilde{Q}(x,y)[f(x) - f(y)]^2 = \sum_{x,y} \tilde{Q}(x,y)\left[\sum_{e \in \gamma_{x,y}} df(e)\right]^2,$$

where for an edge $e = (z,w)$, we write $df(e) = f(w) - f(z)$. Applying the Cauchy-Schwarz inequality yields

$$2\tilde{\mathcal{E}}(f) \le \sum_{x,y} \tilde{Q}(x,y)|\gamma_{xy}| \sum_{e \in \gamma_{x,y}} [df(e)]^2 = \sum_{e \in E}\left[\sum_{\gamma_{xy} \ni e} \tilde{Q}(x,y)|\gamma_{xy}|\right][df(e)]^2.$$

By the definition of the congestion ratio, the right-hand side is bounded above by

$$\sum_{(z,w)\in E} BQ(z,w)[f(w)-f(z)]^2 = 2B\mathcal{E}(f),$$

completing the proof. ∎

Example 13.7 (Comparison for Simple Random Walks on Graphs). If two graphs have the same vertex set but different edge sets E and $\tilde{E}$, then

$$Q(x,y) = \frac{1}{2|E|}, \quad \text{and} \quad \tilde{Q}(x,y) = \frac{1}{2|\tilde{E}|}.$$

Therefor, the congestion ratio is simply

$$B = \left(\max_{e\in E} \sum_{\gamma_{xy}\ni e} |\gamma_{xy}| \right) \frac{|E|}{|\tilde{E}|}.$$

In our motivating example, we only removed horizontal edges at even heights from the torus. Since all odd-height edges remain, we can take $|\gamma_{xy}| \leq 3$ since we can traverse any missing edge in the torus by moving upwards, then across the edge of odd height, and then downwards. The horizontal edge in this path would then be used by at most 3 paths γ (including the edge itself). Since we removed at most one quarter of the edges, $B \leq 12$.

Thus the parameter g for the perturbed torus also satisfies $g^{-1} = O(n^2)$.

{Sec:RATcomp}

13.4.2. Random adjacent transpositions. The Comparison Theorem (Theorem 13.5) can be used to bound the convergence of the random adjacent transposition shuffle, by comparing it with the random transposition shuffle. While this 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 and indeed gives the best known bound in many examples. Note: in the course of this proof, we will introduce several constants $C_1, C_2, \ldots$. Since 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. Theorem 12.7 and Corollary 9.4 imply that the relaxation time of the random transpositions chain is at most $C_1 n \log n$. (We're already off by a factor of $\log n$ here, but we'll lose so much more along the way that it scarcely matters.)

Now, to compare. We must specify two chains on a common state space with transition matrices P and $\tilde{P}$ respectively. Here the state space is the symmetric group $\mathcal{S}_n$, while P corresponds to the random adjacent transposition shuffle and $\tilde{P}$ to the random transposition shuffle. Let $E = \{(x,y) | P(x,y) > 0\}$, and for $e = (\sigma_1, \sigma_2) \in E$ we write $Q(e) = Q(\sigma_1,\sigma_2) = P(\sigma_1,\sigma_2)U(\sigma_1) = P(\sigma_1,\sigma_2)/n!$. Define $\tilde{E}$ and $\tilde{Q}$ in a parallel way.

Because both chains are in fact random walks on the same group, we can exploit the group structure to get a well-distributed collection of paths. Let (a,b) with $a < b$, be a transposition in $\mathcal{S}_n$. Note that

{Eq:GenPaths}

$$(ab) = (a\,a-1)\ldots(b-1\,b-2)(b-1\,b)(b-1\,b-2)\ldots(a+1\,a+2)(a\,a+1). \quad (13.10)$$

Hence there is a path of length at most $2n-1$ using only adjacent transpositions (and using any single adjacent transposition at most twice) from id to (ab). We call these the *generator paths*; note that we have expressed each of the generators of the random transposition walk in terms of the generators of the random adjacent transposition walk.

To obtain a path corresponding to an arbitrary edge $(\sigma_1, \sigma_2) \in \tilde{E}$, write $\sigma_2 = (a,b)\sigma_1$. Then multiply each permutation appearing on the corresponding generator path by σ_1 on the left to get a path $\gamma_{\sigma_1\sigma_2}$ from σ_1 to σ_2.

We must estimate the congestion ratio

$$B = \max_{e\in E}\left(\frac{1}{Q(e)}\sum_{\substack{\sigma_1,\sigma_2\\ \gamma_{\sigma_1\sigma_2}\ni e}} \tilde{Q}(\sigma_1,\sigma_2)|\gamma_{\sigma_1\sigma_2}|\right) = \max_{e\in E}\frac{2(n-1)}{n^2}\sum_{\substack{\{\sigma_1,\sigma_2\}\in\tilde{E}\\ \gamma_{\sigma_1\sigma_2}\ni e}}|\gamma_{\sigma_1\sigma_2}|. \tag{13.11}$$

{Eq:CongRat}

For how many pairs $\{\sigma_1, \sigma_2\} \in \tilde{E}$ can a specific $e \in E$ appear in $\gamma_{\sigma_1\sigma_2}$? Let $e = \{\rho, (i\,i+1)\rho\}$, and let $\{\alpha, (i\,i+1)\alpha\}$ be an edge in a generator path for a transposition (a,b). Then e appears in the path for $\{\alpha^{-1}\rho, (a,b)\alpha^{-1}\rho\} \in \tilde{E}$.

Since the adjacent transposition $(i\,i+1)$ lies on the generator path of (a,b) exactly when $a \le i < i+1 \le b$, and since no generator path uses any adjacent transposition more than twice, the summation on the right-hand-side of (13.11) is bounded by $2i(n-i)(2n-1) \le n^3$. Hence

$$B \le 2n^2,$$

and Theorem 13.5 now 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 12.5 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

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

13.5. Expander Graphs*

A family of graphs $\{G_n\}$ is said to be an (d,α) *expander* family if all of the following three conditions hold for all n:

(i) $\lim_{n\to\infty} |V(G_n)| = \infty$.
(ii) G_n is d-regular.
(iii) The bottleneck ratio of the simple random walk on the graph satisfies $\Phi_\star(G_n) \ge \alpha$.

We now construct a a family of 3-regular expander graphs. This is the first construction of an expander family, due to Pinsker (1973).

The vertices of a *bipartite* graph can be colored red and blue so that red vertices are joined only to blue vertices, and blue vertices are joined only to red vertices. The set of red and blue vertices are called *sides*.

Let $G = (V,E)$ be a bipartite graph with equal sides, A and B, each with n vertices. Denote $A, B = \{1, \dots, n\}$. Let σ_1 and σ_2 be two permutations drawn uniformly at random from the permutations of $\{1,\dots,n\}$, and set the edge set to be

$$E = \{(i,i),(i,\sigma_1(i)),(i,\sigma_2(i)) \ : \ 1 \le i \le n\}.$$

{t.expand}

Theorem 13.8. *With positive probability, γ has a positive bottleneck ratio, i.e., there exists $\delta > 0$ such that for any $S \subset V$ with $|S| \le n$ we have*

$$\frac{|\{\text{edges between } S \text{ and } S^c\}|}{|S|} > \delta.$$

Proof. It is enough to prove that any $S \subset A$ of size $k \le n/2$ has at least $(1+\delta)k$ neighbors in B. This is because for any $S \subset V$ simply consider the side in which S has more vertices, and if this side has more than $n/2$ vertices, just look at an arbitrary subset of size exactly $n/2$ vertices. Let $S \subset A$ be a set of size $k \le n/2$, and denote by $N(S)$ the neighborhood of S. We wish to bound the probability that $|N(S)| \le (1+\delta)k$. Since (i,i) is an edge for any $1 \le i \le k$, we get immediately that $|N(S)| \ge k$. So all we have to enumerate is the surplus δk vertices that a set which contains $N(S)$ will have, and to make sure both $\sigma_1(S)$ and $\sigma_2(S)$ fall within that set. This argument gives

$$\mathbf{P}\{|N(S)| \le (1+\delta)k\} \le \frac{\binom{n}{k}\binom{(1+\delta)k}{k}^2}{\binom{n}{k}^2},$$

so

$$\mathbf{P}\{\text{exists } S,\ |S| \le n/2,\ |N(S)| \le (1+\delta)k\} \le \sum_{k=1}^{n/2} \binom{n}{k} \frac{\binom{n}{\delta k}\binom{(1+\delta)k}{\delta k}^2}{\binom{n}{k}^2},$$

which is strictly less than 1 for $\delta > 0$ small enough by Exercise 13.3. ∎

{Exercise:IntExp}

13.6. Problems

Exercise 13.1. Let Y be a non-negative random variable. Show that

$$\mathbf{E}(Y) = \int_0^\infty \mathbf{P}\{Y > t\}dt.$$

Hint: Write $Y = \int_0^\infty \mathbf{1}_{\{Y>t\}}dt$.

Exercise 13.2. Show that for lazy simple random walk on the box $\{1,\dots,n\}^d$, the parameter g satisfies $g^{-1} = O(n^2)$.

{Exercise:YPN1}

Exercise 13.3. To complete the proof of Theorem 13.8, prove that there exists $\delta > 0$ such that

$$\sum_{k=1}^{n/2} \frac{\binom{n}{\delta k}\binom{(1+\delta)k}{\delta k}^2}{\binom{n}{k}} < 1.$$

13.7. Notes

CHAPTER 14

The Kantorovich Metric and Path Coupling

{Chapter:PC}

We have used the total variation norm to measure distance between probability distributions. In fact, we will see in this chapter that total variation distance defines a metric on the space of probability distributions on Ω. (The reader should consult Appendix B.2 for the definition of a metric space, if needed.) When emphasizing the metric space point-of-view, we will write $\rho_{\mathrm{TV}}(\mu, \nu)$ for $\|\mu - \nu\|_{\mathrm{TV}}$. In this chapter, we introduce a generalization of ρ_{TV} called the Kantorovich metric, which we use to develop the path coupling method for bounding mixing time.

14.1. The Kantorovich Metric

Recall that a coupling of probability distributions μ and ν is a pair of random variables (X, Y), defined on the same probability space, so that X has distribution μ and Y has distribution ν.

For a given distance ρ defined on the state space Ω, the *Kantorovich metric* between two distributions on Ω is defined as

$$\rho_K(\mu, \nu) = \min\{\mathbf{E}(\rho(X, Y)) \ : \ (X, Y) \text{ is a coupling of } \mu \text{ and } \nu\}. \tag{14.1}$$ {Eq:KantorDefn}

For some history on this metric, see Vershik (2004).

By Proposition 5.5, if $\rho = \mathbf{1}_{\{x \neq y\}}$, then $\rho_K = \rho_{\mathrm{TV}}$.

{Rmk:KantorEquiv}

Remark 14.1. It is sometimes convenient to describe couplings using probabilities on the product space $\Omega \times \Omega$, instead of random variables. If q is a probability distribution on $\Omega \times \Omega$, the *projection* onto the first coordinate is the probability distribution on Ω equal to

$$q(\cdot \times \Omega) = \sum_{y \in \Omega} q(\cdot, y).$$

Likewise, the projection onto the second coordinate is the distribution $q(\Omega \times \cdot)$.

Given a coupling (X, Y) of μ and ν as defined above, the distribution of (X, Y) on $\Omega \times \Omega$ has projections μ and ν on the first and second coordinates, respectively. Conversely, given a probability distribution q on $\Omega \times \Omega$ with projections μ and ν, the identity function on the probability space $(\Omega \times \Omega, q)$ is a coupling of ν and μ.

Consequently, observing that $\mathbf{E}(\rho(X, Y)) = \sum_{(x,y) \in \Omega \times \Omega} \rho(x, y) q(x, y)$ when (X, Y) has distribution q, the Kantorovich metric can also be written as

$$\rho_K(\mu, \nu) = \min \left\{ \sum_{(x,y) \in \Omega \times \Omega} \rho(x, y) q(x, y) \ : \ q(\cdot \times \Omega) = \mu, \ q(\Omega \times \cdot) = \nu \right\}. \tag{14.2}$$ {Eq:KantorDefn2}

{Rmk:OptCoup}

Remark 14.2. The set of probability distributions on $\Omega \times \Omega$ can be identified with the $|\Omega|^2$-dimensional simplex, which is a compact subset of $\mathbb{R}^{|\Omega|^2+1}$. The set of distributions on $\Omega \times \Omega$ which project on the first coordinate to μ and project on the second coordinate to ν is a closed subset of this simplex, hence is compact. The function

$$q \mapsto \sum_{(x,y)\in\Omega\times\Omega} \rho(x,y)q(x,y)$$

is continuous on this set, hence there is a $q_\star$ so that

$$\sum_{(x,y)\in\Omega\times\Omega} \rho(x,y)q_\star(x,y) = \rho_K(\mu,\nu).$$

Such a $q_\star$ is called an *optimal* coupling of μ and ν. Equivalently, there is a pair of random variables $(X_\star, Y_\star)$, also called an optimal coupling, so that

$$\mathbf{E}(\rho(X_\star, Y_\star)) = \rho_K(\mu,\nu).$$

{KantorIsMetric}

Lemma 14.1. *ρ_K as defined in (14.1) is a metric on the space of probability distributions on Ω.*

Proof. We check the triangle inequality, and leave to the reader to verify the other two conditions.

Let μ, ν and η be probability distributions on Ω. Let p be a probability distribution on $\Omega\times\Omega$ which is a coupling of μ and ν, and let q be a probability distribution on $\Omega \times \Omega$ which is a coupling of ν and η. Define the probability distribution r on $\Omega \times \Omega \times \Omega$ by

$$r(x,y,z) := \frac{p(x,y)q(y,z)}{\nu(y)}.$$

Note that the projection of r onto its first two coordinates is p, and the projection of r onto its last two coordinates is q. The projection of r onto the first and last coordinates is a coupling of μ and η.

Assume now that p is an optimal coupling of μ and ν. (See 14.2.) Likewise, suppose that q is an optimal coupling of ν and η.

Let (X, Y, Z) be a random vector with probability distribution r. Since ρ is a metric,

$$\rho(X,Z) \leq \rho(X,Y) + \rho(Y,Z).$$

Taking expectation, because (X, Y) is an optimal coupling of μ and ν and (Y, Z) is an optimal coupling of ν and η,

$$\mathbf{E}(\rho(X,Z)) \leq \mathbf{E}(\rho(X,Y)) + \mathbf{E}(\rho(Y,Z)) = \rho_K(\mu,\nu) + \rho_K(\nu,\eta).$$

Since (X,Z) is a coupling (although not necessarily optimal) of μ and η, we conclude that

$$\rho_K(\mu,\eta) \leq \rho_K(\mu,\nu) + \rho_K(\nu,\eta).$$

■

The Kantorovich metric ρ_K "lifts" the metric ρ on Ω to a metric on the space of probability distributions on Ω. In particular, if δ_x denotes the probability distribution which puts unit mass on x, then $\rho_K(\delta_x, \delta_y) = \rho(x,y)$.

14.2. Path Coupling

{Sec:PC}

Suppose the state space of a Markov chain (X_t) has a graph structure: the states Ω form the vertices of a graph, and a collection of edges specify which states are adjacent.

Remark. This graph structure may be different from the structure inherited from the permissible transitions of the Markov chain (X_t).

Given a specification of which states are neighbors, define a *path* in Ω from x to y to be a sequence of states $\xi = (x_0, x_1, \ldots, x_\ell)$ such that the initial vertex $x_0 = x$, the final vertex $x_\ell = y$, and x_{i-1} and x_i are joined by an edge for $i = 1, \ldots, \ell$. The length of the path is ℓ. The *path metric* is defined as

$$\rho(x, y) = \min\{\text{length of } \xi \,:\, \xi \text{ a path in } \Omega \text{ from } x \text{ to } y\}. \tag{14.3}$$

{Eq:PathMetricDefn}

Notice that $\rho(x, y) \geq \mathbf{1}\{x \neq y\}$ when ρ is a path metric. Hence, for any pair (X, Y),

$$\mathbf{P}\{X \neq Y\} = \mathbf{E}\left(\mathbf{1}\{X \neq Y\}\right) \leq \mathbf{E}\left(\rho(X, Y)\right). \tag{14.4}$$

Minimizing over all couplings (X, Y) of μ and ν shows that

$$\rho_{\mathrm{TV}}(\mu, \nu) \leq \rho_K(\mu, \nu). \tag{14.5}$$

{Eq:TVvsK}

While Bubley and Dyer (1997) rediscovered the following theorem and applied it to mixing, the key idea is the fact that the Kantorovich metric *is* a metric, which goes back to Kantorovich (1942).

{Thm:PathCoupling}

Theorem 14.2 (Bubley and Dyer (1997)). *Let ρ be a path metric on the state space Ω and fix $\alpha > 0$. Suppose that for each pair of states $x, y \in \Omega$ with $\rho(x, y) = 1$ there is a coupling (X_1, Y_1) of the distributions $P(x, \cdot)$ and $P(y, \cdot)$ such that*

$$\mathbf{E}_{x,y}\left(\rho(X_1, Y_1)\right) \leq e^{-\alpha}. \tag{14.6}$$

{Eq:NeighborsContract}

Then for any two probability measures μ and ν on Ω,

$$\rho_K(\mu P, \nu P) \leq e^{-\alpha} \rho_K(\mu, \nu). \tag{14.7}$$

{Eq:MeasuresContract}
{Cor:PCMixing}

Corollary 14.3. *Suppose that the hypotheses of Theorem 14.2 hold. Then*

$$d(t) \leq e^{-\alpha t} \mathrm{diam}(\Omega),$$

and consequently

$$t_{\mathrm{mix}}(\varepsilon) \leq \frac{-\log(\varepsilon) + \log(\mathrm{diam}(\Omega))}{\alpha}.$$

Proof. By iterating (14.7), it follows that

$$\rho_K(\mu P^t, \nu P^t) \leq e^{-\alpha t} \rho_K(\mu, \nu) \leq e^{-\alpha t} \max_{x,y} \rho(x, y). \tag{14.8}$$

{Eq:IteratedContractio

Applying (14.5), and setting $\mu = \delta_x$ and $\nu = \pi$ shows that

$$\left\| P^t(x, \cdot) - \pi \right\|_{\mathrm{TV}} \leq e^{-\alpha t} \max_{x,y} \rho(x, y). \tag{14.9}$$

■

Proof of Theorem 14.2. We begin by showing that for arbitrary $x, y \in \Omega$,

$$\rho_K(P(x,\cdot), P(y,\cdot)) \leq e^{-\alpha}\rho(x,y). \tag{14.10}$$

{Eq:PointMassContract}

Fix $x, y \in \Omega$, and let $(x = x_0, x_1, \ldots, x_\ell = y)$ be a minimal-length path from x to y. By the triangle inequality for ρ_K,

{Eq:TriangleForPath}

$$\rho_K(P(x,\cdot), P(y,\cdot)) \leq \sum_{k=1}^{\ell} \rho_K(P(x_{k-1},\cdot), P(x_k,\cdot)). \tag{14.11}$$

Since ρ_K is a minimum over all couplings, the hypotheses of the theorem imply that for any two a, b with $\rho(a,b) = 1$,

{Eq:MetricContracts}

$$\rho_K(P(a,\cdot), P(b,\cdot)) \leq e^{-\alpha}\rho(a,b). \tag{14.12}$$

Since $\rho(x_{k-1}, x_k) = 1$, we can apply (14.12) to each of the terms in the sum appearing on the right-hand side of (14.11) to show

$$\rho_K(P(x,\cdot), P(y,\cdot)) \leq e^{-\alpha} \sum_{k=1}^{\ell} \rho(x_{k-1}, x_k).$$

Since ρ is a path metric, the sum on right-hand side above equals $\rho(x,y)$. This establishes (14.10).

Let p_0 be an optimal coupling of μ and ν. Define the transition matrix Q on $\Omega \times \Omega$ by setting $Q((x,y),\cdot)$ equal to an optimal coupling of $P(x,\cdot)$ and $P(y,\cdot)$. Let $((X_0,Y_0),(X_1,Y_1))$ be one step of the Markov chain with initial distribution p_0 and transition matrix Q. These definitions ensure that

{Eq:CoupleForMuNu}

$$\rho_K(\mu,\nu) = \mathbf{E}(\rho(X_0,Y_0)), \tag{14.13}$$

and for each $(x,y) \in \Omega \times \Omega$,

{Eq:CoupleXY}

$$\rho_K(P(x,\cdot), P(y,\cdot)) = \mathbf{E}\left(\rho(X_1,Y_1) \mid X_0 = x, Y_0 = y\right). \tag{14.14}$$

By Exercise 14.2, (X_1, Y_1) is a coupling of μP and νP, and so

{Eq:K1}

$$\rho_K(\mu P, \nu P) \leq \mathbf{E}\left(\rho(X_1,Y_1)\right). \tag{14.15}$$

We condition on the values of X_0 and Y_0 to decompose the expectation on the right-hand side:

$$\mathbf{E}\left(\rho(X_1,Y_1)\right) = \sum_{x,y\in\Omega} \mathbf{E}\left(\rho(X_1,Y_1) \mid X_0 = x, Y_0 = y\right) \mathbf{P}\{X_0 = x, Y_0 = y\}. \tag{14.16}$$

Using (14.14) we rewrite this as

$$\mathbf{E}\left(\rho(X_1,Y_1)\right) = \sum_{x,y\in\Omega} \rho_K(P(x,\cdot), P(y,\cdot)) \mathbf{P}\{X_0 = x, Y_0 = y\}. \tag{14.17}$$

Using (14.10) shows that

$$\mathbf{E}\left(\rho(X_1,Y_1)\right) \leq \sum_{x,y\in\Omega} e^{-\alpha}\rho(x,y) \mathbf{P}\{X_0 = x, Y_0 = y\}. \tag{14.18}$$

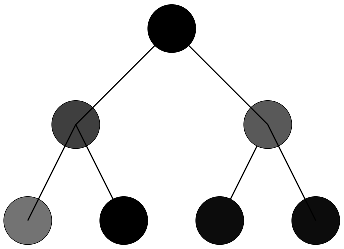


FIGURE 14.1. A proper 3-coloring of a rooted tree. (As is common practice, we have placed the root at the top.)

The right-hand side above is $e^{-\alpha}\mathbf{E}\left(\rho(X_0, Y_0)\right)$, and using (14.13) and (14.15) along with the above equations shows that

$$\rho_K(\mu P, \nu P) \leq e^{-\alpha}\rho_K(\mu, \nu). \tag{14.19}$$

■

14.3. Application: Fast Mixing for Colorings

{SSec:Coloring}

14.3.1. Coloring a graph. Suppose we have q colors, which we will represent in monochromatic text by the integers $\{1, 2, \ldots, q\}$. A *proper coloring* of a graph G is an assignment of colors to the vertices of the graph such that no two neighboring vertices are assigned the same color. The state space Ω is a subset of the set $\{0, 1, \ldots, q\}^V$ of functions $x : V \to \{0, 1, \ldots, q\}$, where the color assigned to vertex v is $x(v)$. We call elements of this state space *configurations*. We also define

$$\mathcal{N}_x(v) = \{x(w) \,:\, w \sim v\}, \tag{14.20}$$

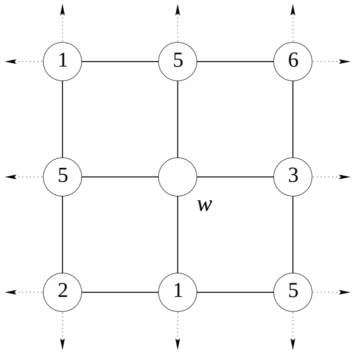
the set of colors assigned to the neighbors of v in configuration x.

Our goal is to sample uniformly from proper colorings of a graph G. In general, this is difficult to do directly, but Markov chain Monte Carlo can be used to generate an approximately uniform sample. In the next section we describe the Glauber dynamics for this distribution. The problem of mixing for this chain was first analyzed in Jerrum (1995).

14.3.2. Coloring trees. It is worth noting that in the special case where the graph is a tree, there is a direct method of sampling proper colorings. Suppose that G is a finite tree. This means that between any two vertices there is a unique connecting path. A vertex is often distinguished as the *root*, and the *depth* of a vertex is its distance from the root. The *children* of a vertex v are the neighbors of v with larger depth.

We proceed inductively, beginning with the root. Choose the color of the root uniformly at random from $\{1, \ldots, q\}$. Suppose colors have been assigned to all vertices up to depth d. For a vertex at depth $d + 1$, assign a color chosen uniformly at random from

$$\{1, 2, \ldots, q\} \setminus \{\text{color of parent}\}. \tag{14.21}$$



Colors: {~~1~~, 2, ~~3~~, 4, ~~5~~, 6}

FIGURE 14.2. Updating at vertex w. The colors of the neighbors are not available, as indicated. Fig:ColorUp

EXERCISE 14.1. Show that this method of coloring a tree samples uniformly from the set of all proper q-colorings of the tree.

{SSec:chain}

14.3.3. Mixing time for Glauber dynamics of random colorings. Glauber dynamics for random colorings of a graph with n vertices operate as follows: at each move, a vertex is chosen uniformly at random and the color of this vertex is updated. To update, a color is chosen uniformly at random from the allowable colors, which are those colors not seen among the neighbors of the chosen vertex.

Recall that $\mathcal{N}_x(w)$ is the collection of colors appearing among the neighboring vertices to w in the configuration x. Since Glauber dynamics dictate that the color of a vertex is updated by a color *not* among the neighboring colors, it is convenient to write $\mathcal{N}'_x(w)$ for the colors available for w:

{Eq:AvailableColors}

$$\mathcal{N}'_x(w) := \{1, 2, \dots, q\} \setminus \mathcal{N}_x. \tag{14.22}$$

The detailed balance equations can be easily verified for this chain. The chain moves between configurations which have the same colors everywhere except possibly at a single vertex. Suppose x is a configuration, and w is a vertex; we will write x^s_w for the configuration which agrees with x everywhere except possibly at w, where it has value s. A typical transition of the chain is from a configuration x to x^s_w, where $s \in \mathcal{N}'_x(w)$. The probability of this transition is $(n\,|\mathcal{N}'_x(w)|\,)^{-1}$, as the vertex w needs to be selected, and then the color s must be selected. The probability of going from x^s_w to x is the same, as again vertex w must be selected, and the color $x(w)$ must be selected out of the $|\mathcal{N}'_x(w)|$ allowable colors.

We will use path coupling to bound the mixing time of this chain.

{Thm:ColoringsUB}

THEOREM 14.4. *Consider the Glauber dynamics chain for random proper q-colorings of a graph with maximum degree Δ. If $q > 2\Delta$, then the mixing time satisfies*

{Eq:tmixColUB}

$$t_{\text{mix}}(\varepsilon) \leq \left(\frac{q-\Delta}{q-2\Delta}\right) n\,(\log n - \log \varepsilon)\,. \tag{14.23}$$

Proof. Let x and y be two configurations which agree everywhere except at vertex v. We describe how to simultaneously evolve two chains, one started in x and the other started in y, so that each chain viewed alone corresponds to the Glauber dynamics for random proper q-colorings.

First, we pick a vertex w uniformly at random from the vertex set of the graph. (We use a lower-case letter for the random variable w to emphasize that its value is a vertex.) We will update the color of w in both the chain started from x and the chain started from y.

If none of the neighbors of w is v, then we can update the two chains with the same color. This is fine because in both chains we pick among the available colors uniformly at random, and the available colors are the same for both chains: $\mathcal{N}'_x(w) = \mathcal{N}'_y(w)$.

Suppose now one of the neighbors of w is v. We will assume that $|\mathcal{N}'_x(w)| \leq |\mathcal{N}'_y(w)|$. If not, run the procedure described below with the roles of x and y reversed.

Generate a random color U from $\mathcal{N}'_y(w)$, and use this to update y at w. If $U \neq x(v)$, then update the configuration x at w to U. We subdivide the case $U = x(v)$ into subcases based on whether or not $|\mathcal{N}'_x(w)| = |\mathcal{N}'_y(w)|$:

case	how to update x at w
$\lvert\mathcal{N}'_x(w)\rvert = \lvert\mathcal{N}'_y(w)\rvert$	set $x(w) = y(v)$
$\lvert\mathcal{N}'_x(w)\rvert < \lvert\mathcal{N}'_y(w)\rvert$	draw a random color from $\mathcal{N}'_x$

The reader should check that this updates x to a color chosen uniformly from $\mathcal{N}'_x(w)$. The probability that the two configurations do not update to the same color is $1/|\mathcal{N}'_y(w)|$, which is bounded above by $1/(q - \Delta)$.

Now given two states x and y which are at unit distance (that is, differ in one vertex only), we have constructed a coupling (X_1, Y_1) of $P(x, \cdot)$ and $P(y, \cdot)$. The distance $\rho(X_1, Y_1)$ increases from 1 only in the case where a neighbor of v is updated and the updates are different in the two configurations. Also, the distance decreases when v is selected to be updated. In all other cases the distance stays at 1. This shows that

$$\mathbf{E}_{x,y}\left(\rho(X_1, Y_1)\right) \leq 1 - \frac{1}{n} + \frac{\deg(v)}{n}\left(\frac{1}{q-\Delta}\right). \tag{14.24}$$

{Eq:PC1}

The right-hand side of (14.24) is bounded by

$$1 - \frac{1}{n}\left(1 - \frac{\Delta}{q-\Delta}\right). \tag{14.25}$$

{Eq:ColoringDrift}

Because $2\Delta < q$, this is not more than 1. Letting $c(q, \Delta) = \left(1 - \frac{\Delta}{q-\Delta}\right)$,

$$\mathbf{E}_{x,y}\left(\rho(X_1, Y_1)\right) \leq \exp\left(-\frac{c(q,\Delta)}{n}\right).$$

Using Corollary 14.3 shows that

$$\max_{x\in\Omega} \left\| P^t(x, \cdot) - \pi \right\|_{\mathrm{TV}} \leq n \exp\left(-\frac{c(q,\Delta)}{n} t\right)$$

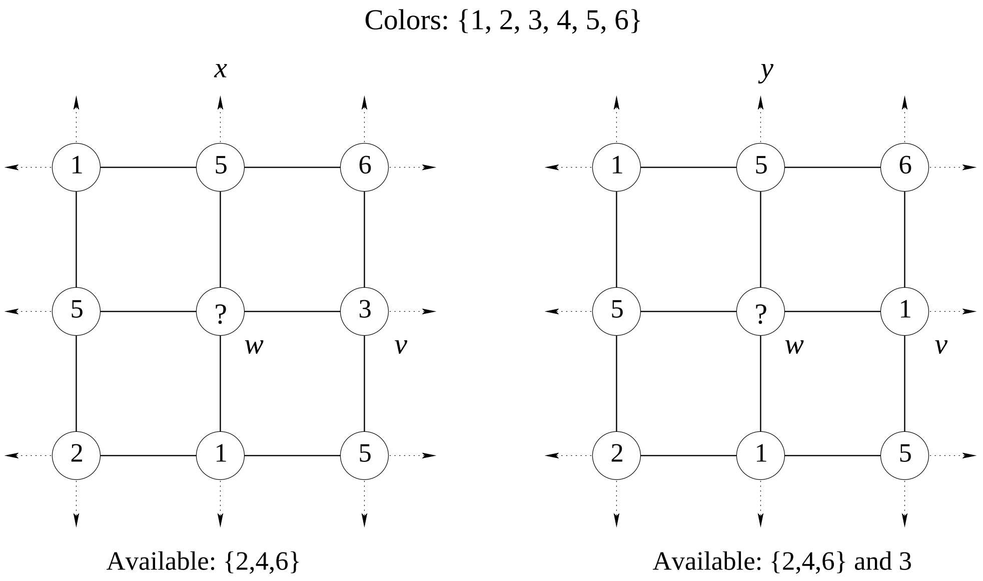


FIGURE 14.3. Jointly updating x and y when they differ only at vertex v and $|\mathcal{N}'_x| < |\mathcal{N}'_y|$ Fig:JntColorUp

and that

$$t_{\text{mix}}(\varepsilon) \le \frac{n}{c(q,\Delta)}\left(\log n + \log \varepsilon^{-1}\right). \qquad (14.26)$$ {Eq:TauForColorings}

(Note that $c(q,\Delta) > 0$ because $q > 2\Delta$.) This establishes (14.23). ■

Some condition on q and Δ is necessary to achieve the fast rate of convergence (order $n \log n$) established in Theorem 14.4, although the condition $q > 2\Delta$ is not the best known. In Example 8.3 it is shown that if Δ grows in n while q remains fixed, then in fact the mixing time is at least exponential in n.

Exercise 8.4 shows that for the graph having no edges, in which case the colors at distinct vertices do not "interact", the mixing time is at least order $n \log n$.

14.3.4. Approximate counting. Many innovations in the study of mixing times for Markov chains came from researchers motivated by the problem of *counting* combinatorial structures. While determining the exact size of a complicated set may be a "hard" problem, an approximate answer is often possible using Markov chains.

In this section, we show how the number of proper colorings can be estimated using the Markov chain analyzed in the previous section. We adapt the method described in Jerrum and Sinclair (1996) to this setting.

THEOREM 14.5. *Let Ω be the set of all proper q-colorings of the graph G of n vertices and maximal degree Δ. Let $c(q,\Delta) = 1 - \Delta/(q - \Delta)$. Given η and ε, there* *is a random variable W which can be simulated using no more than*

{Eq:NoForAC}

$$\left(\frac{n\log n + n\log(3n/\varepsilon)}{c(q,\Delta)}\right)\left(\frac{27n^3}{\eta\varepsilon^2}\right) \tag{14.27}$$

uniform random variables, so that

$$\mathbf{P}\{(1-\varepsilon)|\Omega|^{-1} \le W \le (1+\varepsilon)|\Omega|^{-1}\} \ge 1-\eta.$$

Remark 14.3. This is an example of a *fully polynomial randomized approximation scheme*, an algorithm for approximating values of the function $n \mapsto |\Omega_n|$ having a run-time that is polynomial in both the *instance size* n and the inverse error tolerated, ε^{-1}.

Proof. Let x_0 be a proper coloring of G. Enumerate the vertices of G as $\{v_1, v_2, \dots, v_n\}$. Define for $k = 0, 1, \dots, n$

$$\Omega_k = \{x \in \Omega \,:\, x(v_j) = x_0(v_j) \text{ for } j > k\}.$$

Elements of Ω_k have k "free" vertices, while the $n-k$ vertices $\{v_{k+1}, \dots, v_n\}$ are colored in agreement with x_0.

A random element of Ω_k can be generated using a slight modification to the Markov chain discussed in Section 14.3.3. The chain evolves as before, but only the vertices $\{v_1, \dots, v_k\}$ are permitted to be updated. The other vertices are frozen in the configuration specified by x_0. The bound on $t_{\text{mix}}(\varepsilon)$ in (14.26) still holds, k replacing n. (In particular, since $k \le n$, (14.26) holds.) By definition of $t_{\text{mix}}(\varepsilon)$, if

$$t(n,\varepsilon) := \frac{n\log n + n\log(3n/\varepsilon)}{c(q,\Delta)}$$

then

$$\left\|P^{t(n,\varepsilon)}(x_0,\cdot) - \pi_k\right\|_{TV} < \frac{\varepsilon}{3n}, \tag{14.28}$$

{Eq:MixConseq}

where π_k is uniform on Ω_k.

The ratio $|\Omega_{k-1}|/|\Omega_k|$ can be estimated as follows: A random element from Ω_k can be generated by running the Markov chain for $t(n,\varepsilon)$ steps. Repeating $a_n = 27n^2/\eta\varepsilon^2$ times yields a_n elements of Ω_k; let W_k be the fraction of these which are also in Ω_{k-1}. (Observe that to check if an element x of Ω_k is also an element of Ω_{k-1}, it is enough to determine if $x(v_k) = x_0(v_k)$.) From (14.28),

$$\mathbf{E}(W_k) = \frac{|\Omega_{k-1}|}{|\Omega_k|} + e_k, \quad \text{where } |e_k| \le \varepsilon/(3n).$$

Also,

$$\frac{\mathrm{Var}(W_k)}{\mathbf{E}^2(W_k)} = \frac{1-\mathbf{E}(W_k)}{a_n\mathbf{E}(W_k)} \le \frac{2n}{a_n} = \frac{2\eta\varepsilon^2}{27n}.$$

The inequality follows since $|\Omega_{k-1}|/|X_k| \ge (n-1)^{-1}$. Letting $W = W_1 \cdots W_n$, since the $\{W_k\}$ are independent,

$$\mathbf{E}(W) = \frac{1}{|\Omega|} + e, \quad \text{where } |e| \le \varepsilon/3. \tag{14.29}$$

{Eq:WSize}

Also,

$$\frac{\mathrm{Var}(W)}{\mathbf{E}^2(W)} = \prod_{k=1}^{n}\left[1 + \frac{\mathrm{Var}\, W_k}{\mathbf{E}^2(W_k)}\right] - 1 \leq \prod_{k=1}^{n}\left[1 + \frac{2\eta\varepsilon^2}{27n}\right] - 1 \leq \frac{\eta\varepsilon^2}{9}.$$

By Chebyshev's inequality,

$$\mathbf{P}\left\{\left|\frac{W}{\mathbf{E}(W)} - 1\right| \geq \varepsilon/3\right\} \leq \eta$$

Combining with (14.29),

$$\mathbf{P}\left\{\left|\frac{W}{|\Omega|^{-1}} - 1\right| \geq \varepsilon\right\} \leq \eta.$$

For each of the n variables W_k, $k = 1, \ldots, n$, we need to simulate each of a_n chains for $t(n, \varepsilon)$ steps. This shows that a total of (14.27) steps are needed. ■

14.4. Problems

{Exercise:OneStep}

EXERCISE 14.2. Let P be a transition matrix for a Markov chain. Suppose that p_0 is a coupling of μ and ν, and the transition matrix Q on $\Omega \times \Omega$ has the property that $Q((x, y), \cdot)$ is a coupling of $P(x, \cdot)$ and $P(y, \cdot)$. Let $((X_0, Y_0), (X_1, Y_1))$ be one step of a Markov chain on $\Omega \times \Omega$ started in distribution p_0 and with transition matrix Q. Show that (X_1, Y_1) is a coupling of μP and νP.

{Exer:ZeroOneMetric}

EXERCISE 14.3. Let M be an arbitrary set, and, for $a, b \in M$, define

$$\rho(a, b) = \begin{cases} 0 & \text{if } a = b, \\ 1 & \text{if } a \neq b. \end{cases} \tag{14.30}$$

Check that M is a metric space under the distance ρ.

{Exercise:KantRub}

EXERCISE 14.4. A real valued function f on Ω is called *Lipschitz* if there is a constant c so that for all $x, y \in \Omega$,

{Eq:LipDefn}

$$|f(x) - f(y)| \leq c\rho(x, y), \tag{14.31}$$

where ρ is the distance on Ω. We denote the best constant c in (14.31) by $\mathrm{lip}(f)$:

$$\mathrm{lip}(f) := \max_{\substack{x,y \in \Omega, \\ x \neq y}} \frac{|f(x) - f(y)|}{\rho(x, y)}.$$

For a probability μ on Ω, the integral $\int f d\mu$ denotes the sum $\sum_{x \in \Omega} f(x)\mu(x)$. Define

$$\tilde{\rho_K}(\mu, \nu) = \sup_{\mathrm{lip}(f) \leq 1} \left|\int f d\mu - \int f d\nu\right|.$$

Show that $\tilde{\rho_K} \leq \rho_K$.

{Ex:ColTreeConnect}

EXERCISE 14.5. Consider the space of all proper colorings of a finite tree, and say that two colorings are adjacent if they have identical colors at all vertices but one. Show that for any two 3-colorings x and y that there is a sequence of colorings $x = x_0, x_1, \ldots, x_r = y$ so that x_k and x_{k-1} are adjacent for $k = 1, 2, \ldots, r$.

14.5. Notes

Vigoda (2000) showed that if the number of colors q is larger than $(11/6)\Delta$, then the mixing times for the Glauber dynamics for random colorings is $O(n^2 \log n)$. Dyer, Greenhill, and Molloy (2002) show that the mixing times is $O(n \log n)$ provide $q \geq (2 - 10^{-12})\Delta$.

The inequality in Exercise 14.4 is actually an equality, as was shown in Kantorovich and Rubinstein (1958). In fact, the theorem is valid more generally on compact metric spaces; the proof uses a form of duality.

For more on approximate counting, see Sinclair (1993).

CHAPTER 15

The Ising Model

{Ch:Ising}

15.1. Definitions

15.1.1. Gibbs distribution. The nearest-neighbor *Ising model* is the most widely studied *spin system*, a probability distribution on $\Omega = \{-1, 1\}^V$, where V is the vertex set of a graph. An element σ of Ω is called a *configuration*, and the value $\sigma(x)$ is called the *spin* at x. As usual, we will often write $x \sim y$ if $\{x, y\}$ is an edge. The physical interpretation is that magnets, each having two possible orientations represented by $+1$ and -1, are placed on the vertices of the graph; a configuration specifies the joint orientation of these magnets.

The *energy* of a configuration σ is defined to be

$$H(\sigma) = -\sum_{\substack{v,w \in V \\ v \sim w}} \sigma(v)\sigma(w). \tag{15.1}$$

Clearly, the energy increases with the number of neighboring sites with disagreeing spins. Anyone with experience playing with magnets has observed first hand that it takes some work to place neighboring magnets in opposite orientations and hold them there.

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

$$\mu(\sigma) = \frac{1}{Z(\beta)} e^{-\beta H(\sigma)}. \tag{15.2}$$

{Eq:GibbsDefn}

The parameter β determines the importance of the energy function: if β is zero, then H plays no role and μ is the flat uniform distribution, while the bias of μ towards low-energy configurations increases with β. The physical interpretation is that β equals the reciprocal of temperature. $Z(\beta)$, called the *partition function*, is a normalizing constant required to make μ a probability distribution:

$$Z(\beta) := \sum_{\sigma \in \Omega} e^{-\beta H(\sigma)}. \tag{15.3}$$

At infinite temperature ($\beta = 0$), there is no interaction between the spins at differing vertices, i.e., the random variables $\{\sigma(v)\}_{v \in V}$ are independent.

Exercise 15.1. Show that if $\beta = 0$, the spins $\{\sigma(v) : v \in V\}$ form an independent collection of random variables.

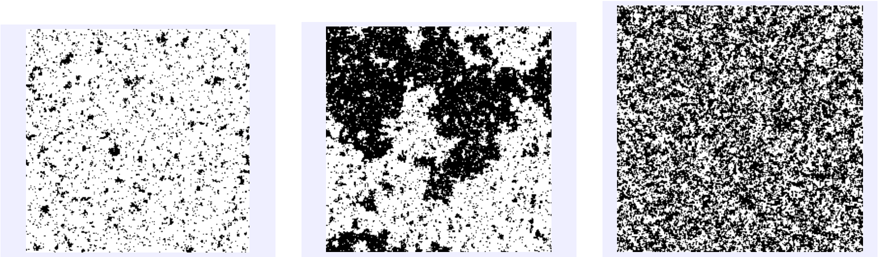


FIGURE 15.1. Glauber dynamics for the Ising model viewed at time t =??? on the 250×250 torus at low, critical, and high temperature. Fig:Ising Simulations and graphics courtesy of Raissa D'Souza.

15.1.2. Glauber dynamics. The (single-site) Glauber dynamics for μ move from a starting configuration σ by picking a vertex w uniformly at random from V and then generating a new configuration according to μ conditioned on the set of configurations agreeing with σ on vertices different from w.

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

{Eq:UpdateProb}

$$p(\sigma, w) := \frac{e^{\beta S(\sigma,w)}}{e^{\beta S(\sigma,w)} + e^{-\beta S(\sigma,w)}} = \frac{1 + \tanh(\beta S(\sigma, w))}{2}, \tag{15.4}$$

where $S(\sigma, w) := \sum_{u\,:\,u\sim w} \sigma(u)$. Note that this probability depends only on the spins at vertices adjacent to w.

REMARK 15.1. Because Glauber dynamics always have stationary distribution given by the measure used to update, the Gibbs distribution is stationary for this transition matrix.

15.2. Fast Mixing at High Temperature

In this section we use the path coupling technique of Chapter 14 to show that, for small values of β, the Glauber dynamics for the Ising model is fast mixing.

{Thm:HighTempIsing}

THEOREM 15.1. *For the Glauber dynamics for the Ising model on a graph with n vertices and maximal degree Δ, if $\tanh(\beta)\Delta < 1$, then $t_{\text{mix}} = O(n \log n)$. In particular, this holds if $\beta < \Delta^{-1}$.*

PROOF. Define the distance ρ on Ω by

$$\rho(\sigma, \tau) = \frac{1}{2} \sum_{u \in V} |\sigma(u) - \tau(u)|.$$

ρ is a path metric as defined in Section 14.2.

Let σ and τ be two configurations with $\rho(\sigma, \tau) = 1$. The spins of σ and τ agree everywhere except at a single vertex v. Assume that $\sigma(v) = -1$ and $\tau(v) = +1$.

Define $\mathcal{N}(v) := \{u\ :\ u \sim v\}$ to be the set of neighboring vertices to v.

We describe now a coupling (X, Y) of one step of the chain started in configuration σ with one step of the chain started in configuration τ.

Pick a vertex w uniformly at random from V. If $w \notin \mathcal{N}(v)$, then the neighbors of w agree in both σ and τ. As the probability of updating the spin at w to $+1$, given in (15.4), depends only on the spins at the neighbors of w, it is the same for the chain started in σ as for the chain started in τ. Thus we can update both chains together.

If $w \in \mathcal{N}(v)$, the probabilities of updating to $+1$ at w are no longer the same for the two chains, so we cannot *always* update together. We do, however, use a single random variable as the common source of noise to update both chains, so the two chains agree as often as is possible. In particular, let U be a uniform random variable on $[0, 1]$ and set

$$X(w) = \begin{cases} +1 & \text{if } U \leq p(\sigma, w), \\ -1 & \text{if } U > p(\sigma, w) \end{cases} \quad \text{and} \quad Y(w) = \begin{cases} +1 & \text{if } U \leq p(\tau, w), \\ -1 & \text{if } U > p(\tau, w). \end{cases}$$

Set $X(u) = \sigma(u)$ and $Y(u) = \tau(u)$ for $u \neq w$.

If $w = v$, then $\rho(X, Y) = 0$. If $w \notin \mathcal{N}(v) \cup \{v\}$, then $\rho(X, Y) = 1$. If $w \in \mathcal{N}(v)$ and $p(\sigma, w) < U \leq p(\tau, w)$, then $\rho(X, Y) = 2$. Thus,

$$\mathbf{E}_{\sigma,\tau}(\rho(X, Y)) \leq 1 - \frac{1}{n} + \frac{1}{n} \sum_{w \in \mathcal{N}(v)} \left[p(\tau, w) - p(\sigma, w) \right]. \tag{15.5}$$

{Eq:IsCo}

Noting that $S(w, \tau) = S(w, \sigma) + 2 = S + 2$, we obtain

$$\begin{aligned} p(\tau, w) - p(\sigma, w) &= \frac{e^{\beta(S+2)}}{e^{\beta(S+2)} + e^{-\beta(S+2)}} - \frac{e^{\beta S}}{e^{\beta S} + e^{-\beta S}} \\ &= \frac{1}{2} \left[\tanh(\beta(S + 2)) - \tanh(\beta S) \right] \\ &\leq \tanh(\beta), \end{aligned} \tag{15.6}$$

{Eq:TanhBound}

where the inequality follows from Exercise 15.2. Combining equation 15.5 with equation 15.6 shows that

$$\mathbf{E}_{\sigma,\tau}(\rho(X, Y)) \leq 1 - \frac{[1 - \Delta \tanh(\beta)]}{n} \leq \exp\left(-\frac{1 - \Delta \tanh(\beta)}{n} \right).$$

Applying Corollary 14.3, since $\operatorname{diam}(\Omega) = n$, if $\Delta \tanh(\beta) < 1$, then

$$t_{\text{mix}} = O\left(\frac{n \log n}{1 - \Delta \tanh(\beta)} \right).$$

By Exercise 15.2, if $\beta < \Delta^{-1}$, then $\Delta \tanh(\beta) < 1$ and $t_{\text{mix}} = O\left(n \log n / (1 - \beta\Delta)\right)$. ■

{Exercise:Tanh}

EXERCISE 15.2. Recall that $\tanh(x) = \sinh(x) / \cosh(x) = (e^{2x} - 1)/(e^{2x} + 1)$.

(a) Show that $\tanh[\beta(x + 2)] - \tanh[\beta x]$ is maximized at $x = -1$.
(b) Show that $\tanh(\beta) \leq \beta$ for $\beta \geq 0$.

15.3. The Complete Graph

Let G be the complete graph on n vertices, the graph which includes all $\binom{n}{2}$ possible edges.

The correct scaling is to allow β to depend on n, in particular set $\beta = \gamma/n$ for $\gamma > 0$.

{Thm:IsingCGSlow}

Theorem 15.2. *Let G be the complete graph on n vertices, and consider Glauber dynamics for the Ising model on G with $\beta = \gamma n^{-1}$.*

(i) *If $\gamma < 1$, then $t_{\mathrm{mix}} = O(n \log n)$.*

(ii) *If $\gamma > 1$, then there is a positive function $r(\gamma)$ so that $t_{\mathrm{mix}} \geq O\left(\exp\left[r(\gamma)n\right]\right)$.*

Proof. Note that $\Delta\beta = \gamma(n-1)/n = \gamma(1-n^{-1}) \leq \gamma$. Thus if $\gamma < 1$, then $\Delta\beta < 1$, and applying Theorem 15.1 shows that $t_{\mathrm{mix}} = O(n \log n)$.

Define $A_k := \{\sigma : |\{v : \sigma(v) = 1\}| = k\}$. By counting, $\pi(A_k) = a_k/Z(\beta)$, where

$$a_k := \binom{n}{k} \exp\left\{\frac{\gamma}{n}\left[\binom{k}{2} + \binom{n-k}{2} - k(n-k)\right]\right\}.$$

Taking logarithms and applying Stirling's formula shows that

$$\log(a_{\lfloor \alpha n \rfloor}) = n\phi_\gamma(\alpha)[1 + o(1)],$$

where

{Eq:IsingCGPhi}

$$\phi_\gamma(\alpha) := -\alpha \log(\alpha) - (1-\alpha)\log(1-\alpha) + \gamma\left[\frac{(1-2\alpha)^2}{2}\right]. \tag{15.7}$$

Taking derivatives shows that

$$\begin{aligned}\phi'_\gamma(1/2) &= 0\\ \phi''_\gamma(1/2) &= -4(1-\gamma).\end{aligned}$$

Hence $\alpha = 1/2$ is a critical point of ϕ_γ, and in particular it is a local maximum or minimum depending on the value of γ. See Figure 15.2 for the graph of ϕ_γ for $\gamma = 0.9$ and $\gamma = 1.1$. Take $\gamma > 1$, in which case ϕ_γ has a local minimum at $1/2$. Define

$$S = \left\{\sigma : \sum_{u \in V} \sigma(u) < 0\right\}.$$

By symmetry, $\pi(S) \leq 1/2$. Observe that the only way to get from S to S^c is through $A_{\lfloor n/2 \rfloor}$, since we are only allowed to change one spin at a time. Thus

$$Q(S, S^c) = \frac{\lceil (n/2) \rceil}{n}\pi(A_{\lfloor n/2 \rfloor}) \quad \text{and } \pi(S) = \sum_{j < \lfloor n/2 \rfloor} \pi(A_j).$$

Let α_1 be the value of α maximizing ϕ_γ over $[0, 1/2]$. Since $1/2$ is a local maximum, $\alpha_1 < 1/2$. Then

$$\Phi_S \leq \frac{\exp\{\phi_\gamma(1/2)n[1+o(1)]\}}{\pi(A_{\lfloor \alpha_1 n \rfloor})} = \frac{\exp\{\phi_\gamma(1/2)n[1+o(1)]\}}{\exp\{\phi_\gamma(\alpha_1)n[1+o(1)]\}}.$$

Since $\phi_\gamma(\alpha_1) > \phi_\gamma(1/2)$, there is a $r(\gamma) > 0$ and constant $c > 0$ so that $\Phi_\star \leq ce^{-nr(\gamma)}$. The conclusion follows from Theorem 8.1. ■

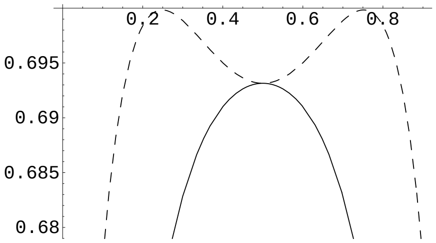


FIGURE 15.2. The function ϕ_γ defined in (15.7). The dashed graph corresponds to $\gamma = 1.1$, the solid line to $\gamma = 0.9$. Fig:IsingCG

15.4. Metastability

15.5. Lower Bound for Ising on Square*

Consider the Glauber dynamics for the Ising model in an $n \times n$ box: $V = \{(j,k) : 0 \leq j,k \leq n-1\}$ and edges connect vertices at unit Euclidean distance.

In this section we prove

{Thm:IsingLBSG}

THEOREM 15.3 (Schonmann (1987) and Thomas (1989)). *The relaxation time* $(1-\lambda_\star)^{-1}$ *of the Glauber dynamics for the Ising model in an* $n \times n$ *square in two dimensions is at least* $\exp(\psi(\beta)n)$*, where* $\psi(\beta) > 0$ *if* β *is large enough.*

More precisely, let $\gamma_\ell < 3^\ell$ *be the number of self-avoiding lattice paths starting from the origin in* $\mathbb{Z}^2$ *that have length* ℓ*, and let* $\gamma < 3$ *be the "connective constant" for the planar square lattice, defined as* $\gamma := \lim_{\ell\to\infty} \sqrt[\ell]{\gamma_\ell}$*. If* $\beta > (1/2)\log(\gamma)$ *then* $\psi(\beta) > 0$.

Much sharper and more general results are known, see the partial history in the notes. We provide here a proof following closely the method used by Dana Randall (2006) for the hardcore lattice gas.

The key idea in Randall (2006) is not to use the usual cut determined by the magnetization (as in the proof of Theorem 15.2), but rather a topological obstruction. As noted by Fabio Martinelli, this idea was already present in Thomas (1989), where contours were directly used to define a cut and obtain the right order lower bound for the relaxation time. Thus the present discussion is purely expository with no claim of originality. The argument in Thomas (1989) works in all dimensions and hence is harder to read.

REMARK 15.2. An upper bound on relaxation time of order $\exp(C(\beta)n^{d-1})$ in all dimensions follows from the "path method" of Jerrum and Sinclair (1989) for all β; The constant $C(\beta)$ obtained that way is not optimal.

To think of Theorem 15.3, it is convenient to attach the spins to the faces (lattice squares) of the lattice rather than the nodes.

DEFINITION 15.1. A *fault line* (with at most k defects) is a self-avoiding lattice path from the left side to the right side, or from the top to the bottom, of $[0,n]^2$, where each edge of the path (with at most k exceptions) is adjacent to two faces

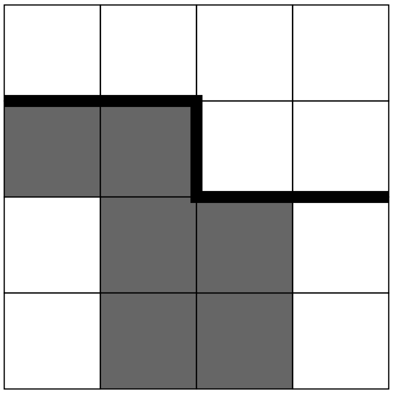


FIGURE 15.3. A faultline with one defect. Positive spins are indicated by shaded squares, while negative spins are indicated by white squares. The fault line is drawn in bold. Figure:FaultLine

with different spins on them. Thus no edges in the fault line are on the boundary of $[0,n]^2$. See Figure 15.3 for an illustration.

{Lem:YL1}

LEMMA 15.4. *Denote by F_k the set of Ising configurations in $[0,n]^2$ that have a fault line with at most k defects. Then $\pi(F_k) \le \sum_{\ell \ge n} 2\ell\gamma_\ell e^{2\beta(2k-\ell)}$. In particular, if k is fixed and $\beta > (1/2)\log(\gamma)$, then $\pi(F_k)$ decays exponentially in n.*

PROOF. For a self avoiding lattice path φ of length ℓ from the left side to the right side (or from top to bottom) of $[0,n]^2$, let F_φ be the set of Ising configurations in $[0,n]^2$ that have φ as a fault line with at most k defects. Reflecting all the spins on one side of the fault line (say, the side that contains the upper left corner) defines a one-to-one mapping from F_φ to its complement that magnifies probability by a factor of $e^{2\beta(\ell-2k)}$. This yields that $\pi(F_\varphi) \le e^{2\beta(2k-\ell)}$.

Summing this over all self-avoiding lattice paths φ of length ℓ from top to bottom and from left to right of $[0,n]^2$, and over all $\ell \ge n$, completes the proof. ■

{Lem:YL2}

LEMMA 15.5.

(i) *If in a configuration σ there is no all-plus crossing from the left side L of $[0,n]^2$ to the right side R , and there is also no all-minus crossing, then there is a fault line with no defects from the top to the bottom of $[0,n]^2$.*

(ii) *Similarly, if Γ_+ is a path of lattice squares (all labeled plus in σ) from a square q in $[0,n]^2$, to the top side of $[0,n]^2$, and Γ_- is a path of lattice squares (all labeled minus) from the same square q to the top of $[0,n]^2$, then there is a lattice path ξ from the boundary of q to the top of $[0,n]^2$ such that every edge in ξ is adjacent to two lattice squares with different labels in σ.*

PROOF.

(i) For the first statement, let A be the collection of lattice squares that can be reached from L by a path of lattice squares of the same label in σ. Let $A^\star$

denote the set of squares that are separated from R by A. Then the boundary of $A^\star$ consists of part of the boundary of $[0,n]^2$ and a fault line.

(ii) Suppose q itself is labeled minus in σ, and Γ_+ terminates in a square q_+ on the top of $[0,n]^2$ which is to the left of the square q_- where Γ_- terminates. Let A_+ be the collection of lattice squares that can be reached from Γ_+ by a path of lattice squares labeled plus in σ and denote by $A_+^\star$ the set of squares that are separated from the boundary of $[0,n]^2$ by A_+. Let ξ_1 be a directed lattice edge with q on its right and a square of Γ_+ on its left. Continue ξ_1 to a directed lattice path ξ leading to the boundary of $[0,n]^2$, by inductively choosing the next edge ξ_j to have a square (labeled plus) of A_+ on its left and a square (labeled minus) not in $A_+^\star$ on its right. It is easy to check that such a choice is always possible (until ξ reaches the boundary of $[0,n]^2$]), the path ξ cannot cycle and it must terminate between q_+ and q_- on the top side of $[0,n]^2$.

■

Proof of Theorem 15.3. Following Randall (2006), let S_+ be the set of configurations that have a top-to-bottom and a left-to-right crossing of pluses. Similarly define S_-. On the complement of $S_+ \cup S_-$ there is either no monochromatic crossing left-to-right (whence there is a top-to-bottom fault line by Lemma 15.5 or there is no monochromatic crossing top-to-bottom (whence there is a left-to-right fault line). By Lemma 15.4, $\pi(S_+) \to 1/2$ as $n \to \infty$.

Let ∂S_+ denote the external vertex boundary of S_+, that is, the set of configurations outside S_+ that are one flip away from S_+. It suffices to show that $\pi(\partial S_+)$ decays exponentially in n for $\beta > \frac{1}{2}\log(\gamma)$. By Lemma 15.4, it is enough to verify that every configuration $\sigma \in \partial S_+$ has a fault line with at most 3 defects.

The case $\sigma \notin S_-$ is handled by Lemma 15.5. Fix $\sigma \in \partial S_+ \cap S_-$ and let q be a lattice square such that flipping $\sigma(q)$ will transform σ to an element of S_+. By Lemma 15.5, there is a lattice path ξ from the boundary of q to the top of $[0,n]^2$ such that every edge in ξ is adjacent to two lattice squares with different labels in σ; by symmetry, there is also such a path $\xi^\star$ from the boundary of q to the bottom of $[0,n]^2$. By adding at most three edges of q, we can concatenate these paths to obtain a fault line with at most three defects.

Lemma 15.4 completes the proof. ■

15.6. Hardcore model

Let $G = (V, E)$ be a graph. A *hardcore configuration* is a placement of particles on V subject to an exclusion law: no pair of adjacent vertices are both occupied by particles. A configuration is represented by an element $\sigma \in \{0,1\}^V$ so that $\sigma(v)\sigma(w) = 1$ only if $\{v,w\} \notin E$.

The *hardcore model* with *fugacity* λ is the probability π on hardcore configurations defined by

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

The factor $Z(\lambda)$ normalizes π to have unit total mass.

The Glauber dynamics for the hardcore model updates a configuration $X_0 = \sigma$ to a new configuration X_1 as follows: A vertex w is chosen at random. Denote the occupied neighbors of w by $\mathcal{N}$, so that

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

If $\mathcal{N}(w) \neq \emptyset$, then $X_1 = \sigma$. If $\mathcal{N}(w) = \emptyset$, then set

$$X_1(w) = \begin{cases} 1 & \text{with probability } \lambda/(1+\lambda), \\ 0 & \text{with probability } 1/(1+\lambda). \end{cases}$$

Set $X_1(v) = \sigma(v)$ for all $v \neq w$.

{Thm:HardcoreFast}

Theorem 15.6. *For the Glauber dynamics for the hardcore model on a graph with maximum degree Δ and n vertices, if $\lambda \leq (\Delta - 1)^{-1}$, then*

$$t_{\text{mix}} = O(n \log n).$$

Proof. We use path-coupling. Let $X_0 = \sigma$ and $Y_0 = \eta$ be two configurations which differ only at a single vertex v. Assume, without loss of generality, that $\sigma(v) = 1$ and $\eta(v) = 0$. We describe how to jointly update (X_0, Y_0) to a new pair of configurations (X_1, Y_1) so that (X_0, X_1) is one step of the Glauber dynamics started from σ and (Y_0, Y_1) is one step of the Glauber dynamics started from η.

Pick a vertex w to update in both X and Y. If w is not a neighbor of v, then update the two configurations at w together.

Suppose that w is a neighbor of v. Since $\sigma(v) = 1$, it must be that $\sigma(w) = 0$ and any permitted configuration must have w unoccupied. Thus the only possibility for X_1 is that it equals σ. For Y_1, if none of the neighbors of w are occupied in η, set $Y_1(w) = 1$ with probability $\lambda/(1+\lambda)$; if w has an occupied neighbor, the only option is to set $Y_1(w) = 0$.

Note that the chance that $Y_1(w) = 1$, given that w is the updated site, is not more than $\lambda/(1-\lambda)$.

Thus,

$$\mathbf{E}_{\sigma,\eta}[\rho(X_1, Y_1)] \leq 1 - \frac{1}{n} + \frac{\Delta}{n}\frac{\lambda}{1+\lambda}.$$

Provided that $\lambda < (\Delta - 1)^{-1}$, there is $\alpha > 0$ so that

$$\mathbf{E}_{\sigma,\eta}[\rho(X_1, Y_1)] \leq 1 - \frac{\alpha}{n}.$$

Applying Corollary 14.3 finished the proof. ■

15.7. The Cycle

Consider the Glauber dynamics for the Ising model on the n-cycles (see example **??**.)

For a configuration σ, let $\phi(\sigma) = \sum_{i=1}^{n} \sigma_i$ be the sum of spins. We show now that ϕ is an eigenvalue.

$$P\phi(\sigma) = \sum_{i=1}^{n}$$

15.8. Notes

Ising's thesis (published as Ising (1925)) concerned the one-dimensional model. For information on the life of Ising, see Kobe (1997).

15.8.1. A partial history. For the Ferromagnetic Ising model with no external field and free boundary, Schonmann (1987) proved

{Thm:A}

Theorem 15.7. *In dimension 2, let $m^\star$ denote the "spontaneous magnetization", i.e., the expected spin at the origin in the plus measure in the whole lattice. Denote by $p(n;a,b)$ the probability that the magnetization (average of spins) in an $n \times n$ square is in an interval (a,b). If $-m^\star < a < b < m^\star$ then $p(n;a,b)$ decays exponentially in n.*

(The rate function was not obtained, only upper and lower bounds.)

Using the easy direction of the Cheeger inequality (an immediate consequence of the variational formula for eigenvalues) this yields Theorem 15.3.

Chayes, Chayes and Schonmann (1987) then extended Theorem 15.7 to all $\beta > \beta_c$. (Recall that for the planar square lattice $\beta_c = \log(1 + \sqrt{2})/2$.)

Theorem 15.3 stated explicitly and proved in Thomas (1989) who extended it to all dimensions $d \geq 2$. He did not use the magnetization to define a cut, but instead his cut was defined by configurations where there is a contour of length (or in higher dimensions $d \geq 3$, surface area) larger than an^{d-1} for a suitable small $a > 0$. Again the rate function was only obtained up to a constant factor and he assumed β was large enough for a Peierls argument to work.

In the breakthrough book of Dobrushin, Kotecký and Shlosman (1992) the correct rate function (involving surface tension) for the large deviations of magnetization in 2 dimensions was identified, and established for large β.

This was extended by Ioffe (1995) to all $\beta > \beta_c$. The consequences for mixing time (a sharp lower bound) and a corresponding sharp upper bound were established in Cesi, Guadagni, Martinelli, and Schonmann (1996).

In higher dimensions, a lower bound for mixing time of the right order (exponential in n^{d-1}) is known for all $\beta > \beta_c(d, slab)$ where $\beta_c(d, slab)$ is conjectured but not proved to coincide with β_c. This follows from the magnetization large deviation bounds of Pisztora (1996).

The correct rate function was not established yet but a related result under plus boundary conditions is in Cerf and Pisztora (2000).

Fast convergence for the hardcore model at low λ was proven by Luby and Vigoda (1999). There upper bound on λ is better than the bound of $(\Delta - 1)^{-1}$ obtained in Theorem 15.6.

CHAPTER 16

Lamplighter walks

{Ch:Lamplighter}

16.1. Introduction

{Sec:LampIntro}

Given a finite graph $G = (V, E)$, imagine placing a lamp at each vertex. Now allow a (possibly intoxicated?) lamplighter to random walk on G, switching lights randomly on and off as he visits them.

We can model this process as a random walk on the *wreath product* $G^* = \{0,1\}^V \times V$, whose vertices are ordered pairs (f, v), where $v \in V$ and $f \in \{0,1\}^V$. There is an edge between (f, v) and (h, w) in the graph G^* if v, w are adjacent in G and $f(u) = h(u)$ for $u \notin \{v, w\}$. We call f the *configuration of the lamps* and v the *position of the lamplighter*. In the configuration function f, zeroes correspond to lamps that are off, and ones to lamps that are on.

We now build a random walk on G^*. Let Q denote the transition probabilities for the lamplighter walk, and P the transition probabilities of the lazy simple random walk on G.

- For $v \neq w$, $Q[(f, v), (h, w)] = P(v, w)/4$ if f and h agree outside of $\{v, w\}$.
- When $v = w$, $Q[(f, v), (h, v)] = P(v, v)/2$ if f and h agree off of $\{v\}$.

That is, at each time step, the current lamp is randomized, the lamplighter moves, and then the new lamp is also randomized. (The second lamp at w is randomized in order to make the chain reversible.) (We have used the lazy walk on G as the basis for the construction to avoid periodicity, or even near-periodicity, problems later).

It should be clear that the configuration of lamps on visited states is fully random. Hence allowing the lamplighter to walk for the cover time of the underlying walk suffices to randomize the lamp configuration—although perhaps not the position of the lamplighter himself!

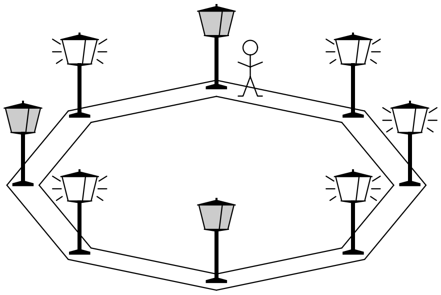


FIGURE 16.1. A lamplighter on a 8-cycle.

{fig:lamp}

Here we study the connections between parameters of the underlying chain G and the lamplighter chain G^*. For a large class of examples, a small constant times the cover time of G bounds the mixing time for G^*.

The relaxation time t_{rel} (defined in Section 12.4) of a lamplighter chain G^* is closely related to the maximal hitting time t_{hit} (defined in Section 11.2) of the underlying walk.

The proofs of these connections between parameters of random walk on a graph and the corresponding lamplighter walk use many of the techniques we have studied in previous chapters.

In all our results in this chapter, we will restrict our attention to situations where the underlying walk G is transitive (defined in Section 7.5). Exercise 7.5 implies that the stationary distributions π and π^* of the walks on G and G^*, respectively, are both uniform.

16.2. A map of many parameters of Markov chains

{Sec:InequalityMap}

We have by now accumulated a large number of time parameters associated with a finite Markov chain. Some of these measure mixing directly. Others, such as the cover time and the various flavors of hitting time, attempt to measure the geometry of the chain.

We have also proved many inequalities relating these parameters, and here we pause give a "map" of the results—most of which will be cited later in this chapter! For now, define $t_1 \lesssim t_2$ if there exists a constant $c > 0$ such that $t_1 \lesssim ct_2$. We have shown:

$$t_{\mathrm{rel}} \lesssim t_{\mathrm{mix}} \lesssim t_{\mathrm{hit}} \lesssim \mathbf{E}C.$$

16.3. Relaxation time bounds

{Sec:LampRelax}{relaxthm}

Theorem 16.1. *Let $\{G_n\}$ be a sequence of vertex transitive graphs such that $|V_n|$ goes to infinity. Then there exist constants $c_1 < c_2$ such that for sufficiently large n,*

{reltimeeq}

$$c_1 t_{\mathrm{hit}}(G_n) \leq t_{\mathrm{rel}}(G_n^*) \leq c_2 t_{\mathrm{hit}}(G_n). \tag{16.1}$$

Proof of Theorem 16.1. The lower bound uses the variational formula (13.4) to show that the spectral gap for the transition kernel Q^t is bounded away from 1 when $t = t_{\mathrm{hit}}(G_n)/4$. For the upper bound, we use the coupling contraction method of Chen (1998), which we have already discussed (Theorem 12.8). The geometry of lamplighter graphs allows us to refine this coupling argument and restrict attention to pairs of states such that the position of the lamplighter is the same in both states.

Let's start with the lower bound. Fix a vertex $w \in G$, and define $\varphi : V^* \to \{0,1\}$ by $\varphi(f,v) = f(w)$. Then $\mathrm{Var}(\varphi) = 1/4$. After running for t steps, started in stationarity, the lamplighter has either visited vertex w or he hasn't. Applying Lemma 13.1 gives

$$\mathcal{E}(\varphi) = \frac{1}{2}\mathbf{E}\left[\varphi(Y_t) - \varphi(Y_0)\right]^2 = \frac{1}{2}\sum_{v \in V} \pi(v)\frac{1}{2}\mathbf{P}_v(\tau_w \leq t) = \frac{1}{4}\mathbf{P}_\pi(\tau_w \leq t),$$

where $\{Y_t\}$ is a stationary Markov chain on G^* and $\mathcal{E}(\varphi) = \mathcal{E}(\varphi|\varphi)$ is the Dirichlet form. For any t,

$$\mathbf{E}_v\tau_w \leq t + t_{\text{hit}}\mathbf{P}_v(\tau_w > t)$$

(if a walk on G started at v has not hit w by time t, the expected additional time to arrive at w is bounded by t_{hit}, regardless of the value of the state at time t). By Lemma 11.2, $t_{\text{hit}} \leq 2\mathbf{E}_\pi\tau_w$. Averaging over π gives

$$t_{\text{hit}} \leq 2t + 2t_{\text{hit}}\mathbf{P}_\pi(\tau_w > t).$$

Substituting $t = t_{\text{hit}}/4$ and rearranging yields

$$\mathbf{P}_\pi[\tau_w \leq t_{\text{hit}}/4] \leq \frac{3}{4}.$$

By (13.4), we thus have

$$1 - |\lambda_2|^{t_{\text{hit}}/4} \leq \frac{3}{4},$$

and so

$$\log 4 \geq \frac{t_{\text{hit}}}{4}(1 - |\lambda_2|),$$

which gives the claimed lower bound on $t_{\text{rel}}(G^*)$, with $c_1 = \frac{1}{\log 4}$.

For the upper bound, we use a coupling argument from Chen (1998). Suppose that φ is an eigenfunction for p with eigenvalue λ_2. To conclude that $t_{\text{rel}}(G^*) \leq \frac{(2+o(1))t_{\text{hit}}}{\log 2}$, it suffices to show that $\lambda_2^{2t_{\text{hit}}} \leq 1/2$. For a configuration h on G, let $|h|$ denote the Hamming length of h. Let

$$M = \sup_{f,g,x} \frac{|\varphi(f,x) - \varphi(g,x)|}{|f - g|}$$

be the maximal amount that φ can vary over two elements with the same lamplighter position. If $M = 0$, then $\varphi(f,x)$ depends only on x, and so $\psi(x) = \varphi(f,x)$ is an eigenfunction for the transition operator on G. Since $t_{\text{rel}}(G) \leq t_{\text{hit}}$ (see Aldous and Fill (in progress), Chapter 4), this would imply that $|\lambda_2^{2t_{\text{hit}}}| \leq e^{-4}$. We may thus assume that $M > 0$.

Consider two walks, one started at (f,x) and one at (g,x). Couple the lamplighter component of each walk and adjust the configurations to agree at each site visited by the lamplighter. Let (f',x') and (g',x') denote the position of the coupled walks after $2t_{\text{hit}}$ steps. Let K denote the transition operator of this coupling. Because φ is an eigenfunction,

$$\begin{aligned}
\lambda_2^{2t_{\text{hit}}} M &= \sup_{f,g,x} \frac{|p^{2t_{\text{hit}}}\varphi(f,x) - p^{2t_{\text{hit}}}\varphi(g,x)|}{|f-g|} \\
&\leq \sup_{f,g,x} \sum_{f',g',x'} K^{2t_{\text{hit}}}[(f,g,x) \to (f',g',x')] \frac{|\varphi(f',x') - \varphi(g',x')}{|f'-g'|} \frac{|f'-g'|}{|f-g|} \\
&\leq M \sup_{f,g,x} \frac{\mathbf{E}|f'-g'|}{|f-g|}.
\end{aligned}$$

But at time $2t_{\text{hit}}$, each lamp that contributes to $|f-g|$ has probability of at least $1/2$ of having been visited, and so $\mathbf{E}|f'-g'| \le |f-g|/2$. Dividing by M gives the required bound of $\lambda_2^{2t_{\text{hit}}} \le 1/2$. ∎

16.4. Mixing time bounds

{Lem:MeanMedianCover}{Sec:LampMix}

Lemma 16.2. *Consider an irreducible finite Markov chain on state space Ω with transition matrix P, and let C be its cover time. Let t_m have the following property: for any $x \in X$,*

$$\mathbf{P}_x(C \le t_m) \ge 1/2.$$

Then $\mathbf{E}_x C \le 2t_m$ for any $x \in \Omega$.

Proof. Consider starting at a state $x \in \Omega$ and running in successive intervals of t_m steps. The probability of states being missed in the first interval is at most $1/2$. If some states are missed in the first interval, then the probability that all are covered by the end of the second interval is at least $1/2$, by the definition of t_m. Hence the probability of not covering by time $2t_m$ is at most $1/4$. In general,

$$\mathbf{P}_x(C > kt_m) < \frac{1}{2^k}.$$

We may conclude that C is dominated by t_m times a geometric(1/2) random variable, and thus $\mathbf{E}_x C$ is at most $2t_m$. ∎

{tvconvthm}

Theorem 16.3. *Let (G_n) be a sequence of vertex transitive graphs with $|V_n| \to \infty$, and let C_n be the cover time for lazy simple random walk on G_n. For any $\varepsilon > 0$, there exist constants c_1, c_2 such that for sufficiently large n,*

{tvconveq}

$$c_1 \mathbf{E} C_n \le t_{\text{mix}}(G_n^*) \le c_2 \mathbf{E} C_n. \tag{16.2}$$

Proof of Theorem 16.3. *Upper bound.* Let (F_t, X_t) denote the state of the lamplighter chain at time t. We will run the lamplighter chain long enough that, with high probability, every lamp has been visited and enough additional steps have been taken to randomize the position of the lamplighter.

Set $t = 8\mathbf{E}C_n + t_{\text{mix}}(G_n, 1/8)$ and fix an initial state $(\mathbf{0}, v)$. We have

$$\left\|Q^t((\mathbf{0},v)) - \pi^*\right\|_{TV} = \sum_s \mathbf{P}(C_n = s)\mathbf{E}\left(\left\|Q^t((\mathbf{0},v)) - \pi^*\right\|_{\text{TV}} \middle| C_n = s\right). \tag{16.3}$$

Since $\mathbf{P}(C_n > 8\mathbf{E}C_n) < 1/8$ and the total variation distance between distributions is bounded by 1, we can bound

{Eq:AlreadyCovered}

$$\left\|Q^t((\mathbf{0},v)) - \pi^*\right\|_{TV} \le 1/8 + \sum_{s<8\mathbf{E}C_n} \mathbf{P}(C_n = s)\mathbf{E}\left(\left\|Q^t((\mathbf{0},v)) - \pi^*\right\|_{\text{TV}} \middle| C_n = s\right). \tag{16.4}$$

Note that when $C_n = s < t$ the strong Markov property implies that the distribution of F_t is uniform on $(0,1)^n$ and the distribution of X_t is $P^{t-s}(X_s, \cdot)$. Hence the total variation distance for the lamplighter walk, conditioned on the cover time, is the same as the total variation distance for the underlying walk started at the last state visited:

{Eq:TVSame}

$$\mathbf{E}\left(\left\|Q^t((\mathbf{0},v)) - \pi^*\right\|_{\text{TV}} \middle| C_n = s\right) = \left\|P^{t-s}(X_s, \cdot) - \pi\right\|_{TV}. \tag{16.5}$$

Combining the estimates (16.4) and (16.5) yields

$$\left\| Q^t((\mathbf{0}, v)) - \pi^* \right\|_{TV} \leq 1/8 + (7/8)(1/8) < 1/4, \tag{16.6}$$

since, by the definition of t, we have $t - C_n \geq t_{\text{mix}}(G_n, 1/8)$ exactly when $C_n \geq 8\mathbf{E}C_n$.

To complete the upper bound, we need only check that $t_{\text{mix}}(G_n, 1/8)$ is bounded by a constant times $\mathbf{E}C_n$.

Lower bound. We break into two cases, depending on whether $t_{\text{hit}} = t_{\text{hit}}(G_n) > (1/200)\mathbf{E}C_n$ or not. If it is in fact true that $t_{\text{hit}} > (1/200)\mathbf{E}C_n$, then there exist constants c_3, c_4, and c_5 such that

$$t_{\text{mix}}(G_n^*) \geq c_3 t_{\text{rel}}(G_n^*) \geq c_4 t_{\text{hit}} \geq c_5 \mathbf{E}C_n,$$

by Theorems 12.7 and 16.1 and our initial assumption, respectively, so we're done.

Otherwise, we may assume that $t_{\text{hit}} \leq (1/200)\mathbf{E}C_n$. In this case we will find a time $\tilde{t}$ such that (i) after $\tilde{t}$ steps, the lamplighter walk is not yet well mixed (ii) after a constant multiple of $\tilde{t}$ steps, the probability of having covered all states is at least $1/2$. Lemma 16.2 will then tell us that we must have made positive progress towards $\mathbf{E}C$.

Define the event B_t by

$$B_t = \{\text{at least } n - 12\sqrt{n} \text{ lamps have been visited by time } t\}.$$

Fix an initial state $(\mathbf{0}, w) \in V^*$, and set

$$\tilde{t} = \max\{t : \mathbf{P}_{(\mathbf{0},w)}(B_t) < 2/3\}. \tag{16.7}$$

{Eq:TildetDef}

Claim 1: $d(\tilde{t}) > 1/4$. (That is, the lamplighter walk is not well mixed after $\tilde{t}$ steps.)

We contrast the number of lamps *off* in stationarity and at time $\tilde{t}$. In particular, let A be the event that at time $\tilde{t}$, *at least* $(n + 5\sqrt{n})/2$ *lamps are off*. In stationarity, the expectation of the number of lamps off is $n/2$, and the variance is $n/4$. By Chebyshev's inequality,

$$\pi^*(A) \leq \frac{n/4}{(5\sqrt{n}/2)^2} = \frac{1}{25}. \tag{16.8}$$

{Eq:LampStat}

Now consider the distribution at time $\tilde{t}$. If we condition on the walk having missed exactly $M \geq 12\sqrt{n}$ lamps, then the number of lamps off has mean $(n - M)/2 + M = (n + M)/2$ and variance $(n - M)/4 < n/4$. By Chebyshev's inequality, the probability of the event A, conditioned on exactly M lamps having been missed, is at least

$$1 - \frac{n/4}{\left(M/2 - 5\sqrt{n}/2\right)^2} \geq 1 - \frac{1}{(12-5)^2} = \frac{48}{49}. \tag{16.9}$$

{Eq:CondLampsOff}

Since the estimate of (16.9) holds for every $M \geq 12\sqrt{n}$, we may conclude that

$$Q^{\tilde{t}}((\mathbf{0}, w), A) \geq \mathbf{P}_{(\mathbf{0},w)}(B^c_{\tilde{t}}) \left(\frac{48}{49}\right) \geq \frac{1}{3}\left(\frac{48}{49}\right) = \frac{16}{49}. \tag{16.10}$$

{Eq:LampLotsMissed}

Finally, the estimates (16.8) and (16.10), together with the definition (5.1) of total variation distance, imply the Claim—note that $\frac{16}{49} - \frac{1}{25} > \frac{1}{4}$.

Claim 2: $\mathbf{P}(C_n > 5\tilde{t} + 5 + 75t_{\text{hit}}) > 1/2$. (That is, we are likely to have covered the base graph after a small multiple of $\tilde{t}$ steps—plus a few more.)

First, define an *experiment* to be the following two-step procedure:

(i) Choose a uniform lamp $v \in V$, independently of the progress of the lamplighter walk so far. Run the lamplighter walk until the lamplighter is at position v.
(ii) Run the lamplighter walk for $\tilde{t} + 1$ steps.

Call an experiment *successful* if, in fact, fewer than $12\sqrt{n}$ lamps are left unvisited during stage (ii). By the definition (16.7) of $\tilde{t}$, the probability an experiment is successful is at least $2/3$.

Fix a start state $(\mathbf{0}, w) \in V^*$. Let τ be the (random) number of steps required to run 5 consecutive experiments. The probability that at least 3 experiments are successful is at least $1 - (1/3)^5 - 5(2/3)(1/3)^4 - \binom{5}{2}(2/3)^2(1/3)^3 = 64/81$.

For any lamp $v \in V$,

$$\mathbf{P}_{(\mathbf{0},w)}(v \text{ not visited by time } \tau | \text{at least 3 experiments succeeded}) < \left(\frac{12\sqrt{n}}{n}\right)^3 = \frac{1728}{n^{3/2}}.$$

since whether v is visited or not during separate successful experiments are independent, and vertex transitivity ensures that each vertex has the same probability of being visited during a successful experiment. Hence the expected number of unvisited lamps, conditioned on at least 3 successful experiments, is bounded by $n(1728/n^{3/2}) = O(1/\sqrt{n})$. Markov's inequality now implies that the probability of one or more unvisited lamps, conditioned on at least 3 successes, is $O(1/\sqrt{n})$.

We have a random time when the probability of covering is high. Now, we find a fixed time such that the probability of covering is also high.

Notice that each stage (ii) is of fixed length. Notice also that the time required for stage (i) of each experiment is the time to hit a uniform state in G, whose expectation is certainly bounded by t_{hit}. Furthermore, the probability that sum of all the stage (i) times is large (at least $75t_{\text{hit}}$) is smaller that the probability that at least one of them is large (at least $15t_{\text{hit}}$). Thus

$$\mathbf{P}_{(\mathbf{0},w)}(\tau > 5\tilde{t} + 5 + 75t_{\text{hit}}) < 5\left(\frac{1}{15}\right) = \frac{1}{3}.$$

Now, the probability that all the lamplighter positions have been covered by time $5\tilde{t} + 5 + 75t_{\text{hit}}$ is at least

$$\left(1 - \frac{1}{3}\right)\left(\frac{64}{81}\right)\left(1 - O\left(\frac{1}{\sqrt{n}}\right)\right),$$

which is larger than $1/2$ for sufficiently large n.

Claim 3: $\tilde{t} > \mathbf{E}C_n/50$ for sufficiently large n.

By Claim 2 and Lemma 16.2,

$$\mathbf{E}C_n < 2(5\tilde{t} + 5 + 75t_{\text{hit}}).$$

By our assumption that $t_{\text{hit}} < (1/200)\mathbf{E}C_n$,

$$\mathbf{E}C_n < 10\tilde{t} + 10 + (3/4)\mathbf{E}C_n$$

Rearranging gives

$$\frac{\mathbf{E}C_n}{40} - 1 < \tilde{t},$$

so for sufficiently large n we have

$$\frac{\mathbf{E}C_n}{50} < \tilde{t}.$$

Combining Claim 1 and Claim 3 gives the desired lower bound on $t_{\mathrm{mix}}(G_n^*)$. ■

Remark. Matthews says that the expected cover time can be greater than t_{hit} by at most a factor of log (size of state space). In fact, in examples we care about, the cover time ends up at one end of the interval or the other; it is perhaps unfortunate that our lower bound proof above has to work so hard in the range in between.

16.5. Examples

{Sec:LampExamples}

16.5.1. The complete graph. When G_n is the complete graph on n vertices, with self-loops, then the chain we study on G_n^* is a random walk on the hypercube—although not quite the standard one, since two bits can change in a single step. This example was analyzed by Häggström and Jonasson (1997). The maximal hitting time is n and the expected cover time is an example of the coupon collector problem. Hence the relaxation time and the mixing time for G_n^* are $\Theta(n)$ and $\Theta(n \log n)$, respectively, just as for the standard walk on the hypercube.

16.5.2. Hypercube. Let $G_n = \mathbb{Z}_2^n$, the n-dimensional hypercube. We showed in Exercise 11.10 that the maximal hitting time is on the order of 2^n and in Exercise 11.22 that the cover time is on the order of $n2^n$. In Example 12.4, we saw that for lazy random walk on G_n, we have $t_{\mathrm{rel}}(G_n) = n$. Finally, in Section 12.6, we showed that $t_{\mathrm{mix}}(\varepsilon, G_n) \sim (n \log n)/2$. By Theorem 16.1, $t_{\mathrm{rel}}(G_n^*)$ is on the order of 2^n, and Theorem 16.3 shows that the convergence time in total variation on G_n^* is on the order of $n2^n$.

16.5.3. Tori. For the one-dimensional case, we note that Häggström and Jonasson (1997) examined lamplighter walks on cycles. Here both the maximal hitting time and the expected cover time of the base graph are $\Theta(n^2)$—see Section 4.1 and Example 11.19. Hence the lamplighter chain on the cycle has both its relaxation time and its mixing time of order $\Theta(n^2)$.

For higher-dimensional tori, we have proved enough about hitting time and cover times to see that the relaxation time and the mixing time grow at different rates in every dimension $d \geq 2$.

{Zdtheorem}

Theorem 16.4. *For the random walk $\{X_t\}$ on $(\mathbb{Z}_n^2)^* = \mathbb{Z}_2 \wr \mathbb{Z}_n^2$ in which the lamplighter performs simple random walk with holding probability $1/2$ on $\mathbb{Z}_n^2$, there exist constants c_2 and C_2 such that the relaxation time satisfies*

$$c_2 n^2 \log n \leq t_{\mathrm{rel}}(\mathbb{Z}_n^2)^* \leq C_2 n^2 \log n. \tag{16.11}$$

{Ztworeltime}

There also exist constants c_2' and C_2' such that the total variation mixing time satisfies

$$c_2' n^2 (\log n)^2 \leq t_{\mathrm{mix}}((\mathbb{Z}_n^2)^*) \leq C_2' n^2 (\log n)^2. \tag{16.12}$$

{Ztwotv}

More generally, for any dimension $d \geq 3$, there are constants c_d, C_d, c'_d and C'_d such that on $\mathbb{Z}_2 \wr \mathbb{Z}_n^d = (\mathbb{Z}_n^d)^*$, the relaxation time satisfies

$$c_d n^d \leq t_{\mathrm{rel}}((\mathbb{Z}_n^d)^*) \leq C_d n^d, \tag{16.13}$$

{Zdreltime}

and the total variation mixing time satisfies

{Zdtv}

$$c'_d n^d \log n \leq t_{\mathrm{mix}}(\varepsilon, (\mathbb{Z}_n^d)^*) \leq C'_d n^d \log n. \tag{16.14}$$

Proof. These follow immediately from combining the bounds on the hitting time and the cover time for tori from Proposition 11.9 and Example 11.22, respectively, with Theorems 16.1 and 16.3. ■

{Sec:LampNotes}

16.6. Notes

The results of this chapter are all taken from Peres and Revelle (2004), which derives sharper versions of the bounds we discuss, especially in the case of the two-dimensional torus, and also considers the time required for convergence in the uniform metric.

Scarabotti and Tolli (2007) study the eigenvalues of lamplighter walks. They compute the spectra for the complete graph and the cycle, and use representations of wreath products to give more general results.

CHAPTER 17

Continuous-time chains and simulation in the continuum*

{Ch:Continuous}

In this chapter, we study two topics: Markov chains in which the time parameter is now continuous (so we have a collection of random variables $(X_t)_{t\in[0,\infty)}$ indexed by the non-negative real numbers), and we introduce some methods for simulating continuous random variables described by density functions on $\mathbb{R}^k$.

17.1. Continuous-Time Chains

Here we will not study the most general type of continuous-time Markov chains, but will restrict ourselves to the following special case: the times between transitions of the chain are i.i.d. exponential random variables of unit rate, and moves are made according to a transition matrix P.

More precisely: let $T_1, T_2, \ldots$ be an independent and indentically distributed sequence of exponential random variables. That is, each takes values in $[0, \infty)$ and has distribution function

$$\mathbf{P}\{T_i \le t\} = \begin{cases} 1 - e^{-t} & t \ge 0, \\ 0 & t < 0. \end{cases}$$

Let $(\Phi_k)_{k=1}^{\infty}$ be a Markov chain with transition matrix P, and define $S_k = \sum_{i=1}^{k} T_i$. Define

$$X_t := \Phi_k \quad \text{for } t \in [S_k, S_{k+1}). \tag{17.1}$$ {Eq:CTDefn}

We will call $(X_t)_{t\ge0}$ the continuous-time Markov chain with transition matrix P.

Letting $N_t = \max\{k : S_k \le t\}$, we have $N_t = k$ if and only if $S_k \le t < S_{k+1}$. From the definition (17.1),

$$\mathbf{P}_x\{X_t = y \mid N_t = k\} = \mathbf{P}_x\{\Phi_k = y\} = P^k(x, y). \tag{17.2}$$ {Eq:GivenN}

Also, the distribution of N_t is Poisson with mean t (Exercise 17.6):

$$\mathbf{P}\{N_t = k\} = \frac{e^{-t}t^k}{k!}. \tag{17.3}$$ {Eq:NPois}

The *heat kernel* is defined as $H_t(x, y) = \mathbf{P}_x\{X_t = y\}$. From (17.2) and (17.3), we have

$$H_t(x, y) = \sum_{k=0}^{\infty} \mathbf{P}_x\{X_t = y \mid N_t = k\} \frac{e^{-t}t^k}{k!} = \sum_{k=0}^{\infty} \frac{e^{-t}t^k}{k!} P^k(x, y).$$

For a $m \times m$ matrix M, define the $m \times m$ matrix $e^M = \sum_{i=0}^{\infty} \frac{M^i}{i!}$. In matrix representation, $H_t = e^{t(P-I)}$.

Theorem 17.1. *Let P be an irreducible transition matrix, and let let H_t be the corresponding heat kernel. Then there exists a unique probability distribution π so that $\pi H_t = \pi$ for all $t \geq 0$, and*

$$\max_{x\in\Omega} \|H_t(x,\cdot) - \pi\|_{\mathrm{TV}} \to 0 \quad \text{as} \quad t \to \infty.$$

Remark 17.1. Note that the above theorem does not require that P is aperiodic, unlike Theorem 5.6. This is one advantage of working with continuous-time chains.

Proof. Let π be a solution to $\pi = \pi P$, which exists by Proposition 3.8.
We have

$$(\pi H_t)(y) = \sum_{x\in\Omega} \pi(x) \sum_{k=0}^{\infty} \frac{e^{-t}t^k}{k!} P^k(x,y) = \sum_{k=0}^{\infty} \frac{e^{-t}t^k}{k!} \sum_{x\in\Omega} \pi(x) P^k(x,y).$$

The change of summation is justified because all terms are non-negative. The inner sum on the right-hand side is $\pi(y)$, which does not depend on k and can be pulled outside the first sum, which adds to unity. This shows that $\pi H_t = \pi$.

Note that if $\widetilde{P} := (P+I)/2$, then $\widetilde{P}$ is aperiodic and irreducible, and $\pi\widetilde{P} = \pi$. By Theorem 5.6, $\max_{x\in\Omega} \|\widetilde{P}^t(x,\cdot) - \pi\|_{\mathrm{TV}} \to 0$. Also,

$$\widetilde{H}_t = e^{t(\widetilde{P}-I)} = e^{t/2(P-I)} = H_{t/2},$$

so we will be done if we can show that

$$\lim_{t\to\infty} \max_{x\in\Omega} \|\widetilde{H}_t(x,\cdot) - \pi\|_{\mathrm{TV}} = 0.$$

We have

$$\begin{aligned} \left\|\widetilde{H}_t(x,\cdot) - \pi\right\|_{\mathrm{TV}} &= \frac{1}{2} \sum_{y\in\Omega} |\widetilde{H}_t(x,y) - \pi(y)| \\ &= \frac{1}{2} \sum_{y\in\Omega} \left| \sum_{k=0}^{\infty} \frac{e^{-t}t^k}{k!} \left[\widetilde{P}^k(x,y) - \pi(y)\right] \right|. \end{aligned}$$

Applying the triangle inequality and interchanging the sums shows that

$$\left\|\widetilde{H}_t(x,\cdot) - \pi\right\|_{\mathrm{TV}} \leq \sum_{k=0}^{\infty} \frac{e^{-t}t^k}{k!} \left\|\widetilde{P}^k(x,\cdot) - \pi\right\|_{\mathrm{TV}} \leq \mathbf{E}(d(N_t)).$$

Since $\mathbf{E}(N_t) = t$ and $\mathrm{Var}(N_t) = t$, by Chebyshev's inequality, $\mathbf{P}\{|N_t - t| \geq \alpha\sqrt{t}\} \leq \alpha^{-2}$. Take $\alpha = \varepsilon^{-1/2}$. Since $d(k) \to 0$, let k_0 be such that $d(k) \leq \varepsilon$ for $k \geq k_0$, and let B be such that $d(k) \leq B$ for all k. Take t large enough so that $t - \alpha\sqrt{t} > k_0$. Then

$$\mathbf{E}(d(N_t)) \leq B\mathbf{P}\{N_t < t - \alpha\sqrt{t}\} + \varepsilon \leq (B+1)\varepsilon.$$

Since $\varepsilon > 0$ was arbitrary, this completes the proof. ■

This leads us to define

{Eq.tmixc}
$$t_{\mathrm{mix}}^{\mathrm{cont}}(\varepsilon) := \inf\left\{t \geq 0 \,:\, \max_{x\in\Omega} \|H_t(x,\cdot) - \pi\|_{\mathrm{TV}} \leq \varepsilon\right\}. \tag{17.4}$$

17.2. Continuous vs. discrete mixing

{Sec:CDMix}

In this section, our goal is to relate the mixing time of the lazy Markov chain and the continuous time Markov chain that correspond to a given transition matrix P. Recall that $\widetilde{P} = \frac{1}{2}(I + P)$ is the corresponding lazy chain. Let H_t be the heat kernel for the corresponding continuous-time chain

Our goal is to show that H_t and $\widetilde{P}$ have about the same mixing time (up to constant). Recall that π denotes the stationary distribution. We do not assume here aperiodicity or reversibility of P.

The following theorem shows that $t_{\text{mix}}(\varepsilon)$ and $t_{\text{mix}}^{\text{cont}}(\varepsilon)$ are comparable:

{Thm:CDMix}

THEOREM 17.2.

(i) *If* $\|\widetilde{P}^k(x,\cdot) - \pi\|_{\text{TV}} < \varepsilon$, *then* $\|H_k(x,\cdot) - \pi\|_{\text{TV}} < 2\varepsilon$ *provided that k is large enough.*

(ii) *If* $\|H_m(x,\cdot) - \pi\|_{\text{TV}} < \varepsilon$ *and m is large enough, then* $\|\widetilde{P}^{4m}(x,\cdot) - \pi\|_{\text{TV}} < 2\varepsilon$.

The proof requires the following lemma:

{Lem:Key}

LEMMA 17.3. *Let Y be a Binomial*$(4m, \frac{1}{2})$ *random variable, and let $\Psi = \Psi_m$ be a Poisson variable with mean m. Then*

$$\eta_m := \|\mathbf{P}\{Y \in \cdot\} - \mathbf{P}\{\Psi + m \in \cdot\}\|_{\text{TV}} \to 0$$

as $m \to \infty$.

PROOF OF LEMMA 17.3. Note that Y and $\Psi + m$ both have mean $2m$ and variance m. Given $\varepsilon > 0$, let $A = 2\varepsilon^{-1/2}$ and deduce from Chebyshev's inequality that

$$\mathbf{P}\left\{|Y - 2m| \geq A\sqrt{m}\right\} \leq \varepsilon/4 \quad \text{and} \quad \mathbf{P}\left\{|\Psi - m| \geq A\sqrt{m}\right\} \leq \varepsilon/4. \tag{17.5}$$

{deviation}

Now, using Stirling's formula and computing directly, we can show that for $|j| \leq A\sqrt{m}$,

$$\mathbf{P}\{Y = 2m + j\} \sim \frac{1}{\sqrt{2\pi m}} e^{-j^2/2m},$$

$$\mathbf{P}\{\Psi + m = 2m + j\} \sim \frac{1}{\sqrt{2\pi m}} e^{-j^2/2m}.$$

Here, we mean that the ratio of the two sides tends to 1 as $m \to \infty$, uniformly for all j such that $|j| \leq A\sqrt{m}$. **Add more details here!**

Thus for large m we have

$$\sum_{|j| \leq A\sqrt{m}} [\mathbf{P}\{Y = 2m + j\} - \mathbf{P}\{\Psi + m = 2m + j)\}] \leq \sum_{|j| \leq A\sqrt{m}} \varepsilon \mathbf{P}\{Y = 2m + j\} \leq \varepsilon$$

Dividing this by 2 and using (17.5) establishes the lemma. ■

PROOF OF THEOREM 17.2. *Step 1.* First we show that shortly after the original chain is close to equilibrium, so is the continuous time chain. Suppose that $\|P^k(x,\cdot) - \pi\|_{\text{TV}} < \varepsilon$. Then for $\delta > 0$ and $t \geq k(1 + \delta)$, conditioning on the value of Ψ_t and applying the triangle inequality give

$$\|H_t(x,\cdot) - \pi\|_{\text{TV}} \leq \sum_{j \geq 0} \mathbf{P}\{\Psi_t = j\} \|P^j(x,\cdot) - \pi\|_{\text{TV}} \leq \mathbf{P}\{\Psi_t < k\} + \varepsilon,$$

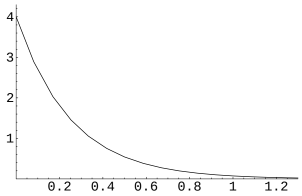


FIGURE 17.1. $f(x) = 4e^{-4x}$, the exponential probability density function with rate 4. Fig:ExpDen

where the right-hand inequality used monotonicity of $\|P^j(x,\cdot) - \pi\|_{\mathrm{TV}}$ in j. By the law of large numbers, $\mathbf{P}\{\Psi_t < k\} \to 0$ as $k \to \infty$ for $t \geq k(1+\delta)$. Thus if k is sufficiently large, then $\|H_t(x,\cdot) - \pi\|_{\mathrm{TV}} < 2\varepsilon$ for such t.

Step 2. Let $\widetilde{H}_t$ be the continuous time version of the lazy chain $\widetilde{P}$. We claim that $\widetilde{H}_t = H_{t/2}$. There are several ways to see this. One is to observe that that H_t involves Ψ_t steps of the lazy chain $\widetilde{P}$. Each of these steps is a step of P with probability $1/2$, and a delay otherwise; thinning a Poisson process of rate 1 this way yields a Poisson process of rate $1/2$ (Exercise!).

Alternatively, matrix exponentiation yields a very short proof of the claim:

$$\widetilde{H}_t = e^{t(\widetilde{P}-I)} = e^{t(\frac{P+I}{2}-I)} = e^{\frac{t}{2}(P-I)}.$$

Step 3. Now suppose that the lazy chain is close to equilibrium after k steps, that is $\|\widetilde{P}^k(x,\cdot) - \pi\|_{\mathrm{TV}} < \varepsilon$. We then claim that the continuous time chain is close to equilibrium shortly after time $k/2$. This is an easy corollary of Steps 1 and 2. If k is large enough, then for $t = \frac{k}{2}(1+\delta)$, we have

$$\|H_t(x,\cdot) - \pi\|_{\mathrm{TV}} = \|\widetilde{H}_{2t} - \pi\|_{\mathrm{TV}} < 2\varepsilon.$$

Step 4. Suppose that $\|H_m(x,\cdot) - \pi\|_{\mathrm{TV}} < \varepsilon$; we claim that $\|\widetilde{P}^{4m}(x,\cdot) - \pi\|_{\mathrm{TV}} < 2\varepsilon$ for large m.

After the discrete-time chain has been run for Ψ_m steps, running it for another m steps will not increase the distance to π, so $\|H_mP^m(x,\cdot) - \pi\|_{\mathrm{TV}} < \varepsilon$. (Observe that the matrices H_m and P^m commute.) Now

$$H_mP^m = \sum_{k\geq 0} \mathbf{P}\{\Psi + m = k\}P^k, \widetilde{P}^{4m} \qquad\qquad = \sum_{k\geq 0} \mathbf{P}\{Y = k\}P^k,$$

so Lemma 17.3 and the definition of total variation, or its coupling description, give

$$\|H_mP^m(x,\cdot) - \widetilde{P}^{4m}(x,\cdot)\|_{\mathrm{TV}} \leq \eta_m,$$

whence

$$\begin{aligned}\|\widetilde{P}^{4m}(x,\cdot) - \pi\|_{\mathrm{TV}} &\leq \|P^mH_mP^m(x,\cdot) - \pi\|_{\mathrm{TV}} + \eta_m \\ &\leq \varepsilon + \eta_m.\end{aligned}$$

as needed. ■

{Sec:InvCDF}

17.3. Continuous Simulation

17.3.1. Inverse distribution function method.

{Example:Exponential}

Example 17.4. Let U be a uniform random variable on $[0,1]$, and define $Y = -\lambda^{-1}\log(1-U)$. The distribution function of Y is

$$F(t) = \mathbf{P}\{Y \le t\} = \mathbf{P}\{-\lambda^{-1}\log(1-U) \le t\} = \mathbf{P}\{U \le 1 - e^{-\lambda t}\}. \tag{17.6}$$

As U is uniform, the rightmost probability above equals $1 - e^{-\lambda t}$, the distribution function for an exponential random variable with rate λ. (The graph of an exponential density with $\lambda = 4$ is shown in Figure 17.1.)

This calculation leads to the following algorithm:

(1) Generate U.
(2) Output $Y = -\lambda^{-1}\log(1-U)$.

The algorithm in Example 17.4 is a special case of the *inverse distribution function method* for simulating a random variable with distribution function F, which is practical *provided that F can be inverted efficiently*. Unfortunately, there are not very many examples where this is the case.

Suppose that F is strictly increasing, so that its inverse function $F^{-1} : [0,1] \to \mathbb{R}$ is defined everywhere. Recall that F^{-1} is the function so that $F^{-1} \circ F(x) = x$ and $F \circ F^{-1}(y) = y$.

We now show how, using a uniform random variable U, to simulate X with distribution function F. For a uniform U, let $X = F^{-1}(U)$. Then

$$\mathbf{P}\{X \le t\} = \mathbf{P}\{F^{-1}(U) \le t\} = \mathbf{P}\{U \le F(t)\}. \tag{17.7}$$

The last equality follows because F is strictly increasing, so $F^{-1}(U) \le t$ if and only if $F\left(F^{-1}(U)\right) \le F(t)$. Since U is uniform, the probability on the right can be easily evaluated to get

$$\mathbf{P}\{X \le t\} = F(t). \tag{17.8}$$

That is, the distribution function of X is F.

17.3.2. Acceptance-rejection sampling. Suppose that we have a black box which on demand produces a uniform sample from a region R' in the plane, but what we really want is to sample from another region R which is contained in R' (see Figure 17.2.)

If independent points are generated, each uniformly distributed over R', until a point falls in R, then this point is a uniform sample from R. (Exercise 17.9.)

Now we want to use this idea to simulate a random variable X with density function f given that we know how to simulate a random variable Y with density function g.

We will suppose that

$$f(x) \le Cg(x) \text{ for all } x, \tag{17.9}$$

{Eq:RejectionAssumptio

for some constant C. We will see that good choices for the density g minimizes the constant C. Because f and g both integrate to unity, $C \ge 1$.

Here is the algorithm:

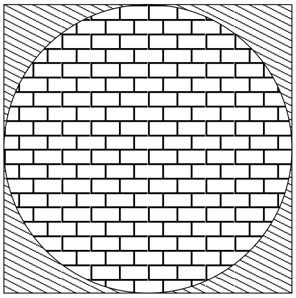


FIGURE 17.2. R' is the diagonally hatched square, and R is the bricked circle. Fig:Regions

(1) Generate a random variable Y having probability density function g.
(2) Generate a uniform random variable U.
(3) Conditional on $Y = y$, if $Cg(y)U \leq f(y)$, output the value y and halt.
(4) Repeat.

{Exercise:UnderDensity}

EXERCISE 17.1. Show that if (Y, U_Y) is the pair generated in one round of the rejection sampling algorithm, then (Y, U_Y) is uniformly distributed over the region bounded between the graph of Cg and the horizontal axis. Conversely, if g is a density, and a point is sampled from the region under the graph of g, then the projection of this point onto the x-axis has distribution g.

We now show that this method generates a random variable with probability density function f. Given that $Y = y$, the random variable $U_y := Cg(y)U$ is uniform on $[0, Cg(y)]$. By Exercise 17.1, the point (Y, U_Y) is uniform over the region bounded between the graph of Cg and the horizontal axis. We halt the algorithm if and only if this point is also underneath the graph of f. By Exercise 17.9, in this case, the point is uniformly distributed over the region under f. But again by Exercise 17.1, the horizontal coordinate of this point has distribution f. (See Figure 17.3.)

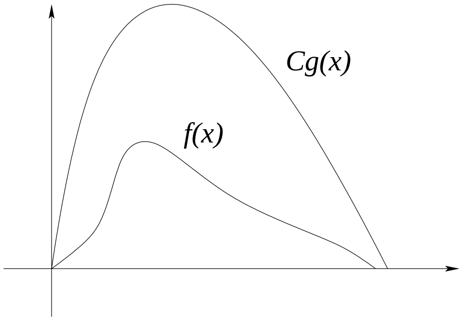


FIGURE 17.3. The probability density function f lies below the scaled probability density function of g. Fig:TwoPdfs

The value of C determines the efficiency of the algorithm. The probability the algorithm terminates on any trial, given that $Y = y$ is $f(y)/Cg(y)$. Using the law of total probability, the unconditional probability is C^{-1}. The number of trials required is geometric, with success probability C^{-1}, and so the expected number of trials before terminating is C.

We comment here that there is a version of this method for discrete random variables; the reader should work on the details for herself.

{Exa:Gamma}

Example 17.5. Consider the gamma distribution with parameters α and λ. Its probability density function is

$$f(x) = \frac{x^{\alpha-1}\lambda^{\alpha}e^{-\lambda x}}{\Gamma(\alpha)}. \tag{17.10}$$

(The function Γ in the denominator is defined to normalize the density so that it integrates to unity. It has several interesting properties, notably that $\Gamma(n) = (n-1)!$ for integers n.)

The distribution function does not have a nice closed-form expression, so inverting the distribution function does not provide an easy method of simulation.

We can use the rejection method here, when $\alpha > 1$, bounding the density by a multiple of the exponential density

$$g(x) = \mu e^{-\mu x}.$$

The constant C depends on μ, and

$$C = \sup_x \frac{[\Gamma(\alpha)]^{-1}(\lambda x)^{\alpha-1}\lambda e^{-\lambda x}}{\mu e^{-\mu x}}.$$

A bit of calculus shows that the supremum is attained at $x = (\alpha - 1)/(\lambda - \mu)$, and

$$C = \frac{\lambda^{\alpha}(\alpha-1)^{\alpha-1}e^{1-\alpha}}{\Gamma(\alpha)\mu(\lambda-\mu)^{\alpha-1}}.$$

Some more calculus shows that the constant C is minimized for $\mu = \lambda/\alpha$, in which case

$$C = \frac{\alpha^{\alpha}e^{1-\alpha}}{\Gamma(\alpha)}.$$

The case of $\alpha = 2$ and $\lambda = 1$ is shown in Figure 17.4, where $4e^{-1}\frac{1}{2}e^{-x/2}$ bounds the gamma density.

We end the example by commenting that the exponential is easily simulated by the inverse distribution function method, as the inverse to $1-e^{-\mu x}$ is $(-1/\mu)\ln(1-u)$.

17.3.3. Simulating Normal random variables. Recall that a standard normal random variable has the "bell-shaped" probability density function specified by

$$f(x) = \frac{1}{\sqrt{2\pi}}e^{-\frac{1}{2}x^2}. \tag{17.11}$$ {Eq:NormalPdf}

The corresponding distribution function Φ is the integral

$$\Phi(x) = \int_{-\infty}^{x} \frac{1}{\sqrt{2\pi}}e^{-\frac{1}{2}t^2}\,dt, \tag{17.12}$$

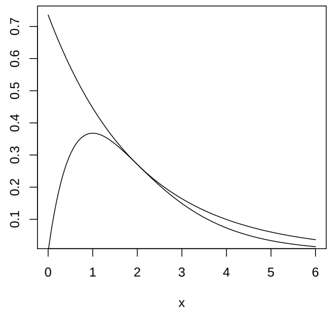


FIGURE 17.4. The Gamma density for $\alpha = 2$ and $\lambda = 1$, along with $4e^{-1}$ times the Exponential density of rate $1/2$. Fig:Gamma

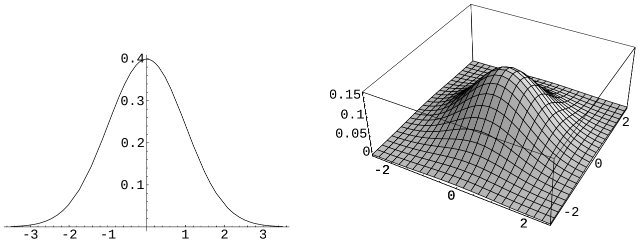


FIGURE 17.5. The standard normal density on the left, and on the right the joint density of two independent standard Normal variables. Fig:BVNorm

which cannot be evaluated in closed form. The inverse of Φ likewise cannot be expressed in terms of elementary functions. As a result the inverse distribution function method requires numerical evaluation of Φ^{-1}. We present here another method of simulating from Φ which does not require evaluation of the inverse of Φ.

Let X and Y be independent standard normal random variables. Geometrically, the ordered pair (X, Y) is a random point in the plane. The joint probability density function for (X, Y) is shown in Figure 17.5.

We will write (R, Θ) for the representation of (X, Y) in polar coordinates, and define $S := R^2 = X^2 + Y^2$ to be the squared distance of (X, Y) to the origin.

The distribution function of S is

$$\mathbf{P}\{S \leq t\} = \mathbf{P}\{X^2 + Y^2 \leq t\} = \iint_{D(\sqrt{t})} \frac{1}{2\pi} e^{-\frac{x^2+y^2}{2}} dxdy, \tag{17.13}$$

where $D(\sqrt{t})$ is the disc of radius $\sqrt{t}$ centered at the origin. Changing to polar coordinates, this equals

$$\int_0^{\sqrt{t}} \int_0^{2\pi} \frac{1}{2\pi} e^{-\frac{r^2}{2}} r dr d\theta = 1 - e^{-t/2}. \tag{17.14}$$

We conclude that S has an exponential distribution with mean 2.

{Exercise:RotationInvar

EXERCISE 17.2. Argue that since the joint density $(2\pi)^{-1}\exp[-(x^2+y^2)/2]$ is a function of $s = x^2 + y^2$, the distribution of Θ must be uniform and independent of S.

To summarize, the squared radial part of (X, Y) has an exponential distribution, its angle has a uniform distribution, and these are independent.

Our standing assumption is that we have available independent uniform variables; here we need two, U_1 and U_2. Define $\Theta := 2\pi U_1$ and $S := -2\log(1 - U_2)$, so that Θ is uniform on $[0, 2\pi]$, and S is independent of Θ and has an exponential distribution.

Now let (X, Y) be the Cartesian coordinates of the point with polar representation $(\sqrt{S}, \Theta)$. Our discussion shows that X and Y are independent standard normal variables.

17.3.4. Sampling from the simplex.

Let Δ_n be the $n-1$-dimensional simplex:

$$\Delta_n := \left\{ (x_1, \ldots, x_n) \; : \; x_i \geq 0, \; \sum_{i=1}^n x_i = 1 \right\} \tag{17.15}$$

This is the collection of probability vectors of length n. We consider here the problem of sampling from Δ_n.

Let $U_1, U_2, \ldots, U_{n-1}$ be i.i.d. uniform variables in $[0, 1]$, and define $U_{(k)}$ to be the k-th smallest among these.

{Exercise:UnifOrder}

EXERCISE 17.3. Show that the vector $(U_{(1)}, \ldots, U_{(n-1)})$ is uniformly distributed over the set $A_{n-1} = \{(u_1, \ldots, u_{n-1}) \; : \; u_1 \leq u_2 \leq \cdots \leq u_{n-1} \leq 1\}$.

Let $T : \mathbb{R}^{n-1} \to \mathbb{R}^n$ be the linear transformation defined by

$$T(u_1, \ldots, u_{n-1}) = (u_1, u_2 - u_1, \ldots, u_{n-1} - u_{n-2}, 1 - u_{n-1}).$$

{Exercise:UniformLinear

EXERCISE 17.4. Suppose that X is uniformly distributed on a region A of $\mathbb{R}^d$, and the map $T : \mathbb{R}^d \to \mathbb{R}^r, d \leq r$ is a linear transformation. A useful fact is that for a region $R \subset \mathbb{R}^d$,

$$\text{Volume}_d(TR) = \sqrt{\det(T^t T)}\,\text{Volume}(R),$$

where $\text{Volume}_d(TR)$ is the d-dimensional volume of $TR \subset \mathbb{R}^r$. Use this to show that $Y = TX$ is uniformly distributed over TA.

Note that T maps A_{n-1} linearly to Δ_n, so Exercise 17.3 and Exercise 17.4 together show that $(X_1, \ldots, X_n) = T(U_{(1)}, \ldots, U_{(n-1)})$ is uniformly distributed on Δ_n.

We can now easily generate a sample from Δ_n: throw down $n - 1$ points uniformly in the unit interval, sort them along with the points 0 and 1, and take the vector of successive distances between the points.

This requires sorting n variables, which in fact can be avoided. The following exercise requires knowledge of the change-of-variables formula for d-dimensional random vectors.

{Exercise:ExpSimplex}

EXERCISE 17.5. Let $Y_1, \ldots, Y_n$ be i.i.d. exponential variables, and define

$$X_i = \frac{Y_i}{Y_1 + \cdots + Y_n}. \tag{17.16}$$

Show that $(X_1, \ldots, X_n)$ is uniformly distributed on Δ_n

{Exer:PP}

17.4. Problems

EXERCISE 17.6. Let $T_1, T_2, \ldots$ be an i.i.d. sequence of exponential random variables of unit rate, let $S_k = \sum_{i=1}^k T_i$, and let $N_t = \max\{k : S_k \leq t\}$.

(a) Show that S_k has a Gamma distribution with shape parameter n and rate parameter 1, i.e. its density function is

$$f_k(s) = \frac{s^{k-1}e^{-s}}{(k-1)!}.$$

(b) Show by computing $\mathbf{P}\{S_k \leq t < S_{k+1}\}$ that N_t is a Poisson random variable with mean t.

[SOLUTION]

EXERCISE 17.7. Here we outline an alternative proof that N_t has a Poisson distribution with mean t.

(a) Divide the interval $[0, t]$ into t/Δ intervals of length Δ.

EXERCISE 17.8. Describe how to use the inverse distribution function method to simulate from the probability density function

$$f(x) = \begin{cases} 2x & \text{if } 0 < x < 1, \\ 0 & \text{otherwise.} \end{cases}$$

{Exercise:Subregion}

EXERCISE 17.9. Let $R \subset R' \subset \mathbb{R}^k$. Show that if points uniform in R' are generated until a point falls in R, that this point is uniformly distributed over R. Recall that this means that the probability of falling in any subregion B of R is equal to $\text{Area}(B)/\text{Area}(R)$.

EXERCISE 17.10. Find a method for simulating the random variable Y with density

$$g(x) = e^{-|x|/2}.$$

Use the rejection method to simulate a random variable X with the standard Normal density given in (17.11).

{Exercise:OrderStats}

EXERCISE 17.11. Let $U_1, U_2, \ldots, U_n$ be independent random variables, each uniform on the interval $[0, 1]$. Let $U_{(k)}$ be the kth *order statistic*, the k-th smallest among $\{U_1, \ldots, U_n\}$, so that

$$U_{(1)} < U_{(2)} < \cdots < U_{(n)}.$$

The purpose of this exercise is to give several different arguments that

{Eq:ExOrdStat}
$$\mathbf{E}\left(U_{(k)}\right) = \frac{k}{n+1}. \tag{17.17}$$

Fill in the details for the following proofs of (17.17):

(a) Find the density of $U_{(k)}$ and integrate.

(b) Find the density of $U_{(n)}$, and observe that given $U_{(n)}$ the other variables are the order statistics for uniforms on the interval $[0, U_{(n)}]$. Then apply induction.

(c) Let $Y_1, \dots, Y_n$ be independent and identically distributed exponential variables with mean 1, and let $S_1 = Y_1, S_2 = Y_1 + Y_2, \dots$, be their partial sums. Show that the random vector

$$\frac{1}{S_{n+1}} \left(S_1, S_2, \dots, S_n\right) \tag{17.18}$$
{Eq:SString}

has constant density on the simplex

$$\mathcal{A}_n = \{(x_1, \dots, x_n) \ : \ 0 < x_1 < x_2 < \cdots < x_n < 1\}.$$

Conclude that (17.18) has the same law as the vector of order statistics.

17.5. Notes

To make the estimates in Section 17.2 more quantitative, one needs an estimate of the convergence rate for η_m in the Lemma 17.3. This can be done in at least three ways:

- We could apply a version of Stirling's formula with error bounds (see Equation B.11) in conjunction with large deviation estimates for Y and Ψ.
- We could replace Stirling's formula with a precise version of the local central limit theorem, see e.g. Spitzer (1976).
- One can also use Stein's method, see Chyakanavichyus and Vaĭtkus (2001) or Röllin (2006).

These methods all show that η_m is of order $m^{-1/2}$.

For a stimulating and much wider discussion of univariate simulation techniques, Devroye (1986) is an excellent reference.

CHAPTER 18

Countable State-Space Chains*

{Ch:Infinite}

In this chapter we treat the case where Ω is not necessarily finite, although we assume it is a countable set. A classical example is the simple random walk on $\mathbb{Z}^d$. This walker moves on $\mathbb{Z}^d$ by choosing uniformly at random among her $2d$ nearest neighbors. There is a striking dependence on the dimension d: when $d \geq 3$, the walker may wonder off "to infinity", never returning to her starting place, while this is impossible in dimensions $d \leq 2$. We will return to this example later.

As before, P is a function from $\Omega \times \Omega$ to $[0, 1]$ satisfying $\sum_{y \in \Omega} P(x, y) = 1$ for all $x \in \Omega$. We still think of P as a matrix, except now it has countably many rows and columns. The matrix arithmetic in the finite case extends to the countable case without any problem. The joint distribution of the infinite sequence (X_t) is still specified by P along with a starting distribution μ on Ω.

18.1. Recurrence and Transience

{Xmpl:SRWRecur0}

Example 18.1 (Simple random walk on $\mathbb{Z}$). Let (X_t) have transition matrix

$$P(j, k) = \begin{cases} 1/2 & \text{if } k = j \pm 1, \\ 0 & \text{otherwise.} \end{cases}$$

Let A_k be the event that the walker started from zero reaches absolute value 2^k before it returns to zero. By symmetry, $\mathbf{P}_0(A_1) = 1/2$ and $\mathbf{P}_0(A_{k+1} \mid A_k) = 1/2$. Thus $\mathbf{P}_0(A_k) = 2^{-k}$, and in particular

$$\mathbf{P}_0\{\tau_0^+ = \infty\} = \mathbf{P}_0\left(\bigcap_{k=1}^{\infty} A_k\right) = \lim_{k \to \infty} \mathbf{P}_0(A_k) = 0.$$

The penultimate equality follows since the events $\{A_k\}$ are decreasing.

{Xmple:BRWTrans}

Example 18.2 (Biased random walk on $\mathbb{Z}$). Suppose now that a walker on $\mathbb{Z}$ makes biased moves, so that

$$P(j, k) = \begin{cases} q & \text{for } k = j - 1, \\ p & \text{for } k = j + 1, \end{cases}$$

where $q < p$ and $q + p = 1$. Recall the gambler's ruin formula for biased random walk, c.f. Equation 10.21,

$$\mathbf{P}_k\{\tau_n < \tau_0\} = \frac{1 - (q/p)^k}{1 - (q/p)^n}.$$

Thus,

$$\mathbf{P}_1\{\tau_0 = \infty\} \geq \mathbf{P}_1\left(\bigcap_{n=2}^{\infty}\{\tau_n < \tau_0\}\right) = \lim_n \frac{1-(q/p)}{1-(q/p)^n} = \frac{p-q}{p} > 0.$$

Since $\mathbf{P}_0\{\tau_0 = \infty\} = \mathbf{P}_1\{\tau_0 = \infty\}$, there is positive chance that the biased random walker never returns to her starting position.

This is also a consequence of the Strong Law of Large Numbers; see Exercise 18.1.

We have seen that the unbiased random walk (Example 18.1) and the biased random walk (Example 18.2) have quite different behavior. We make the following definition to describe this difference.

We define a state $x \in \Omega$ as *recurrent* if $\mathbf{P}_x\{\tau_x^+ < \infty\} = 1$. Otherwise, x is called *transient*.

{Prop:Communication}

Proposition 18.3. *Suppose that P is an irreducible transition matrix of a Markov chain (X_t). Define $G(x,y) := \mathbf{E}_x\left(\sum_{t=0}^{\infty}\mathbf{1}_{\{X_t=y\}}\right) = \sum_{t=0}^{\infty}P^t(x,y)$, the expected number of visits to y starting from x. The following are equivalent:*

{It:Grxx} (i) *$G(x,x) = \infty$, for some $x \in \Omega$.*
{It:Grxy} (ii) *$G(x,y) = \infty$ for all $x,y \in \Omega$.*
{It:Retxx} (iii) *$\mathbf{P}_x\{\tau_x^+ < \infty\} = 1$ for some $x \in \Omega$.*
{It:Retxy} (iv) *$\mathbf{P}_x\{\tau_y^+ < \infty\} = 1$ for all $x,y \in \Omega$.*

Proof. Every time the chain visits x, it has the same probability of eventually returning to x, independent of the past. Thus the number of visits to x is a geometric random variable with success probability $1 - \mathbf{P}_x\{\tau_x^+ < \infty\}$. It follows that (i) and (iii) are equivalent.

Suppose $G(x_0,x_0) = \infty$, and let $x,y \in \Omega$. By irreducibility, there exists r and s so that $P^r(x,x_0) > 0$ and $P^s(x_0,y) > 0$. Then

$$\begin{aligned}P^r(x,x_0)P^t(x_0,x_0)P^s(x_0,y) &= \mathbf{P}_x\{X_r = x_0,\ X_{r+t} = x_0,\ X_{r+t+s} = y\}\\ &\leq \mathbf{P}_x\{X_{r+t+s} = y\} = P^{r+t+s}(x,y).\end{aligned}$$

Thus,

{Eq:FinGF}

$$G(x,y) \geq \sum_{t=0}^{\infty}P^{r+t+s}(x,y) = P^r(x,x_0)P^s(x_0,y)\sum_{t=0}^{\infty}P^t(x_0,x_0). \tag{18.1}$$

Since $P^r(x,x_0)P^s(x_0,y) > 0$, Equation 18.1 shows that conditions (i) and (ii) are equivalent.

Suppose that $\mathbf{P}_{x_0}\{\tau_{x_0}^+ < \infty\} = 1$ for some $x_0 \in \Omega$, and let $x,y \in \Omega$.

If $\mathbf{P}_{x_0}\{\tau_x < \tau_{x_0}^+\} = 0$, then x is never hit when starting from x_0, contradicting the irreducibility of the chain. We have

$$0 = \mathbf{P}_{x_0}\{\tau_{x_0}^+ = \infty\} \geq \mathbf{P}_{x_0}\{\tau_x < \tau_{x_0}^+\}\mathbf{P}_x\{\tau_{x_0}^+ = \infty\}.$$

Since $\mathbf{P}_{x_0}\{\tau_x < \tau_{x_0}^+\} > 0$, it must be that $\mathbf{P}_x\{\tau_{x_0}^+ = \infty\} = 0$. Each time the chain visits x_0, it has positive probability of visiting y, independent of the past. Since the chain visits x_0 infinitely often, it will eventually visit y. To summarize: starting from x, the chain is certain to visit x_0, and starting from x_0, the chain is certain to visit y. Consequently, $\mathbf{P}_x\{\tau_y < \infty\} = 1$. We conclude that (iii) and (iv) are equivalent. ■

By Proposition 18.3, for an irreducible chain, a single state is recurrent if and only if all states are recurrent. For this reason, an irreducible chain can be classified as either recurrent or transient.

{Xmpl:SRWRecur1}

Example 18.4 (Simple random walk on $\mathbb{Z}$ revisited). Another proof that the simple random walker on $\mathbb{Z}$ discussed in Example 18.1 is recurrent uses Proposition 18.3.

When started at 0, the walk can return to 0 only at even times, with the probability of returning after $2t$ steps equal to $\mathbf{P}_0\{X_{2t} = 0\} = \binom{2t}{t}2^{-2t}$. By application of Stirling's formula (Equation B.10), $\mathbf{P}_0\{X_{2t} = 0\} \sim ct^{-1/2}$. Then

$$G(0,0) = \sum_{t=0}^{\infty} \mathbf{P}_0\{X_{2t} = 0\} = \infty,$$

so by Proposition 18.3 the chain is recurrent.

Example 18.5. The simple random walk on $\mathbb{Z}^2$ moves at each step by selecting each of the four neighboring locations with equal probability. Instead, consider at first the "corner" walk, which at each move adds with equal probability one of $\{(1,1),(1,-1),(-1,1),(-1,-1)\}$ to the current location. The advantage of this walk is that its coordinates are independent simple random walks on $\mathbb{Z}$. So

$$\mathbf{P}_{(0,0)}\{X_{2t} = (0,0)\} = \mathbf{P}_{(0,0)}\left\{X^1_{2t} = 0\right\}\mathbf{P}_{(0,0)}\left\{X^2_{2t} = 0\right\} \sim \frac{c}{n}.$$

Again by Proposition 18.3, the chain is recurrent. Now notice that the usual nearest-neighbor simple random walk is a rotation of the corner walk by $\pi/4$, so it is recurrent.

For random walks on infinite graphs, the electrical network theory of chapter 10 is very useful for deciding if a chain is recurrent.

18.2. Infinite Networks

For an infinite graph G containing vertex a, let $\{G_n\}$ be a collection of finite connected subgraphs containing a and satisfying $\cup_n G_n = G$. If all the vertices in $G \setminus G_n$ are replaced by a single vertex z_n, then

$$\mathcal{R}(a \leftrightarrow \infty) := \lim_{n\to\infty} \mathcal{R}\left(a \leftrightarrow z_n \text{ in } G_n \cup \{z_n\}\right).$$

Also, define $C(a \leftrightarrow \infty) := [\mathcal{R}(a \leftrightarrow \infty)]^{-1}$. By (10.16),

$$\mathbf{P}_a\{\tau_a^+ = \infty\} = \lim_{n\to\infty} \mathbf{P}_a\{\tau_{z_n} < \tau_a^+\} = \lim_{n\to\infty} \frac{C(a \leftrightarrow z_n)}{\pi(a)} = \frac{C(a \leftrightarrow \infty)}{\pi(a)}.$$

A flow on G from a to infinity is an antisymmetric edge function obeying the node law at all vertices except a. Thomson's Principle (Theorem 10.6) remains valid for infinite networks:

$$\mathcal{R}(a \leftrightarrow \infty) = \inf\{\mathcal{E}(\theta) \,:\, \theta \text{ a unit flow from } a \text{ to } \infty\}. \qquad (18.2)$$ {eq:tpi}

As a consequence, Rayleigh's Monotonicity Law (Theorem 10.7) also holds for infinite networks

The following summarizes the connection of resistance with recurrence.

{prop:tranrw}

Proposition 18.6. *Let $\langle G, \{c(e)\}\rangle$ be a network. The following are equivalent:*

(i) *The weighted random walk on the network is transient.*
(ii) *There is some node a with $C(a \leftrightarrow \infty) > 0$. (Equivalently, $\mathcal{R}(a \leftrightarrow \infty) < \infty$.)*
(iii) *There is a flow θ from some node a to infinity with $\|\theta\| > 0$ and $\mathcal{E}(\theta) < \infty$.*

In an infinite network $\langle G, \{c_e\}\rangle$, a version of Proposition 10.10 (the Nash-Williams inequality) is valid.

{Prop:NWI}

Proposition 18.7 (Nash-Williams). *If there exist disjoint edge-cutsets $\{\Pi_n\}$ that separate a from ∞ and satisfy*

$$\sum_n \left(\sum_{e \in \Pi_n} c(e) \right)^{-1} = \infty,$$

then the weighted random walk on $\langle G, \{c_e\}\rangle$ is recurrent.

Example 18.8 ($\mathbb{Z}^2$ is recurrent). Take $c(e) = 1$ for each edge of $G = \mathbb{Z}^2$ and consider the cutsets consisting of edges joining vertices in $\partial\Box_n$ to vertices in $\partial\Box_{n+1}$, where $\Box_n := [-n, n]^2$. Then by the Nash-Williams inequality,

$$\mathcal{R}(a \leftrightarrow \infty) \geq \sum_n \frac{1}{4(2n+1)} = \infty.$$

Thus, simple random walk on $\mathbb{Z}^2$ is recurrent. Moreover, we obtain a lower bound for the resistance from the center of a square $\Box_n = [-n, n]^2$ to its boundary:

$$\mathcal{R}(0 \leftrightarrow \partial\Box_n) \geq c \log n.$$

{ex:z3}

Example 18.9 ($\mathbb{Z}^3$ is transient). To each directed edge $\vec{e}$ in the lattice $\mathbb{Z}^3$, attach an orthogonal unit square $\Box_e$ intersecting $\vec{e}$ at its midpoint m_e. Define $\theta(\vec{e})$ to be the area of the radial projection of $\Box_e$ onto the sphere $\partial B(0, 1/4)$, taken with a positive sign if $\vec{e}$ points in the same direction as the radial vector from 0 to m_e, and with a negative sign otherwise. By considering a unit cube centered at each lattice point and projecting it to $\partial B(0, 1/4)$, we can easily verify that θ satisfies the node law at all vertices except the origin. Hence θ is a flow from 0 to ∞ in $\mathbb{Z}^3$. It is easy to bound its energy:

$$\mathcal{E}(\theta) \leq \sum_n C_1 n^2 \left(\frac{C_2}{n^2} \right)^2 < \infty.$$

By Proposition 18.6, $\mathbb{Z}^3$ is transient. This works for any $\mathbb{Z}^d$, $d \geq 3$. An analytic description of the same flow was given by T. Lyons (1983).

18.3. Positive Recurrence and Convergence

The convergence theorem as stated in Theorem 5.6 does not hold for all irreducible and aperiodic chains on infinite state-spaces. If the chain is transient, then by Proposition 18.3, $\sum_{t=0}^{\infty} \mathbf{P}_x\{X_t = y\} < \infty$ for all $x, y \in X$. This implies that for all $x, y \in \Omega$,

{Eq:ConvToZero}

$$\lim_{t\to\infty} \mathbf{P}_x\{X_t = y\} = 0. \tag{18.3}$$

That is, if there is a probability π on Ω so that $(\mu P^t)(x) \to \pi(x)$ for all $x \in \Omega$, then the chain must be recurrent.

However, recurrence is not sufficient. For example, the simple random walker of Example 18.4, a recurrent chain, also satisfies Equation 18.3. A condition stronger than recurrence is required.

{Example:RWNullRecurrer

Example 18.10. We have already seen that the simple random walker on $\mathbb{Z}$ is recurrent. Let $\alpha = \mathbf{E}_1(\tau_0)$. By conditioning on the first move of the walk,

$$\alpha = \frac{1}{2} + \frac{1}{2}[1 + \mathbf{E}_2(\tau_0)] = 1 + \alpha.$$

The last equality follows since the time to go from 2 to 0 equals the time to go from 2 to 1 plus the time to go from 1 to 0, and the time to go from 2 to 1 has the same distribution as the time to go from 1 to 0. There is no finite number α which satisfies this equation, so we must have $\alpha = \infty$. From this follows that $\mathbf{E}_0(\tau_0) = \infty$. Thus, although τ_0 is a finite random variable with probability one, it has infinite expectation.

A state x is called *positive recurrent* if $\mathbf{E}_x(\tau_x^+) < \infty$. As Example 18.10 shows, this property is strictly stronger than recurrence.

Proposition 18.11. *If (X_t) is a Markov chain with irreducible transition matrix P, then the following are equivalent:*

(i) $\mathbf{E}_x(\tau_x^+) < \infty$ *for some* $x \in \Omega$,
(ii) $\mathbf{E}_x(\tau_y^+) < \infty$ *for all* $x, y \in \Omega$.

Proof. Suppose that $\mathbf{E}_{x_0}(\tau_{x_0}^+) < \infty$. Define $\tau_{x_0,0}^+ := 0$ and

$$\tau_{x_0,k}^+ := \min\{t > \tau_{x_0,k-1}^+ \ : \ X_t = x_0\}, \quad k \geq 1.$$

Denote by L_k the time $\tau_{x_0,k}^+ - \tau_{x_0,k-1}^+$, the length of the k-th excursion from x_0. Because the chain starts anew at every visit to x, the random variables L_k form an i.i.d. sequence. In particular, $\mathbf{E}_{x_0}(L_k) = \mathbf{E}_{x_0}\{\tau_x^+\} < \infty$. By irreducibility, $\mathbf{P}_{x_0}\{\tau_y < \tau_{x_0}^+\} > 0$ and the chain has positive probability to hit y during each of these excursions. If T is the number of excursions from x_0 until the chain first hits y, then T is a geometric random variable and hence has finite mean. Also, if $\tau_{y\to x_0}$ is defined to be the first time after first visiting y that the chain returns to x_0, then when starting from x_0,

$$\tau_{y\to x_0} = \sum_{k=1}^{T} L_k.$$

Since the event that $\{T \geq k\} = \{T \leq k-1\}^c$ depends only on the chain up to $\tau^+_{x_0,k}$, it is independent of L_{k+1}. Thus by Exercise 7.10,

$$\mathbf{E}_{x_0}(\tau_{y\to x_0}) \leq \mathbf{E}_{x_0}(T)\mathbf{E}_{x_0}(L_k) < \infty.$$

Now let x and y be any two states in Ω. Note that

$$\infty > \mathbf{E}_{x_0}(\tau_{x\to x_0}) = \mathbf{E}_{x_0}(\tau^+_x) + \mathbf{E}_x(\tau^+_{x_0}).$$

Consequently, both $\mathbf{E}_{x_0}(\tau^+_x)$ and $\mathbf{E}_x(\tau^+_{x_0})$ are finite for any x. It follows that

$$\mathbf{E}_x(\tau^+_y) \leq \mathbf{E}_x(\tau^+_{x_0}) + \mathbf{E}_{x_0}(\tau^+_y) < \infty.$$

■

Thus if a single state of the chain is positive recurrent, all states are positive recurrent. We can therefore classify an irreducible chain as positive recurrent if one state, and hence all states, is positive recurrent. A chain which is recurrent but not positive recurrent is called *null recurrent*.

The following relates positive recurrence to the existence of a stationary distribution:

{Thm:PosRecStat}

Theorem 18.12. *An irreducible Markov chain with transition matrix P is positive recurrent if and only if there exists a probability distribution π on Ω so that $\pi = \pi P$.*

One direction of Theorem 18.12 is a consequence of the following Lemma together with Exercise 18.2.

{Lem:Kac}

Lemma 18.13 (Kac). *Let (X_t) be an irreducible Markov chain with transition matrix P. Suppose that there is a stationary distribution π solving $\pi = \pi P$. Then for any set $S \subset \Omega$,*

{Eq:Kac}

$$\sum_{x\in S} \pi(x)\mathbf{E}_x(\tau^+_S) = 1. \qquad (18.4)$$

In other words, the expected return time to S when starting at the stationary distribution conditioned on S is $\pi(S)^{-1}$.

Proof. Let (Y_t) be the reversed chain with transition matrix $\hat{P}$, defined in (3.30). First we show that both (X_t) and (Y_t) are recurrent. Define

$$\alpha(t) := \mathbf{P}_\pi\{X_t = x,\ X_s \neq x \text{ for } s > t\}.$$

By stationarity,

{Eq:AlphaStat}

$$\alpha(t) = \mathbf{P}_\pi\{X_t = x\}\mathbf{P}_x\{\tau^+_x = \infty\} = \pi(x)\mathbf{P}_x\{\tau^+_x = \infty\}. \qquad (18.5)$$

Since the events $\{X_t = x,\ X_s \neq x \text{ for } s > t\}$ are disjoint for distinct t,

$$\sum_{t=0}^{\infty} \alpha(t) \leq 1.$$

Since it is clear from (18.5) that $\alpha(t)$ does not depend on t, it must be that $\alpha(t) = 0$ for all t. Again from the identity (18.5), it follows that $\mathbf{P}_x\{\tau^+_x < \infty\} = 1$. The same argument works for reversed chain as well, so (Y_t) is also recurrent.

For $x \in S, y \in \Omega$ and $t \geq 0$, sum the identity

$$\pi(z_0)P(z_0, z_1)P(z_1, z_2)\cdots P(z_{t-1}, z_t) = \pi(z_t)\hat{P}(z_t, z_{t-1})\cdots \hat{P}(z_1, z_0)$$

over all sequences where $z_0 = x$, the states $z_1, \ldots, z_{t-1}$ are not in S, and $z_t = y$ to obtain

$$\pi(x)\mathbf{P}_x\{\tau_S^+ \geq t,\ X_t = y\} = \pi(y)\hat{\mathbf{P}}_y\{\tau_S^+ = k,\ Y_t = x\}. \tag{18.6}$$ {Eq:KacSS}

(We write $\hat{\mathbf{P}}$ for the probability measure corresponding to the reversed chain.) Summing over all $x \in S$, $y \in \Omega$, and $t \geq 0$ shows that

$$\sum_{x \in S} \pi(x) \sum_{t=1}^{\infty} \mathbf{P}_x\{\tau_S^+ \geq t\} = \hat{\mathbf{P}}_\pi\{\tau_S^+ < \infty\} = 1.$$

(The last equality follows from recurrence of (Y_t).) By Exercise 3.12(a), this simplifies to

$$\sum_{x \in S} \pi(x)\mathbf{E}_x\{\tau_S^+\} = 1. \tag{18.7}$$ {Eq:Kac1}

■

Proof of Theorem 18.12. That the chain is positive recurrent when a stationary distribution exists follows from Lemma 18.13.

The key fact needed to show that $\tilde{\pi}$ defined in Equation 3.18 can be normalized to yield a stationary distribution is that $\mathbf{E}_z(\tau_z^+) < \infty$, which holds now by positive recurrence. Thus the proof that a stationary distribution exists goes through as in the finite case. ■

{Thm:ConvInfinite}

Theorem 18.14. *Let P be an irreducible and aperiodic transition matrix for a Markov chain (X_t). If the chain is positive recurrent, then there is a unique probability distribution π on Ω so that $\pi = \pi P$ and for all $x \in \Omega$,*

$$\lim_{t \to \infty} \|P^t(x, \cdot) - \pi\|_{TV} = 0. \tag{18.8}$$ {Eq:InfConv}

Proof. The existence of π solving $\pi = \pi P$ is one direction of Theorem 18.12.

We now show that for any two states x and y we can couple together the chain started from x with the chain started from y so that the two chains eventually meet with probability one.

Consider the chain on $\Omega \times \Omega$ with transition matrix

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

This chain makes independent moves in the two coordinates, each according to the matrix P. Aperiodicity implies that this chain is irreducible (see Exercise 18.5). If (X_t, Y_t) is a chain started with product distribution $\mu \times \nu$ and run with transition matrix $\tilde{P}$, then (X_t) is a Markov chain with transition matrix P and initial distribution μ, and (Y_t) is a Markov chain with transition matrix P and initial distribution ν.

Note that

$$(\pi \times \pi)\tilde{P}(z,w) = \sum_{(x,y)\in\Omega\times\Omega} (\pi \times \pi)(x,y)P(x,z)P(y,w)$$
$$= \sum_{x\in X} \pi(x)P(x,z) \sum_{y\in Y} \pi(y)P(y,w).$$

Since $\pi = \pi P$, the right-hand side equals $\pi(z)\pi(w) = (\pi \times \pi)(z,w)$. Thus $\pi \times \pi$ is a stationary distribution for $\tilde{P}$. By Theorem 18.12, the chain (X_t, Y_t) is positive recurrent. In particular, for any fixed x_0, if

$$\tau := \min\{t > 0 \,:\, (X_t, Y_t) = (x_0, x_0)\},$$

then

{Eq:FinHit}
$$\mathbf{P}_{x,y}\{\tau < \infty\} = 1 \quad \text{for all } x, y \in \Omega. \tag{18.9}$$

To obtain Equation 18.8, note that if the chain (X_t, Y_t) is started with the distribution $\delta_x \times \pi$, then for fixed t the pair of random variables X_t and Y_t is a coupling of $P^t(x,\cdot)$ with π. Thus

{Eq:CoupHit}
$$\left\|P^t(x,\cdot) - \pi\right\|_{TV} \leq \mathbf{P}_{\delta_x\times\pi}\{X_t \neq Y_t\} \leq \mathbf{P}_{\delta_x\times\pi}\{\tau > t\}. \tag{18.10}$$

From (18.9),

$$\mathbf{P}_{\delta_x\times\pi}\{\tau > t\} = \sum_{y\in\Omega} \pi(y)\mathbf{P}_{x,y}\{\tau > t\} = 1.$$

This and (18.10) imply Equation 18.8.

■

{Example:RRW}

Example 18.15. Consider a nearest-neighbor random walk on $\mathbb{Z}^+$ which moves up with probability p and down with probability q. If the walk is at 0, it remains at 0 with probability q. Assume that $q > p$.

The equation $\pi = \pi P$ reads as

$$\pi(0) = q\pi(1) + q\pi(0)$$
$$\pi(k) = p\pi(k-1) + q\pi(k+1).$$

Solving, $\pi(1) = \pi(0)(p/q)$ and working up the ladder,

$$\pi(k) = (p/q)^k \pi(0)$$

π can be normalized to be a probability distribution, in which case $\pi(k) = (p/q)^k(1 - p/q)$. Since there is a solution to $\pi P = \pi$ which is a probability distribution, the chain is positive recurrent.

If a solution can be found to the detailed balance equations,

$$\pi(x)P(x,y) = \pi(y)P(y,x), \quad x, y \in \Omega,$$

then provided π is a probability distribution, the chain is positive recurrent.

Example 18.16 (Birth-and-Death Chains). A *birth-and-death* chain on $\{0, 1, \ldots, \}$ is a nearest-neighbor chain which moves up when at k with probability p_k and down with probability $q_k = 1 - p_k$. The detailed balance equations are, for $j \geq 1$,

$$\pi(j)p_j = \pi(j+1)q_j.$$

Thus $\pi(j+1)/\pi(j) = p_j/q_j$ and so

$$\pi(k) = \pi(0)\prod_{j=0}^{k-1}\frac{\pi(j+1)}{\pi(j)} = \pi(0)\prod_{j=0}^{k-1}\frac{p_j}{q_j}.$$

This can be made into a probability distribution provided that

$$\sum_{k=1}^{\infty}\prod_{j=0}^{k-1}\frac{p_j}{q_j} < \infty, \tag{18.11}$$ {Eq:BDSum}

in which case we take $\pi(0)^{-1}$ to equal this sum.

If the sum in (18.11) is finite, the chain is positive recurrent.

18.4. Problems

{Xsz:BRWSLN}

Exercise 18.1. Use the Strong Law of Large numbers to give a proof that the biased random walk in Example 18.2 is transient. [Solution]

{Exercise:PiPos}

Exercise 18.2. Suppose that P is irreducible. Show that if $\pi = \pi P$ for a probability distribution π, then $\pi(x) > 0$ for all $x \in \Omega$. [Solution]

{ex:fuzz}

Exercise 18.3. Fix $k > 1$. Define the *k-fuzz* of an undirected graph $G = (V, E)$ as the graph $G_k = (V, E_k)$ where for any two distinct vertices $v, w \in V$, the edge $\{v, w\}$ is in E_k if and only if there is a path of at most k edges in E connecting v to w. Show that for G with bounded degrees, G is transient if and only if G_k is transient.

A solution can be found in Doyle and Snell (1984, section 8.4).

{Exercise:SubGRec}

Exercise 18.4. Show that any subgraph of a recurrent graph must be recurrent. [Solution]

{Exercise:ProdIred}

Exercise 18.5. Let P be an irreducible and aperiodic transition matrix on Ω. Let $\tilde{P}$ to be the matrix on $\Omega \times \Omega$ defined by

$$\tilde{P}((x,y),\,(z,w)) = P(x,z)P(y,z), \quad (x,y) \in \Omega\times\Omega,\ (z,w) \in \Omega\times\Omega.$$

Show that $\tilde{P}$ is irreducible. [Solution]

{Exercise:FIFO}

Exercise 18.6. Consider the discrete-time single server FIFO (first in, first out) queue: At every step, if there is a customer waiting, exactly one of the following happens:

(1) a new customer arrives (with probability α), or {It:arrive}
(2) an existing customer is served (with probability $\beta = 1 - \alpha$), {It:served}

If there are no customers waiting, then (1) still has probability α, but (2) is replaced by "nothing happens". Let X_t be the number of customers in the queue at time t.

Show that (X_t) is

(a) positive recurrent if $\alpha < \beta$,
(b) null recurrent if $\alpha = \beta$,
(c) transient if $\alpha > \beta$.

[Solution]

{Exercise:FIFO2}

Exercise 18.7. Consider the same set-up as Exercise 18.6. In the positive recurrent case, determine the stationary distribution π and the π-expectation of the time T from the arrival of a customer until he is served. [Solution]

Remark. In communication theory one talks of *packets* instead of customers.

{Exer:FiniteTransRec}

Exercise 18.8. Consider a not-necessarily-irreducible Markov chain on a *finite* state space, Ω. Recall the communication classes defined on Section 3.7, and the partial order $\rightarrow$ on communication classes defined in Exercise 3.24.

Prove that a state $x \in X$ is recurrent if and only if $[x]$ is a maximal element in this partial order.

{Exer:Liouville1D}

Exercise 18.9. Let P be the transition matrix for simple random walk on $\mathbb{Z}$. Show that the walk is not positive recurrent by showing there are no probability distributions π on $\mathbb{Z}$ satisfying $\pi P = \pi$. [Solution]

CHAPTER 19

Martingales

{Ch:MG}

19.1. Definition and Examples

Let $(Y_t)_{t=0}^{\infty}$ be a sequence of random variables. In what follows, (Y_t) will serve as a basic source of randomness. For example, (Y_t) could be an i.i.d. sequence of $\{-1, +1\}$-valued random variables, or a Markov chain. We make no assumptions about the distribution of this sequence.

A *martingale* with respect to (Y_t) is a sequence of random variables (M_t) satisfying the following: {it:integrable}

(i) $\mathbf{E}(M_t) < \infty$ for all t; {it:adapted}
(ii) M_t is *adapted* to (Y_t), meaning for each t there exists a function g_t so that $M_t = g_t(Y_0, \ldots, Y_t)$ for all t; {it:fair}
(iii) $\mathbf{E}(M_{t+1} \mid Y_0, \ldots, Y_t) = M_t$.

Condition (ii) says that M_t is determined by $(Y_1, \ldots, Y_t)$, the underlying randomness up to and including time t. If we assume that an observer at time t knows the random vector $(Y_0, \ldots, Y_t)$, then she can compute the value of M_t from this information. In particular, she does not need any any of the future variables $(Y_s)_{s>t}$.

Condition (iii) says that given the data $(Y_1, \ldots, Y_t)$, the best prediction for M_{t+1} is M_t.

{Example:RWMart}

Example 19.1. The familiar unbiased random walk is a martingale.

Let $(Y_s)_{s=1}^{\infty}$ be a sequence of independent random variables with $\mathbf{E}(Y_s) = 0$ for all s, and $M_t := \sum_{s=1}^{t} Y_s$.

The conditions (i) and (ii) are manifest, and (iii) also holds:

$$\begin{aligned}\mathbf{E}(M_{t+1} \mid Y_0, \ldots, Y_t) &= \mathbf{E}(Y_{t+1} + M_t \mid Y_0, \ldots, Y_t) \\ &= \mathbf{E}(Y_{t+1} \mid Y_0, \ldots, Y_t) + M_t = M_t.\end{aligned}$$

The penultimate equality follows since M_t is a function of $(Y_0, \ldots, Y_t)$, and the last equality follows since Y_{t+1} is independent of $(Y_0, \ldots, Y_t)$ and has $\mathbf{E}(Y_{t+1}) = 0$.

In the previous example, the *increments* $\Delta M_t := M_{t+1} - M_t$ form an independent sequence with $\mathbf{E}(\Delta M_t) = 0$. For a general martingale, the increments also have mean zero, and although not necessarily independent, they are uncorrelated: for $s < t$,

$$\begin{aligned}\mathbf{E}(\Delta M_t \Delta M_s) &= \mathbf{E}\left(\mathbf{E}\left(\Delta M_t \Delta M_s \mid Y_0, Y_1, \ldots, Y_t\right)\right) \\ &= \mathbf{E}\left(\Delta M_s \mathbf{E}\left(\Delta M_t \mid Y_0, Y_1, \ldots, Y_t\right)\right) \qquad (19.1) \\ &= 0.\end{aligned}$$

{Eq:UncorrInc}

We have used here the fact, immediate from condition (iii) in the definition of a martingale, that

$$\mathbf{E}(\Delta M_t \mid Y_0, \ldots, Y_t) = 0, \tag{19.2}$$

{eq:mginc}

which is stronger than the statement that $\mathbf{E}(\Delta M_t) = 0$.

To summarize, martingales are very similar to sums of i.i.d. random variables: A martingale (M_t) can be written as

{eq:inc}

$$M_t = M_0 + \sum_{s=0}^{t-1} \Delta M_s \tag{19.3}$$

where the elements of the sequence $(\Delta M_s)_{s=0}^{\infty}$ are uncorrelated and satisfy (19.2).

Example 19.2. Let (Y_t) be a random walk which moves up one unit with probability p, and down one unit with probability $q = 1 - p$, where $p \neq 1/2$. In other words, given $Y_0, \ldots, Y_t$,

$$\Delta Y_t := Y_{t+1} - Y_t = \begin{cases} 1 & \text{with probability } p \\ -1 & \text{with probability } q. \end{cases}$$

If $M_t := (q/p)^{Y_t}$, then (M_t) is a martingale with respect to (Y_t). Condition (ii) is clear, and

$$\begin{aligned} \mathbf{E}\left[(q/p)^{Y_{t+1}} \,\middle|\, Y_0 = y_0, \ldots, Y_t = y_t\right] &= \mathbf{E}\left[(q/p)^{y_t}(q/p)^{Y_{t+1}-Y_t} \,\middle|\, Y_0 = y_0, \ldots, Y_t = y_t\right] \\ &= (q/p)^{y_t}\left[p(q/p) + q(q/p)^{-1}\right] \\ &= (q/p)^{y_t}. \end{aligned}$$

Example 19.3. Let (Y_t) be as in the previous example. Let $\mu := p - q$, and $M_t := Y_t - \mu t$. Then

$$\begin{aligned} \mathbf{E}(M_{t+1} - M_t \mid Y_0, \ldots, Y_t) &= p - q - \mu \\ &= 0, \end{aligned}$$

so (M_t) is a martingale.

A sequence of random variables (A_t) is called *previsible* if for each t there is a function f_t so that $A_t = f_t(Y_0, \ldots, Y_{t-1})$. The random variable A_t is determined by what has happened strictly before time t.

Suppose that (M_t) is a martingale with respect to (Y_t), and (A_t) is a previsible sequence. Imagine that a gambler can wager on a sequence of games so that he receives $M_t - M_{t-1}$ for each unit bet on the t-th game. The interpretation of the martingale property $E(M_t - M_{t-1} \mid Y_0, \ldots, Y_t) = 0$ is that the games are fair. Let A_t be the amount wagered on the t-th game; the fact that the player sizes his bet based only on the outcomes of previous games forces (A_t) to be a previsible sequence. At time t, the gambler's fortune is

{eq:winnings}

$$F_t = M_0 + \sum_{s=0}^{t-1} A_{s+1}(M_{s+1} - M_s). \tag{19.4}$$

Is it possible, by a suitably clever choice of bets $(A_1, A_2, \ldots)$, to generate an advantage for the player? By this, we mean is it possible that $E(F_t) > 0$ for some t? Many gamblers say so. Unfortunately, they are wrong! The next theorem proves it.

Define for a martingale (M_t) and a previsible sequence (A_t),

$$(A \circ M)_t := M_0 + \sum_{s=0}^{t-1} A_{s+1}(M_{s+1} - M_s).$$

{thm.discstochint}

Theorem 19.4. *For any previsible sequence (A_t), the sequence of random variables $(A \circ M)_t$ is a martingale.*

Proof.

$$\mathbf{E}\left((A \circ M)_{t+1} - (A \circ M)_t \mid Y_0, \ldots, Y_t\right) = \mathbf{E}(A_{t+1}(M_{t+1} - M_t) \mid Y_0, \ldots, Y_t).$$

Since A_{t+1} is a function of $Y_0, \ldots, Y_t$, the right-hand side equals

$$A_{t+1}\mathbf{E}(M_{t+1} - M_t \mid Y_0, \ldots, Y_t) = 0.$$

■

Recall from Section 7.2.1 that a stopping time is a random variable τ with values in $\{0, 1, \ldots\} \cup \{\infty\}$ so that the event $\{\tau = t\}$ is determined by the random variables $Y_0, \ldots, Y_t$. More precisely, the sequence $(\mathbf{1}_{\{\tau=t\}})$ is adapted to the sequence (Y_t).

For a martingale, $\mathbf{E}(M_t) = \mathbf{E}(M_0)$ for all *fixed* times t. Does this remain valid if we replace t by a random time? In particular, for stopping times τ, is $\mathbf{E}(M_\tau) = \mathbf{E}(M_0)$? Under some additional conditions, the answer is “yes”. However, these conditions cannot be ignored, as it is false in general.

Example 19.5. Taking (Y_s) to be the i.i.d. sequence with

$$\mathbf{P}\{Y_1 = +1\} = \mathbf{P}\{Y_1 = -1\} = \frac{1}{2},$$

in Example 19.1, the partial sum $M_t := \sum_{s=1}^{t} Y_s$ is a martingale. The first-passage time to 1,

$$\tau = \min\{t \,:\, M_t = +1\}$$

is a stopping time, and clearly

$$\mathbf{E}(M_\tau) = 1 \neq \mathbf{E}(M_0).$$

Note that if τ is a stopping time, then so is $\tau \wedge t$. for any fixed t.

{thm:ost1}

Theorem 19.6. *Let τ be a stopping time and (M_t) a martingale. Then $(M_{t\wedge\tau})$ is a martingale. Consequently, $\mathbf{E}(M_{t\wedge\tau}) = \mathbf{E}(M_0)$.*

{cor:ost2}

Corollary 19.7. *Let (M_t) be a martingale and τ a stopping time so that $|M_{t\wedge\tau}| \leq K$ for all t, where K is a fixed number. Then $\mathbf{E}(M_\tau) = \mathbf{E}(M_0)$.*

Proof of Theorem 19.6. Let $A_t = \mathbf{1}_{\{\tau>t\}}$. Then

$$A_t = 1 - \mathbf{1}_{\{\tau \leq t-1\}} = 1 - \sum_{s=1}^{t-1} \mathbf{1}_{\{\tau=t\}},$$

and since τ is a stopping time, A_t can be written as a function of $Y_0, \ldots, Y_{t-1}$. Thus (A_t) is previsible. Check that

$$(A \circ M)_t = M_{t \wedge \tau} - M_0.$$

Thus $M_{t\wedge\tau} - M_0$ is a martingale. The reader should check that $M_{t\wedge\tau} - M_0 + M_0 = M_{t\wedge\tau}$ is still a martingale. ■

Proof of Corollary 19.7. Since $(M_{\tau\wedge t})$ is a martingale, $\mathbf{E}\,(M_{\tau\wedge t}) = \mathbf{E}\,(M_0)$. Thus

$$\lim_{t\to\infty} \mathbf{E}(M_{\tau\wedge t}) = \mathbf{E}(M_0).$$

By Proposition B.5, we are allowed to take a limit inside the expectation and conclude that $\mathbf{E}(M_\tau) = \mathbf{E}(M_0)$. ■

{cor:ost3}

Corollary 19.8. *Let (M_t) be a martingale with bounded increments, that is $|M_{t+1} - M_t| \leq B$ for all t, where B is a non-random constant. Suppose that τ is a stopping time with $\mathbf{E}(\tau) < \infty$. Then $\mathbf{E}(M_\tau) = \mathbf{E}(M_0)$.*

Proof. Note that

$$|M_{\tau\wedge n}| = \left|\sum_{s=1}^{\tau\wedge n}(M_s - M_{s-1}) + M_0\right| \leq \sum_{s=1}^{\tau\wedge n} |M_s - M_{s-1}| + |M_0| \leq B\tau + |M_0|.$$

Since $\mathbf{E}(B\tau + |M_0|) < \infty$, by the Dominated Convergence Theorem (Proposition B.5) and Theorem 19.6,

$$\mathbf{E}(M_0) = \lim_{n\to\infty} \mathbf{E}(M_{\tau\wedge n}) = \mathbf{E}(M_\tau).$$

■

Example 19.9. Let $Y_0 \equiv 0$, and let $Y_1, Y_2, \ldots$ be a sequence of independent and identically distributed random variables with

$$\mathbf{P}\{Y_s = 1\} = \mathbf{P}\{Y_s = -1\} = \frac{1}{2}.$$

$S_t := \sum_{s=0}^{t} Y_s$ is a martingale. Let $B_1 \equiv 1$, and for $t > 1$, let

$$B_t = \begin{cases} 2^t & \text{if } Y_1 = Y_2 = \cdots = Y_{t-1} = -1 \\ 0 & \text{if } Y_s = 1 \text{ for some } s < t. \end{cases}$$

Thus, provided we have not won a single previous game, we bet 2^t, and as soon as we win, we stop playing. If τ is the first time that we win, τ is a stopping time.

$$M_t := (B \circ S)_t = \begin{cases} 0 & \text{if } t = 0, \\ -2^{(t-1)} & \text{if } 1 \leq t < \tau, \\ 1 & \text{if } t \geq \tau. \end{cases}$$

Since we are assured that $Y_s = 1$ for some s eventually, $\tau < \infty$ and $M_\tau = 1$. Thus $\mathbf{E}(M_\tau) = 1$. But $\mathbf{E}(M_0) = 0$, and (M_t) is a martingale! By doubling our bets every time we lose, we have assured ourselves of a profit. This at first glance seems to contradict Corollary 19.7. But notice that the condition $|M_{\tau\wedge t}| < K$ is not satisfied, so we cannot apply the Corollary.

19.2. Applications

19.2.1. Gambler's Ruin. Let (Y_t) be a random walk, and let $\alpha(x) = \mathbf{P}_x\{\tau_0 < \tau_N\}$, where $0 \leq x \leq N$. Suppose that $p \neq q$. We have seen before that $M_t := (q/p)^{Y_t}$ is a martingale. Let $\tau := \tau_0 \wedge \tau_N$ be the first time the walk hits either 0 or N. Then τ is a stopping time.

Since $M_{\tau \wedge t}$ is bounded, we can apply Corollary 19.7 to get

$$\mathbf{E}_x\left((q/p)^{Y_\tau}\right) = (q/p)^x.$$

We can break up the expectation above to get

$$\mathbf{E}_x\left((q/p)^{Y_\tau}\right) = \alpha(x) + (q/p)^N(1 - \alpha(x)).$$

Combining these two equations and solving for $\alpha(x)$ yields

$$\alpha(x) = \frac{(q/p)^x - (q/p)^N}{1 - (q/p)^N}.$$

In the case where $p = q = \frac{1}{2}$, we can apply the same argument to get that $\alpha(x) = 1 - (x/N)$.

Now consider again the unbiased random walk. Notice that

$$\begin{aligned}\mathbf{E}(Y_{t+1}^2 - Y_t^2 \mid Y_0, \ldots, Y_t) &= (Y_t + 1)^2\frac{1}{2} + (Y_t - 1)^2\frac{1}{2} - Y_t^2 \\ &= 1.\end{aligned}$$

Thus $M_t := S_t^2 - t$ is a martingale. By Theorem 19.6 we have that

$$\mathbf{E}_x(S_{t\wedge\tau}^2) = \mathbf{E}_x(\tau \wedge t).$$

Now since $S_{t\wedge\tau}^2$ is bounded by N^2 for all n, if we take the limit as $t \to \infty$ on the left-hand side above, we can take it inside the expectation. Also, $T \wedge t$ does not decrease as t increases, so we are allowed to take the limit inside the expectation. Thus

$$\mathbf{E}_x(S_\tau^2) - x^2 = \mathbf{E}_x(T).$$

Now conditioning on whether $\tau = \tau_0$ or $\tau = \tau_N$ yields

$$(1 - \alpha(x))N^2 - x^2 = \mathbf{E}_x(T).$$

Hence,

$$\mathbf{E}_x(T) = x(N - x).$$

19.2.2. Waiting times for patterns in coin tossing. Consider a sequence of independent fair coin tosses, $X_1, X_2, \ldots$, and define

$$\tau_{HTH} = \min\{t : X_{t-2}X_{t-1}X_t = HTH\}.$$

We wish to determine $\mathbf{E}(\tau_{HTH})$.

Gamblers are allowed to place bets on each individual coin toss. On each bet, the gambler is allowed to pay \$$k$ dollars, and then either wins \$$2k$ dollars or \$0 dollars.

We suppose that at each unit of time until the word HTH first appears, a new gambler enters, and employs the following strategy: On his first bet, he wagers \$1 on the outcome H. If he looses, he stops. If he wins and the sequence HTH still has not yet appeared, he wagers his payoff of \$2 on T. Again, if he looses, he stops playing. As before, if he wins and the sequence HTH has yet to occur, he takes his payoff (now \$4) and wagers on H. He then stops playing.

We describe the situation a bit more precisely: Let (B_t) be an i.i.d. sequence of $\{0,1\}$-valued random variables, with $\mathbf{E}(B_t) = 1/2$, and then define $M_t = \sum_{s=1}^{t}(2B_s - 1)$. Clearly (M_t) is a martingale. Let $\tau_{101} = \min\{t : X_{t-2}X_{t-1}X_t = 101\}$, and define

$$A_t^s = \begin{cases} 1 & t = s, \\ -2 & t = s+1, \tau > t, \\ 4 & t = s+2, \tau > t, \\ 0 & \text{otherwise.} \end{cases}$$

Then $(A^s \circ M)_t$ is the profit of the sth gambler at the tth game. By Theorem 19.4, $(A^s \circ M)$ is a martingales, and by Corollary 19.8,

$$\mathbf{E}((A^s \circ M)_\tau) = 0.$$

Suppose that $\tau_{HTH} = t$. The gambler who started at t is paid \$2, the gambler who started at $t-2$ is paid \$8, and every gambler has paid an initial \$1 wager. Since the game is fair, we must have the expected winnings is 0, so

$$10 - \mathbf{E}(\tau_{HTH}) = 0.$$

That is, $\mathbf{E}(\tau_{HTH}) = 10$.

It is (sometimes) suprising to the non-expert that the expected time to see HHH is longer than HTH. Running the same arguments as above, the bettor entering at time $\tau - 2$ is paid \$8, the bettor entering at time $\tau - 1$ is paid \$4, and the bettor entering at τ is paid \$2. Again, the total outlay is \$$\tau$, and fairness requires that $\mathbf{E}(\tau) = 8 + 4 + 2 = 14$.

19.3. Problems

{Exer:CondGR}

EXERCISE 19.1. Let (X_t) be the simple random walk on $\mathbb{Z}$.

(a) Show that $M_t = X_t^3 - 3tX_t$ is a martingale.
(b) If τ is the expected time until the walker hits either 0 or n, find $\mathbf{E}_k(\tau \mid X_\tau = n)$. (Here, $0 \le k \le n$.)

EXERCISE 19.2. Let (X_t) be a Markov chain with transition matrix P. A function h on Ω is called *harmonic* with respect to P if $Ph = h$. Show that if h is harmonic, then the sequence (M_t) is a martingale, where $M_t = h(X_t)$.

CHAPTER 20

Coupling from the Past

by James G. Propp and David B. Wilson

This chapter is based in part on the expository article “Coupling from the Past: a User’s Guide,” which appeared in *Microsurveys in Discrete Probability*, volume 41 of the *DIMACS Series in Discrete Mathematics and Computer Science*, published by the AMS.

20.1. Introduction

In Markov chain Monte Carlo studies, one attempts to sample from a probability distribution π by running a Markov chain whose unique steady-state distribution is π. Ideally, one has proved a theorem that guarantees that the time for which one plans to run the chain is substantially greater than the mixing time of the chain, so that the distribution $\tilde{\pi}$ that one’s procedure actually samples from is known to be close to the desired π in variation distance. More often, one merely hopes that this is the case, and the possibility that one’s samples are contaminated with substantial initialization bias cannot be ruled out with complete confidence.

The “coupling from the past” procedure introduced in Propp and Wilson (1996) provides one way of getting around this problem. Where it is applicable, this method determines on its own how long to run, and delivers samples that are governed by π itself, rather than $\tilde{\pi}$. Many researchers have found ways to apply the basic idea in a wide variety of settings (see `http://www.dbwilson.com/exact/` for pointers to this research). Our aim here is to explain the basic method, and to give a sampling of some of its varied applications.

It is worth stressing at the outset that CFTP is especially valuable as an alternative to standard Markov chain Monte Carlo when one is working with Markov chains for which one suspects, but has not proved, that rapid mixing occurs. In such cases, the availability of CFTP makes it less urgent that theoreticians obtain bounds on the mixing time, since CFTP (unlike Markov chain Monte Carlo) cleanly separates the issue of efficiency from the issue of quality of output. That is to say, one’s samples are guaranteed to be uncontaminated by initialization bias, regardless of how quickly or slowly they are generated.

Before proceeding we mention that there are other algorithms that may be used for generating perfect samples from the stationary distribution of a Markov chain, including Fill’s algorithm (Fill, 1998, Fill, Machida, Murdoch, and Rosenthal, 2000), “read-once CFTP” (Wilson, 2000), and the “randomness recycler” (Fill and Huber, 2000). Each of these has its merits, but since CFTP is conceptually the simplest of these, it is the one that we shall focus our attention on here.

As an historical aside, we mention that the conceptual ingredients of CFTP were in the air even before the versatility of the method was made clear in Propp and Wilson (1996). Precursors include Letac (1986), Thorisson (1988), and Borovkov and Foss (1992). Even back in the 1970's, one can find foreshadowings in the work of Ted Harris (on the contact process, the exclusion model, random stirrings, and coalescing and annihilating random walks), David Griffeath (on additive and cancellative interacting particle systems), and Richard Arratia (on coalescing Brownian motion). One can even see traces of the idea in the work of Loynes (1962) forty-five years ago. See also the survey Diaconis and Freedman (1999).

20.2. Monotone CFTP

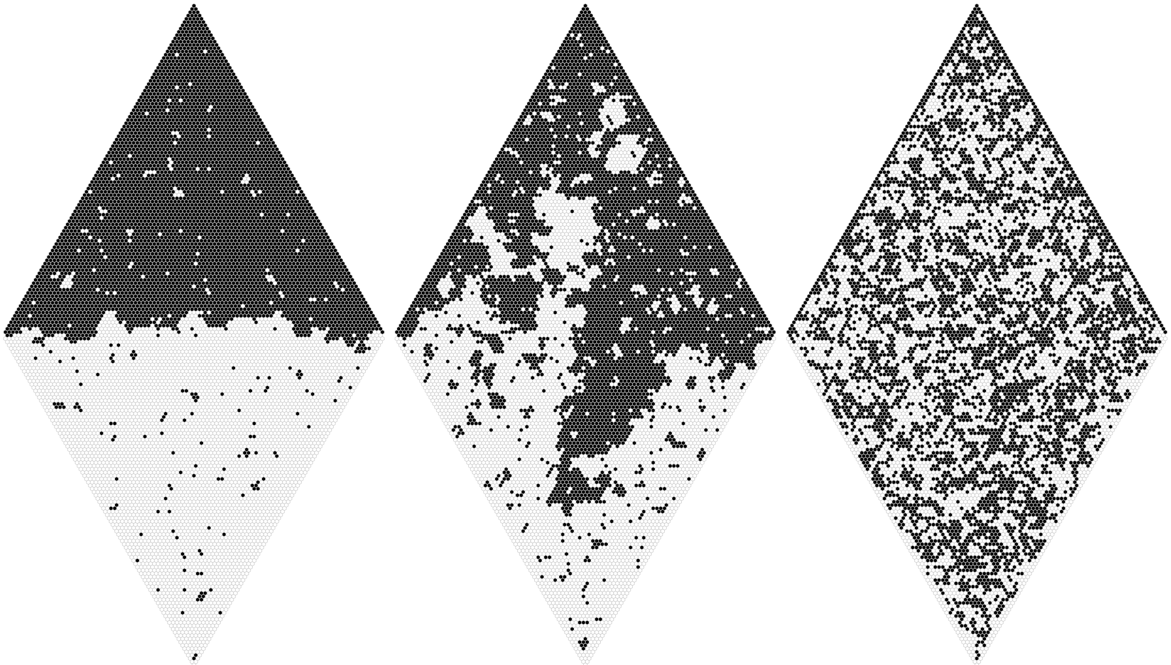
The basic idea of coupling from the past is quite simple. Suppose that there is an ergodic Markov chain that has been running either forever or for a very long time, long enough for the Markov chain to have reached its steady-state distribution. So the state that the Markov chain is currently in is a sample from the stationary distribution. If we can figure out what that state is, by looking at the recent randomizing operations of the Markov chain, then we have a sample from its stationary distribution. To illustrate these ideas, we show how to apply them to the Ising model of magnetism.

Recall that an Ising system consists of a collection of n interacting spins, possibly in the presence of an external field. Each spin may be aligned up or down. Spins that are close to each other prefer to be aligned in the same direction, and all spins prefer to be aligned with the external magnetic field (which sometimes varies from site to site). These preferences are quantified in the total energy E of the system

$$E(\sigma) = -\sum_{i<j} \alpha_{i,j}\sigma_i\sigma_j - \sum_i B_i\sigma_i,$$

where B_i is the strength of the external field as measured at site i, σ_i is 1 if spin i is aligned up and -1 if it is aligned down, and $\alpha_{i,j} \geq 0$ represents the interaction strength between magnets i and j. The probability of a given spin configuration is given by $Z^{-1}\exp[-E(\sigma)/T]$ where T is the "temperature," and Z is a normalizing constant that makes the probabilities add up to 1. Often the n spins are arranged in a 2D or 3D lattice, and $\alpha_{i,j}$ is 1 if spins i and j are adjacent in the lattice, and 0 otherwise. The Ising model has been used to model certain substances such as crystals of $FeCl_2$ and $FeCO_3$, and certain phases of carbon dioxide, xenon, and brass — see Baxter (1982) for further background.

We may use the single-site heat bath algorithm, also known as Glauber dynamics, to sample Ising spin configurations. A single move of the heat-bath algorithm may be summarized by a pair of numbers (i, u), where i represents a spin location (say that i is a uniformly random spin), and u is a uniformly random real number between 0 and 1. The heat-bath algorithm randomizes the alignment of spin i, holding all of the remaining magnets fixed, and uses the number u when



{ising} FIGURE 20.1. The Ising model at three different temperatures. Here the spins lie at the vertices of the triangular lattice and are shown as black or white hexagons. The spins along the upper boundaries were forced to be black and the spins along lower boundaries were forced to be white (using an infinite magnetic field on these boundary spins).

deciding whether the new spin should be up or down. There are two possible choices for the next state, denoted by $\sigma_\uparrow$ and $\sigma_\downarrow$. We have $\Pr[\sigma_\uparrow]/\Pr[\sigma_\downarrow] = e^{-(E(\sigma_\uparrow)-E(\sigma_\downarrow))/T} = e^{-(\Delta E)/T}$. The update rule is that the new spin at site i is ↑ if $u < \Pr[\sigma_\uparrow]/(\Pr[\sigma_\uparrow] + \Pr[\sigma_\downarrow]$, and otherwise the new spin is ↓. It is easy to check that this defines an ergodic Markov chain with the desired stationary distribution.

Recall our supposition that the randomizing process, in this case the single-site heat bath, has been running for all time. Suppose that someone has recorded all the randomizing operations of the heat bath up until the present time. They have not recorded what the actual spin configurations or Markov chain transitions are, but merely which sites were updated, and which random number was used to update the spin at the given site. Given this recorded information, our goal is to determine the state of the Markov chain at the present time (time 0), since, as we have already determined, this state is a sample from the stationary distribution of the Markov chain.

To determine the state at time 0, we make use of a natural partial order with which the Ising model is equipped: we say that two spin-configurations σ and τ satisfy $\sigma \leq \tau$ when each spin-up site in σ is also spin-up in τ. Notice that if we update both σ and τ with the same heat-bath update operation (i, u), then because site i has at least as many spin-up neighbors in τ as it does in σ, and because of our assumption that the $\alpha_{i,j}$'s are nonnegative, we have $\Pr[\tau_\uparrow]/\Pr[\tau_\downarrow] \geq \Pr[\sigma_\uparrow]/\Pr[\sigma_\downarrow]$, and so the updated states σ' and τ' also satisfy $\sigma' \leq \tau'$. (We say that the randomizing operation respects the partial order $\leq$.) Notice also that the partial order $\leq$ has a maximum state $\hat{1}$, which is spin-up at every site, and a minimum state $\hat{0}$, which is spin-down at every site.

This partial order enables us to obtain upper and lower bounds on the state at the present time. We can look at last T randomizing operations, figure out what would happen if the Markov chain were in state $\hat{1}$ at time $-T$, and determine where it would be at time 0. Since the Markov chain is guaranteed to be in a state which is $\leq \hat{1}$ at time $-T$, and the randomizing operations respect the partial order, we obtain an upper bound on the state at time 0. Similarly we can obtain a lower bound on the state at time 0 by applying the last T randomizing operations to the state $\hat{0}$. It could be that we are lucky and the upper and lower bounds are equal, in which case we have determined the state at time 0. If we are not so lucky, we could look further back in time, say at the last $2T$ randomizing operations, and obtain better upper and lower bounds on the state at the present time. So long as the upper and lower bounds do not coincide, we can keep looking further and further back in time (see Figure 20.2). Because the Markov chain is ergodic, when it is started in $\hat{1}$ and T is large enough, there is some positive chance that it will reach $\hat{0}$, at which time the upper and lower bounds are guaranteed to coincide. In the limit that $T \to \infty$, the probability that the upper and lower bounds agree tends to 1, so almost surely we eventually succeed in determining the state at time 0.

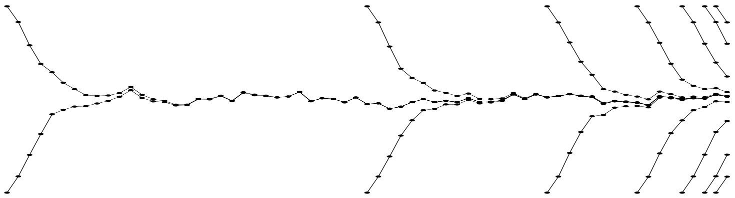


FIGURE 20.2. Illustration of CFTP in the monotone setting. Shown are the heights of the upper and lower trajectories started at various starting times in the past. When a given epoch is revisited later by the algorithm, it uses the same randomizing operation.

{fig:mcftp}

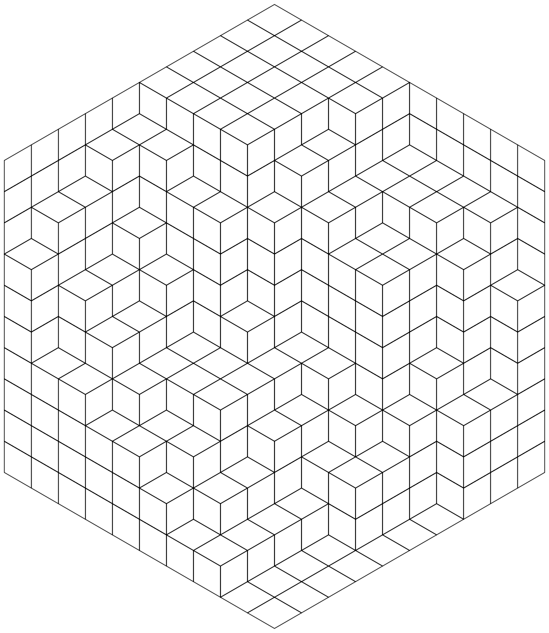
The randomizing operation (the heat-bath in the above Ising model example) defines a coupling of the Markov chain, also sometimes called a *stochastic flow* since it couples not just two states but all the states in the state space. For CFTP, the choice of the coupling is as important as the choice of the Markov chain. To illustrate this we consider another example, tilings of a hexagon by lozenges, which are $60°/120°$ rhombuses (see Figure 20.3). The set of lozenge tilings comes equipped with a natural partial order $\preceq$: we say that one tiling lies below another tiling if, when we view the tilings as collections of little boxes contained within a large box, the first collection of boxes is a subset of the other collection of boxes. The minimum configuration $\hat{0}$ is just the empty collection of little boxes, and the maximum configuration $\hat{1}$ is the full collection of little boxes.

A site in the tiling is just a vertex of one of the rhombuses that is contained within the interior of the hexagon. For each possible tiling, these sites form a triangular lattice. If a site is surrounded by three lozenges, then the three lozenges will have three different orientations; there are two different ways for a site to be surrounded by three lozenges — the horizontal will lie either above the site or below it. One possible randomizing operation would with probability $1/2$ do nothing, and with probability $1/2$ pick a uniformly random site in the tiling, and if that site is surrounded by three lozenges, rearrange those three lozenges. Another possible randomizing operation would pick a site uniformly at random, and then (viewing the tiling as a collection of boxes) with probability $1/2$ tries to add a little box at the site and with probability $1/2$ tries to remove a little box at the site. (These attempts to add or remove a little box only succeed when the resulting configuration of little boxes would be stable under gravity, otherwise the randomizing operation leaves the configuration alone.) It is straightforward to check that both of these randomizing operations give rise to the same Markov chain, i.e., a given tiling can be updated according to the first randomizing operation or the second randomizing operation, and either way, the distribution of the resulting tiling will be precisely the same. However, for purposes of CFTP the second randomizing operation is much better, because it respects the partial order $\preceq$, whereas the first randomizing operation does not.

With the Ising model and tiling examples in mind, we give pseudocode for "monotone CFTP," which is CFTP when applied to state spaces with a partial $\preceq$ (with a top state $\hat{1}$ and bottom state $\hat{0}$) that is preserved by the randomizing operation:

```
$T \leftarrow 1$
repeat
    $upper \leftarrow \hat{1}$
    $lower \leftarrow \hat{0}$
    for $t = -T$ to $-1$
        $upper \leftarrow \phi(upper, U_t)$
        $lower \leftarrow \phi(lower, U_t)$
    $T \leftarrow 2T$
until $upper = lower$
return $upper$
```

Here the variables U_t represent the intrinsic randomness used in the randomizing operations. In the Ising model heat-bath example above, U_t consists of a random number representing a site together with a random real number between 0 and 1.



In the tiling example, U_t consists of the random site together with the outcome of a coin toss. The procedure ϕ deterministically updates a state according to the random variable U_t.

Recall that we are imagining that the randomizing operation has been going on for all time, and that someone has recorded the random variables U_t that drive the randomizing operations, and our goal is to determine the state at time 0. Clearly if we read the random variable U_t more than one time, it would have the same value both times. Therefore, when the random mapping $\phi(\cdot, U_t)$ is used in one iteration of the repeat loop, for any particular value of t, it is essential that the same mapping be used in all subsequent iterations of the loop. We may accomplish this by storing the U_t's; alternatively, if (as is typically the case) our U_t's are given by some pseudo-random number generator, we may simply suitably reset the random number generator to some specified seed $seed(i)$ each time t equals -2^i.

Remark: Many people ask about different variations of the above procedure, such as what happens if we couple into the future, or what happens if we use fresh randomness each time we need to refer to the random variable U_t. There is a simple example that rules out the correctness of all such variations that have been suggested. Consider the state space $\{1, 2, 3\}$, where the randomizing operation with probability $1/2$ increments the current state by 1 (unless the state is 3), and with probability $1/2$ decrements the current state by 1 (unless the state is 1). We leave it as an exercise to verify that this example rules out the correctness of the above two variants. There are in fact other ways to obtain samples from the stationary distribution of a monotone Markov chain, such as by using Fill's algorithm (Fill, 1998) or "read-once CFTP" (Wilson, 2000), but these are not the sort of procedures that one will discover by randomly mutating the above procedure.

It is worth noting that monotone-CFTP is efficient whenever the underlying Markov chain is rapidly mixing. If H denotes the length of the longest totally ordered chain of states between $\hat{0}$ and $\hat{1}$, then in Propp and Wilson (1996) we proved that the number of randomizing operation updates that monotone-CFTP performs before returning a sample as at least T_{mix} and at most $O(T_{\mathrm{mix}} \log H)$, where T_{mix} is the mixing time of the Markov chain when measured with the total variation distance.

There are a surprisingly large number of Markov chains for which monotone-CFTP may be used (see Propp and Wilson (1996) and other articles listed in `http://www.dbwilson.com/exact/`). In the remainder of this chapter we describe a variety of scenarios in which CFTP has been used even when monotone-CFTP cannot be used.

20.3. Perfect Sampling via Coupling from the past

Computationally, one needs three things in order to be able to implement the CFTP strategy: a way of generating (and representing) certain maps from the state space $\mathbf{X}$ to itself; a way of composing these maps; and a way of ascertaining whether total coalescence has occurred, or equivalently, a way of ascertaining whether a certain composite map (obtained by composing many random maps) collapses all of $\mathbf{X}$ to a single element.

The first component is what we call the random map procedure; we model it as an oracle that on successive calls returns independent, identically distributed functions f from $\mathbf{X}$ to $\mathbf{X}$, governed by some selected probability distribution $\mathbf{P}$ (typically supported on a very small subset of the set of all maps from $\mathbf{X}$ to itself). We use the oracle to choose independent, identically distributed maps $f_{-1}, f_{-2}, f_{-3}, \ldots, f_{-N}$, where how far into the past we have to go (N steps) is determined during run-time itself. (In the notation of the previous section, $f_t(x) = \phi(x, U_t)$.) The defining property that N must have is that the composite map

$$F^0_{-N} \overset{\text{def}}{=} f_{-1} \circ f_{-2} \circ f_{-3} \circ \cdots \circ f_{-N}$$

must be collapsing. Finding such an N thus requires that we have both a way of composing f's and a way of testing when such a composition is collapsing. (Having the test enables one to find such an N, since one can iteratively test ever-larger values of N, say by successive doubling, until one finds an N that works. Such an N will be a random variable that is measurable with respect to $f_{-N}, f_{-N+1}, \ldots, f_{-1}$.)

Once a suitable N has been found, the algorithm outputs $F^0_{-N}(x)$ for any $x \in \mathbf{X}$ (the result will not depend on x, since F^0_{-N} is collapsing). We call this output the CFTP sample. It must be stressed that when one is attempting to determine a usable N by guessing successively larger values and testing them in turn, one must use the *same* respective maps f_i during each test. That is, if we have just tried starting the chain from time $-N_1$ and failed to achieve coalescence, then, as we proceed to try starting the chain from time $-N_2 < -N_1$, we must use the same maps $f_{-N_1}, f_{-N_1+1}, \ldots, f_{-1}$ as in the preceding attempt. This procedure is summarized below:

```
$T \leftarrow 1$
while $f_{-1} \circ \cdots \circ f_{-T}$ is not collapsing
    Increase $T$
return the value to which $f_{-1} \circ \cdots \circ f_{-T}$ collapses $\mathbf{X}$
```

As long as the nature of $\mathbf{P}$ guarantees (almost sure) eventual coalescence, and as long as $\mathbf{P}$ bears a suitable relationship to the distribution π, the CFTP sample will be distributed according to π. Specifically, it is required that $\mathbf{P}$ preserve π in the sense that if a random state x is chosen in accordance with π and a random map f is chosen in accordance with $\mathbf{P}$, then the state $f(x)$ will be distributed in accordance with π. In the next several sections we give examples.

20.4. The hard-core model

The states of this model are given by subsets of the vertex-set of a finite graph G, or equivalently, by $0, 1$-valued functions on the vertex-set. We think of 1 and 0 as respectively denoting the presence or absence of a particle. In a legal state, no two adjacent vertices may both be occupied by particles. The probability of a particular legal state is proportional to λ^m, where m is the number of particles (which depends on the choice of state) and λ is some fixed parameter-value. We denote this probability distribution by π. That is, $\pi(S) = \lambda^{|S|}/Z$ where S is a state, $|S|$ is the number of particles in that state, and $Z = \sum_S \lambda^{|S|}$.

Luby and Vigoda (1995) provide a simple Markov chain Monte Carlo procedure for randomizing an initial hard-core state. The random moves they consider are determined by a pair of adjacent vertices u, v and a pair of numbers i, j with (i, j) equal to $(0, 0)$, $(0, 1)$, or $(1, 0)$. They assume that the pair u, v is chosen uniformly from the set of pairs of adjacent vertices in G, and that (i, j) is $(0, 0)$ with probability $\frac{1}{1+2\lambda}$, $(0, 1)$ with probability $\frac{\lambda}{1+2\lambda}$, and $(1, 0)$ with probability $\frac{\lambda}{1+2\lambda}$. Once such a quadruple u, v, i, j is chosen, the algorithm proposes to put a vacancy (respectively particle) at vertex u if i is 0 (respectively 1), and similarly for v and j; if the proposed move would lead to an illegal state, it is rejected, otherwise it is accepted. It is not hard to show that this randomization procedure has π as its unique steady-state distribution.

Luby and Vigoda show that as long as $\lambda \leq \frac{1}{\Delta-3}$, where $\Delta \geq 4$ is the maximum degree of G, this Markov chain is rapidly mixing. They do this by using a coupling argument: two initially distinct states, evolved in tandem, tend to coalesce over time. That is, the authors implicitly embed the Markov chain in a stochastic flow. As such, the method cries out to be turned into a perfect sampling scheme via CFTP.

This is easy to do. Following Häggström and Nelander (1998) and Huber (1998), one can associate with each *set* of hard-core states a three-valued function on the vertex-set, where the value “1” means that all states in the set are known to have a particle at that vertex, the value “0” means that all states in the set are known to have a vacancy at that vertex, and the value “?” means that it is possible that some of the states in the set have a particle there while others have a vacancy. We can operate directly on this three-valued state-model by means of simple rules that mimic the Luby-Vigoda algorithm on the original two-valued model.

More specifically, we start with a three-valued configuration in which the adjacencies 0–0, 0–?, and ?–? are permitted but in which a 1 can only be adjacent to 0’s. Proposals are still of the form $(0, 0)$, $(0, 1)$, $(1, 0)$, and they still have respective probabilities $\frac{1}{2\lambda+1}$, $\frac{\lambda}{2\lambda+1}$, and $\frac{\lambda}{2\lambda+1}$, but proposals are implemented differently. When it is proposed to put 0’s at u and v, the proposal is always accepted. When it is proposed to put 0 at u and 1 at v, there are three cases. If all the vertices adjacent to v (other than u) have a 0, the proposal is accepted. If any vertex adjacent to v (other than u) has a 1, the proposal is simply rejected and nothing happens. However, if vertex v has a neighbor (other than u) that has a ? but no neighbor (other than u) that has a 1, then v gets marked with ? and u also gets marked with ? (unless u was already 0, in which case the marking of u does not change). When it is proposed to put 1 at u and 0 at v, the same procedure is followed, but with the roles of u and v reversed.

In short, we can take the work of Luby and Vigoda and, without adding any new ideas, check that their way of coupling two copies of the Luby-Vigoda Markov chain extends to a stochastic flow on the whole state-space. Moreover, this flow can be simulated in such a way that coalescence is easily detected: it is not hard to show that if the 0,1,? Markov chain, starting from the all-?’s state, ever reaches a state in which there are no ?’s, then the Luby-Vigoda chain, using the same random proposals, maps all initial states into the same final state. Hence we might want to call the 0,1,? Markov chain the “certification chain”, for it tells us when the stochastic flow of primary interest has achieved coalescence.

One might fear that it would take exponentially long for the certification chain to certify coalescence, but the proof that Luby and Vigoda give carries over straightforwardly to the three-valued setting, and shows that the number of ?’s tends to shrink to zero in polynomial time (relative to the size of the system).

We mention that Häggström and Nelander (1998) and Huber (1998) originally used the more natural single-site heat-bath randomizing operation, in which only one vertex at a time is modified. Work of Randall and Tetali (2000), in conjunction with the Luby-Vigoda result, implies that the single-site heat-bath Markov chain is also rapidly mixing for $\lambda \leq \frac{1}{\Delta-3}$.

20.5. Random state of an unknown Markov chain

Now we come to a problem that in a sense encompasses all the cases we have discussed so far: the problem of sampling from the steady-state distribution $\pi(\cdot)$ of a general Markov chain. Of course, in the absence of further strictures this problem admits a trivial “solution”: just solve for the steady-state distribution analytically! In the case of the systems studied in sections 3 through 5, this is not practical, since the state spaces are large. We now consider what happens if the state space is small but the analytic method of simulation is barred by imposing the constraint that the transition probabilities of the Markov chain are unknown: one merely has access to a black box that simulates the transitions.

It might seem that, under this stipulation, no solution to the problem is possible, but in fact a solution was found by Asmussen, Glynn, and Thorisson (1992). However, their algorithm was not very efficient. Subsequently Aldous (1995) and Lovász and Winkler (1995) found faster procedures (although the algorithm of Aldous involves controlled but non-zero error). The CFTP-based solution given below is even faster than that of Lovász and Winkler.

For pictorial concreteness, we envision the Markov chain as biased random walk on some directed graph G whose arcs are labeled with weights, where the transition probabilities from a given vertex are proportional to the weights of the associated arcs (as in the preceding section). We denote the vertex set of G by $\mathbf{X}$, and denote the steady-state distribution on $\mathbf{X}$ by π. Propp and Wilson (1998) give a CFTP-based algorithm that lets one sample from this distribution π.

Our goal is to define suitable random maps from $\mathbf{X}$ to $\mathbf{X}$ in which many states are mapped into a single state. We might therefore define a random map from $\mathbf{X}$ to itself by starting at some fixed vertex r, walking randomly for some large number N of steps, and mapping all states in $\mathbf{X}$ to the particular state v that one has landed in after N steps. However, v is subject to initialization bias, so this random map procedure typically does not preserve π in the sense defined in section 2.

What actually works is a multi-phase scheme of the following sort: Start at some vertex r and take a random walk for a *random* amount of time T_1, ending at some state v; then map every state that has been visited during that walk to v. In the second phase, continue walking from v for a further random amount of time T_2, ending at some new state v'; then map every state that was visited during the second phase but not the first to v'. In the third phase, walk from v' for a random time to a new state v'', and map every hitherto-unvisited state that was visited during that phase to the state v''. And so on. Eventually, every state gets visited, and every state gets mapped to some state. Such maps, like tree-maps, are easy to compose, and it is easy to recognize when such a composition is coalescent (it maps every state to one particular state).

There are two constraints that our random durations $T_1, T_2, \ldots$ must satisfy if we are planning to use this scheme for CFTP. (For convenience we will assume henceforth that the T_i's are i.i.d.) First, the distribution of each T_i should have the property that, at any point during the walk, the (conditional) expected time until the walk terminates does not depend on where one is or how one got there. This ensures that the stochastic flow determined by these random maps preserves π. Second, the time for the walk should be neither so short that only a few states get visited by the time the walk ends nor so long that generating even a single random map takes more time than an experimenter is willing to wait. Ideally, the expected duration of the walk should be on the order of the cover-time for the random walk. Propp and Wilson (1998) show that by using the random walk itself to estimate its own cover-time, one gets an algorithm that generates a random state distributed according to π in expected time at most 15 times the cover time.

At the beginning of this section, we said that one has access to a black box that simulates the transitions. This is, strictly speaking, ambiguous: Does the black box have an "input port" so that we can ask it for a random transition from a specified state? Or are we merely passively observing a Markov chain in which we have no power to intervene? This ambiguity gives rise to two different versions of the problem, of separate interest. Our CFTP algorithm works for both of them.

For the "passive" version of the problem, it is not hard to show that no scheme can work in expected time less than the expected cover time of the walk, so in this setting our algorithm runs in time that is within a constant factor of optimal. It is possible to do better in the active setting, but no good lower bounds are currently known for this case.

APPENDIX A

Notes on notation

The $\subset$ symbol includes the possibility of equality: hence, $\Omega \subset \Omega$ is true. (Equation 5.1)

$a \wedge b = \min(a, b)$. (Proposition 5.5)

$\mathbb{Z}_n = \{0, \ldots, n-1\}$ = set of remainders mod n. (definition of random walk on n-cycle, chapter 3.)

$a_n = O(b_n)$ mean that there is a constant c so that $a_n/b_n \leq c$ for all n.

$a_n = o(b_n)$ means that $\lim_{n\to\infty} a_n/b_n = 0$.

$a_n \asymp b_n$ means that $a_n = O(b_n)$ and $b_n = O(a_n)$. In other words, there are constants $0 < c_1, c_2 < \infty$ so that $c_1 \leq a_n/b_n \leq c_2$ for all n.

For a real-valued function $f : \Omega \to \mathbb{R}$ and a probability distribution μ on Ω, we write $E_\mu(f)$ for $\sum_{x\in\Omega} f(x)\mu(x)$.

The symbol := means *defined as*. For example, $f(x) := x^2$ means that f is the function defined at x to be x^2.

APPENDIX B

Background Material

B.1. Probability Spaces and Random Variables

{Apx:RVs}

For a comprehensive account of *measure theory*, the mathematical theory underlying modern probability, the interested reader should consult one of the many textbooks on the subject, for example Billingsley (1995). We will need very little of this theory in this book, but for the purpose of establishing notation and terminology we record a few definitions here.

A *probability space* is a set Ξ, together with a family of subsets of Ξ whose elements are called *events*. When Ξ is a finite or countable set, all subsets are events, but when Ξ is uncountable, for example a subinterval of $\mathbb{R}$, not *every* subset is an event. Events satisfy the following *closure properties*:

(i) Ξ is an event,
(ii) if $B_1, B_2, \ldots$ are all events, then the union $\bigcup_{i=1}^{\infty} B_i$ is also an event, and
(iii) if B is an event, so is $\Xi \setminus B$.

The following are two very important examples of probability spaces.

Example B.1. When Ξ is a subinterval of $\mathbb{R}$, the set of events is the smallest collection containing all open subiniterval of Ξ and satisfying the closure properties. In this case, events are called *Borel* sets.

Example B.2. When Ξ is the sequence space S^{∞} for a finite set S, a set of the form

$$A_1 \times A_2 \times \cdots \times A_n \times S \times S \cdots , \quad A_k \subset S \text{ for all } k = 1, \ldots, n$$

is called a *cylinder* set. The events in S^{∞} is the smallest collection of sets satisfying the closure properties and containing the cylinder sets.

Given a probability space, a *probability measure* is a non-negative function $\mathbf{P}$ defined on events and satisfying the *probability axioms*:

(i) $\mathbf{P}(\Xi) = 1$,
(ii) for any sequence of events $B_1, B_2, \ldots$ which are mutually disjoint, meaning $B_i \cap B_j = \varnothing$ for $i \neq j$,

$$\mathbf{P}\left(\bigcup_{i=1}^{\infty} B_i\right) = \sum_{i=1}^{\infty} \mathbf{P}(B_i).$$

If Ξ is a countable set, a *probability distribution* on Ξ is a function $p : \Xi \to [0,1]$ so that $\sum_{\xi\in\Xi} p(\xi) = 1$. A probability distribution induces a probability measure on the events of Ξ via the definition

$$\mathbf{P}(B) = \sum_{\xi\in B} p(\xi). \tag{B.1}$$

{Eq:distdefn}

If Ξ is a subinterval of $\mathbb{R}$, and $f : \Xi \to [0,\infty)$ satisfies $\int_\Xi f(\xi)d\xi = 1$, then f is called a *density function*. Given a density function, a probability measure can be defined for events B by

$$\mathbf{P}(B) = \int_B f(\xi)d\xi.$$

Given a probability space, a *random variable* X is a function defined on Ξ. The notation $\{X \in A\}$ means $\{\xi \in \Xi \,:\, X(\xi) \in A\} = X^{-1}(A)$. *Any set of the form* $\{X \in A\}$ *encountered in this book will be an event.* The *distribution* of a random variable X is the probability measure μ_X on $\mathbb{R}$ defined for events B by

$$\mu_X(B) = \mathbf{P}\{X \in B\}.$$

Suppose that X is a real-valued random variable. X is called *discrete* if there is a finite or countable set S so that $\mu_X(S) = 1$. In this case, the function

$$p_X(a) = \mathbf{P}\{X = a\}$$

is a probability distribution on S.

A real-valued random variable X is called *continuous* if there is a density function f on $\mathbb{R}$ so that

$$\mu_X(A) = \int_A f(x)dx.$$

For a discrete real-valued random variable X, the *expectation* $\mathbf{E}(X)$ can be computed by the formula

$$\mathbf{E}(X) = \sum_{x\in\mathbb{R}} x\mathbf{P}\{X = a\}.$$

(Note that there are at most countably non-zero summands.) For a continuous real-valued random variable X,

$$\mathbf{E}(X) = \int_{\mathbb{R}} x f_X(x)dx.$$

A sequence of random variables (X_t) *converge in probability* to a random variable X if

$$\lim_{t\to\infty} \mathbf{P}\{|X_t - X| > \varepsilon\} = 0, \tag{B.2}$$

for all ε. This is denoted by $X_t \xrightarrow{\text{pr}} X$.

Theorem B.3 (Weak Law of Large Numbers). *If (X_t) is a sequence of independent random variable so that* $\mathbf{E}(X_t) = \mu$ *and* $\operatorname{Var}(X_t) = \sigma^2$ *for all t, then*

$$\frac{1}{T}\sum_{t=1}^{T} X_t \xrightarrow{\text{pr}} \mu \quad \textit{as } T \to \infty.$$

Proof. By linearity of expectation, $\mathbf{E}(T^{-1}\sum_{t=1}^{T} X_t) = \mu$, and by independence, $\mathrm{Var}(T^{-1}\sum_{t=1}^{T} X_t) = \sigma^2/T$. Applying Chebyshev's inequality,

$$\mathbf{P}\left\{\left|\frac{1}{T}\sum_{t=1}^{T} X_t - \mu\right| > \varepsilon\right\} \leq \frac{\sigma^2}{T\varepsilon^2}.$$

For every $\varepsilon > 0$ fixed, the right-hand side tends to zero as $T \to \infty$. ■

{Thm:SLLN}

Theorem B.4 (Strong Law of Large Numbers). *Let $Z_1, Z_2, \ldots$ be a sequence of random variables with $E(Z_s) = 0$ for all s and*

$$\mathrm{Var}(Z_{s+1} + \cdots + Z_{s+k}) \leq Ck$$

for all s and k. Then

$$\mathbf{P}\left\{\lim_{t\to\infty} \frac{1}{t}\sum_{s=0}^{t-1} Z_s = 0\right\} = 1. \tag{B.3}$$ {Eq:SLLN}

Proof. Let $A_t := t^{-1}\sum_{s=0}^{t-1} Z_s$. Then

$$\mathbf{E}(A_t^2) = \frac{\mathbf{E}\left[\left(\sum_{s=0}^{t-1} Z_s\right)^2\right]}{t^2} \leq \frac{C}{t}.$$

Thus, $\mathbf{E}\left(\sum_{m=1}^{\infty} A_{m^2}^2\right) < \infty$, which in particular implies that

$$\mathbf{P}\left\{\sum_{m=1}^{\infty} A_{m^2}^2 < \infty\right\} = 1, \quad \text{and} \quad \mathbf{P}\left\{\lim_{m\to\infty} A_{m^2} = 0\right\} = 1. \tag{B.4}$$ {Eq:ASquare}

For a given t, let m_t be such that $m_t^2 \leq t < (m_t + 1)^2$. Then

$$A_t = \frac{1}{t}\left(m_t^2 A_{m_t^2} + \sum_{s=m_t^2}^{t-1} Z_s\right). \tag{B.5}$$ {Eq:ASum}

Since $\lim_{t\to\infty} t^{-1}m_t^2 = 1$, by (B.4),

$$\mathbf{P}\left\{\lim_{t\to\infty} t^{-1}m_t^2 A_{m_t^2} = 0\right\} = 1. \tag{B.6}$$ {Eq:ASum1}

Defining $B_t := t^{-1}\sum_{s=m_t^2}^{t-1} Z_s$,

$$\mathbf{E}(B_t^2) = \frac{\mathrm{Var}\left(\sum_{s=m_t^2}^{t-1} Z_s\right)}{t^2} \leq \frac{2Cm_t}{t^2} \leq \frac{2C}{t^{3/2}}.$$

Thus $\mathbf{E}(\sum_{t=0}^{\infty} B_t^2) < \infty$, and

$$\mathbf{P}\left\{\lim_{t\to\infty} \frac{\sum_{s=m_t^2+1}^{t} Z_s}{t} = 0\right\} = 1. \tag{B.7}$$ {Eq:ASum2}

Putting together (B.6) and (B.7), from (B.5) we conclude that (B.3) holds. ■

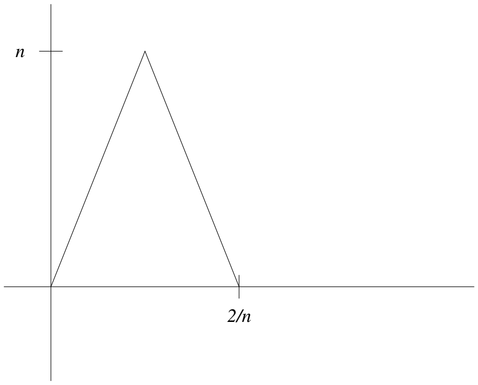


FIGURE B.1. A sequence of functions whose integrals do no converge to the integral of the limit. fig:nonunif

B.1.1. Limits of Expectations. We know from calculus that if (f_n) is a sequence of functions defined on an interval I, satisfying for every $x \in I$,

$$\lim_{n\to\infty} f_n(x) = f(x)$$

then it is not necessarily the case that

$$\lim_{n\to\infty} \int_I f_n(x)dx = \int_I f(x)dx\,.$$

As an example, consider the function whose graph is shown in Figure B.1. The integral of this function is always 1, but each $x \in [0,1]$, the limit $\lim_n g(x) = 0$. That is,

{eq:noconverge}

$$\int_0^1 \lim_n g_n(x)dx = 0 \neq 1 = \lim_n \int_0^1 g_n(x)dx. \tag{B.8}$$

We can turn this into a story about random variables. Let U be a uniform random variable, and let $Y_n = g_n(U)$. Notice that $Y_n \to 0$. Then

$$\mathbf{E}(Y_n) = \mathbf{E}(g_n(U)) = \int g_n(x) f_U(x)dx = \int_0^1 g_n(x)dx,$$

as the density of U is $f_U(x) = \mathbf{1}_{[0,1]}$. Then by (B.8) we see that

$$\lim_{n\to\infty} \mathbf{E}(Y_n) \neq \mathbf{E}\left(\lim_{n\to\infty} Y_n\right).$$

Now that we have seen that we cannot always move a limit inside an expectation, can we ever? The answer is "yes", given some additional assumptions.

{prop:dconv}

PROPOSITION B.5. *Let Y_n be a sequence of random variables and Y a random variable so that* $\mathbf{P}\{\lim_{n\to\infty} Y_n = Y\} = 1$.

{it:dom}

(i) *If there is a constant K independent of n so that $|Y_n| < K$ for all n, then*

$$\lim_{n\to\infty} \mathbf{E}(Y_n) = \mathbf{E}(Y).$$

{it:mon}

(ii) *If* $\mathbf{P}\{Y_n \leq Y_{n+1}\} = 1$ *for all n, then*

$$\lim_{n\to\infty} \mathbf{E}(Y_n) = \mathbf{E}(Y).$$

Proposition B.5(i) is called the *Dominated Convergence Theorem*, and Proposition B.5(ii) is called the *Monotone Convergence Theorem*.

Proof. For any $\varepsilon > 0$,

$$|Y_n - Y| \leq 2K\mathbf{1}_{\{|Y_n - Y| > \varepsilon/2\}} + \varepsilon/2,$$

and taking expectation above shows that

$$\begin{aligned} |\mathbf{E}(Y_n) - \mathbf{E}(Y)| &\leq \mathbf{E}\left(|Y_n - Y|\right) \\ &\leq 2K\mathbf{P}\left\{|Y_n - Y| > \varepsilon/2\right\} + \varepsilon/2. \end{aligned}$$

Since $\mathbf{P}\left\{|Y_n - Y| \geq \varepsilon/2\right\} \to 0$, by taking n sufficiently large,

$$|\mathbf{E}(Y_n) - \mathbf{E}(Y)| \leq \varepsilon.$$

That is, $\lim_{n\to\infty} \mathbf{E}(Y_n) = \mathbf{E}(Y)$. ■

For a proof of (ii), see Billingsley (1995, Theorem 16.2)

B.2. Metric Spaces

{App:MS}

A set M equipped with a function ρ measuring the distance between its elements is called a *metric space*. In Euclidean space $\mathbb{R}^k$, the distance between vectors is measured by the norm $\|x - y\| = \sqrt{\sum_{i=1}^n (x_i - y_i)^2}$. On a graph, distance can be measured as the length of the shortest path connecting x and y. These are examples of metric spaces.

The function ρ must satisfy some properties to reasonably be called a distance. In particular, it should be symmetric, in the sense that there should be no difference between measuring from a to b and measuring from b to a. Distance should never be negative, and there should be no two distinct elements which have distance zero. Finally, the distance $\rho(a, c)$ from a to c should never be greater than proceeding via a third point b and adding the distances $\rho(a, b) + \rho(b, c)$. For obvious reasons, this last property is called the *triangle inequality*.

We summarize here these properties:

(i) $\rho(a, b) = \rho(b, a)$ for all $a, b \in M$,
(ii) $\rho(a, b) \geq 0$ for all $a, b \in M$, and $\rho(a, b) = 0$ only if $a = b$,
(iii) For any three elements $a, b, c \in M$,

$$\rho(a, c) \leq \rho(a, b) + \rho(b, c). \tag{B.9}$$

B.3. Linear Algebra

{App:LA} {Thm:SpectralThm}

Theorem B.6 (Spectral Theorem for Symmetric Matrices). *If M is a symmetric $m \times m$ matrix, then there exists a matrix U with $U'U = I$ and a diagonal matrix Λ so that $M = U'\Lambda U$.*

(The matrix U' is the *transpose* of U, defined as $U'_{i,j} := U_{j,i}$.) A proof of Theorem B.6 can be found, for example, in Horn and Johnson (1990, Theorem 4.1.5).

Another way of formulating the Spectral Theorem is to say that there is an orthonormal basis of eigenvectors for M. The columns of U form one such basis, and the eigenvalue associated to the ith column is $\lambda_i = \Lambda_{ii}$.

The variational characterization of the eigenvalues of a symmetric matrix is very useful:

{Thm:RayleighRitz}

Theorem B.7 (Rayleigh-Ritz). *Let M be a symmetric matrix with eigenvalues*

$$\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_n$$

and associated eigenvectors $x_1, \ldots, x_n$. Then

$$\lambda_k = \max_{\substack{x \neq 0 \\ x \perp x_1, \ldots, x_{k-1}}} \frac{\langle x, Ax \rangle}{\langle x, x \rangle}.$$

See Horn and Johnson (1990, p. 178) for a discussion.

B.4. Miscellaneous

Stirling's formula says that

{Eq:SF}

$$n! \sim \sqrt{2\pi} e^{-n} n^{n+1/2}, \tag{B.10}$$

where $a_n \sim b_n$ means that $\lim_{n\to\infty} a_n b_n^{-1} = 1$.

More precise results are known, for example,

{Eq:SFBounds}

$$n! = \sqrt{2\pi} e^{-n} n^{n+1/2} e^{\varepsilon_n}, \qquad \frac{1}{12n+1} \leq \varepsilon_n \leq \frac{1}{12n}. \tag{B.11}$$

APPENDIX C

Solutions to Selected Exercises

Chapter 2

Solution to 2.4. Assume that n is even and let $q = 1 - p$. For $y \in \{0,1\}^m$, the probability that exactly the first m pairs are discordant and yield the word y as output is

$$[pq]^m[1-2pq]^{n/2-m}.$$

Given that $L = m$, there are $\binom{n/2}{m}$ possibilities for the locations of the m disagreeing pairs. By symmetry, we have

$$\mathbf{P}\{(Y_1,\ldots,Y_m) = y,\ L = m\} = \binom{n/2}{m}[pq]^m[1-2pq]^{n/2-m}. \tag{C.1}$$

{Eq:JointYEll}

The marginal distribution of L is

$$\mathbf{P}\{L = m\} = \binom{n/2}{m}[2pq]^m[1-2pq]^{n/2-m}. \tag{C.2}$$

{Eq:MarginalEll}

Together (C.1) and (C.2) show that

$$\mathbf{P}\{Y = y \mid L = m\} = 2^{-m}.$$

■

Solution to 2.10. We proceed by induction on n. The base case $n = 1$ is clear. Assume that the $(n-1)$-step algorithm indeed produces a uniformly distributed $\xi_{n-1} \in \Xi^{\mathrm{nr}}_{n-1}$. Extend ξ_{n-1} to ξ_n according to the algorithm, picking one of the three available extensions at random. Note that $|\Xi^{\mathrm{nr}}_n| = 4 \cdot 3^{n-1}$. For h any path in Ξ^{nr}_n, let h_{n-1} be the projection of h to Ξ^{nr}_{n-1}, and observe that

$$\mathbf{P}\{\xi_n = h\} = \mathbf{P}\{\xi_n = h \mid \xi_{n-1} = h_{n-1}\}\mathbf{P}\{\xi_{n-1} = h_{n-1}\} = \frac{1}{3}\left(\frac{1}{4\cdot 3^{n-2}}\right) = \frac{1}{4\cdot 3^{n-1}}.$$

■

Solution to 2.11. Since the number of self-avoiding walks of length n is clearly bounded by $c_{n,4}$, and our method for generating non-reversing paths is uniform over Ξ^{nr}_n which has size $4 \cdot 3^{n-1}$, the second part follows from the first.

There are $4(3^3) - 8$ walks of length 4 starting at the origin which are non-reversing and do not return to the origin. At each 4-step stage later in the walk, there are 3^4 non-reversing paths of length 4, of which six create loops. This establishes (2.13). ■

Solution to 2.12. This is established by induction. The cases $n = 0$ and $n = 1$ are clear. Suppose it holds for $n \leq k - 1$. The number of configurations $\omega \in \Omega_k$ with $\omega(k) = 0$ is the same as the total number of configurations in Ω_{k-1}. Also, the number of configurations $\omega \in \Omega_k$ with $\omega(k) = 1$ is the same as the number of configurations in Ω_{k-1} having no particle at $k - 1$, which is the same as the number of configurations in Ω_{k-2}. ■

Solution to 2.13. Let ω be an element of Ω_n, and let X be the random element of Ω_n generated by the algorithm. If $\omega(n) = 1$, then

$$\mathbf{P}\{X = \omega\} = \frac{1}{f_{n-2}}\left(\frac{f_{n-2}}{f_n}\right) = \frac{1}{f_n}.$$

Similarly, if $\omega(n) = 0$, then $\mathbf{P}\{X = \omega\} = 1/f_n$. ■

Solution to 2.1. σ is a permutation if all of the images are distinct, which occurs with probability

$$p_n := \frac{n!}{n^n}.$$

where $a_n \sim b_n$ means that $\lim_{n\to\infty} a_n/b_n = 1$. Using Stirling's Formula shows that

$$p_n \sim \sqrt{2\pi n} e^{-n}.$$

Since the number of trials needed is geometric with parameter p_n, the expected number of trials needed is asymptotic to

$$\frac{e^n}{\sqrt{2\pi n}}.$$

■

Solution to 2.2. The proposed method clearly yields a uniform permutation when $n = 1$ or $n = 2$. However, it fails to do for for all larger values of n. One way to see this is to note that at each stage in the algorithm, there are n options. Hence the probability of each possible permutation must be an integral multiple of $1/n^n$. For $n \geq 3$, $n!$ is not a factor of n^n, so no permutation can have probability $1/n!$ of occurring. ■

Solution to 2.3. We proceed by induction. Let H_j be the function defined in the first j steps described above; the domain of H_j is $[j]$. Clearly H_1 is uniform on $\Omega_{k,1}$. Suppose H_{j-1} is uniform on $\Omega_{k,j-1}$. Let $h \in \Omega_{k,j}$. Write h_{j-1} for the restriction of h to the domain $[j-1]$. Then

$$\mathbf{P}\{H_{j-1} = h_{j-1}\} = |\Omega_{k,j-1}|^{-1},$$

by the induction hypothesis. Note that

$$|\Omega_{k,j}| = (k-1)|\Omega_{k,j-1}|,$$

since for each element of $\Omega_{k,j-1}$ there are $k - 1$ ways to extend it to an element of $\Omega_{k,j}$, and every element of $\Omega_{k,j}$ can be obtained as such an extension. By the construction and the induction hypothesis,

$$\begin{aligned}\mathbf{P}\{H_j = h\} &= \mathbf{P}\{H_{j-1} = h_{j-1}\}\mathbf{P}\{H_j = h \mid H_{j-1} = h_{j-1})\\ &= \frac{1}{|\Omega_{k,j-1}|}\frac{1}{(k-1)}\\ &= |\Omega_{k,j}|^{-1}.\end{aligned}$$

■

Chapter 3

Solution to 3.1. Since the lily pad the frog is sitting on shows a head, it must be morning, and the frog must be about to jump to the other pad. ■

Solution to 3.8. We show that the tree can be recolored, vertex-by-vertex, so that it has all odd depth vertices with the color 2, and all even depth vertices with the color 1.

Without loss of generality, assume all leaves are the same distance to the root.

Start at the leaves, and one-by-one, change them to the color of their grandparent. Because their parent has a color different from their grandparent, this is always legal.

Let the *height* of a vertex be its distance from the closest leaf. Suppose all vertices of height at most k have the same color as the grandparents. Let v be a vertex at height $k + 1$. Change it to the same color as its grandparent, say color c_1. This is possible because the parent and children of v have the same color, say c_2, which is necessarily different from c_1. If this causes v to have a different color than its grandchildren, recolor them, their grandchildren, their grandchildren's grandchildren, on so on, also with c_1. This is possible because these vertices are connected to vertices only of color c_2. All the vertices at height k can then be recolored in turn, so that all vertices at height at most $k+1$ share the same color as their grandparents.

When the level just below the root is reached, recolor these vertices, making sure to recolor any "even" descendant with the same color.

At this point, all vertices at even height have a single color, and all vertices at odd height have a single color. This configuration can be recolored, again vertex-by-vertex, so that all odd heights have color 2 and all even heights have color 1. ■

Solution to 3.10.

(a) This is by now a standard application of the parity of permutations. Note that any sequence of moves in which the empty space ends up in the lower right corner must be of even length. Since every move is a single transposition, the permutation of the tiles (including the empty space as a tile) in any such position must be even. However, the desired permutation (switching two adjacent tiles in the bottom row) is odd.

(b) In fact, all even permutations of tiles can be achieved, but it is not entirely trivial to demonstrate. See Archer (1999) for an elementary proof and some historical discussion. Zhentao Lee discovered a new and elegant elementary proof during our 2006 MSRI workshop.

■

Solution to 3.12. (b) Since $\mathbf{P}_x\{\tau^+ > t\}$ is a decreasing function of t, (3.17) suffices to bound the entire sum:

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

■

Solution to 3.13. ■

Solution to 3.14.

$$\begin{aligned}
\pi(x)P^2(x,y) &= \pi(x)\sum_{x\in\Omega} P(x,z)P(z,y)\\
&= \sum_{x\in\Omega} \pi(z)P(z,x)P(z,y)\\
&= \sum_{x\in\Omega} \pi(z)P(z,y)P(z,x)\\
&= \sum_{x\in\Omega} \pi(y)P(y,z)P(z,x)\\
&= \pi(y)\sum_{x\in\Omega} P(y,z)P(z,x)\\
&= \pi(y)P^2(y,x).
\end{aligned}$$

■

Solution to 3.18.

(a) Compute:

$$\nu_n P(x) - \mu_n(x) = \frac{1}{n}\left(\mu P^n(x) - \mu(x)\right) \leq \frac{2}{n},$$

since any probability measure has weight at most 1 at x.

(b) Bolzano-Weierstrass, applied either directly in $\mathbb{R}^{|\Omega|}$ or iteratively: first take a subsequence that converges at x_1, then take a subsequence of that which converges at x_2, and so on. Either way, it's key that the weights of the measure are bounded and that the state space is finite.

(c) Part (a) gives stationarity, while the fact that the set of probability measures on Ω (viewed as a set in $\mathbb{R}^{|\Omega|}$) is closed gives that ν is a probability distribution.

■

Chapter 4

Solution to 17.11.

(a) $x \leq U_{(k)} \leq x + dx$ if and only if among $\{U_1, U_2, \dots, U_n\}$, exactly $k-1$ lie to the left of x, one is in $[x, x+dx]$, and $n-k$ variables exceed $x+dx$. This occurs with probability

$$\binom{n}{(k-1),1,(n-k)} x^{k-1}(1-x)^{n-k}dx.$$

Thus,

$$\begin{aligned}\mathbf{E}\left(U_{(k)}\right) &= \int_0^1 \frac{n!}{(k-1)!(n-k)!} x^k (1-x)^{n-k} dx \\ &= \frac{n!}{(k-1)!(n-k)!} \frac{(n-k)!k!}{(n+1)!} \\ &= \frac{k}{n+1}.\end{aligned}$$

[The integral can be evaluated by observing that the function $\frac{k!(n-k)!}{(n+1)!} x^k (1-x)^{n-k}$ is the density for a Beta random variable with parameters $k+1$ and $n-k+1$.]

(b) The distribution function for $U_{(n)}$ is

$$F_n(x) = \mathbf{P}\{U_1 \leq x, U_2 \leq x, \ldots, U_n \leq x\} = \mathbf{P}\{U_1 \leq x\}^n = x^n.$$

Differentiating, the density function for $U_{(n)}$ is

$$f_n(x) = nx^{n-1}.$$

Consequently,

$$\mathbf{E}\left(U_{(n)}\right) = \int_0^1 xnx^{n-1} dx = \frac{n}{n+1} x^{n+1}\Big|_0^1 = \frac{n}{n+1}.$$

We proceed by induction, showing that

$$\mathbf{E}\left(U_{(n-k)}\right) = \frac{n-k}{n+1}. \tag{C.3}$$ {Eq:RevOS}

We just established the case $k = 0$. Now suppose (C.3) holds for $k = j$. Given $U_{(n-j)}$, the order statistics $U_{(i)}$ for $i = 1, \ldots, n-j-1$ have the distribution of the order statistics for $n-j-1$ independent variables uniform on $[0, U_{(n-j)}]$. Thus,

$$\mathbf{E}\left(U_{(n-j-1)} \mid U_{(n-j)}\right) = U_{(n-j)} \frac{n-j-1}{n-j},$$

and so

$$\mathbf{E}\left(U_{(n-j-1)}\right) = \mathbf{E}\left(\mathbf{E}\left(U_{(n-j-1)} \mid U_{(n-j)}\right)\right) = \mathbf{E}\left(U_{(n-j)}\right) \frac{n-j-1}{n-j}.$$

Since (C.3) holds for $k = j$ by assumption,

$$\mathbf{E}\left(U_{(n-j-1)}\right) = \frac{n-j}{n+1}\frac{n-j-1}{n-j} = \frac{n-j-1}{n+1}.$$

This establishes (C.3) for $j = k$.

(c) The joint density of $(S_1, S_2, \ldots, S_{n+1})$ is $e^{-s_{n+1}} \mathbf{1}_{\{0 < s_1 < \cdots < s_{n+1}\}}$, as can be verified by induction:

$$\begin{aligned}f_{S_1,S_2,\ldots,S_{n+1}}(s_1, \ldots, s_{n+1}) &= f_{S_1,S_2,\ldots,S_n}(s_1, \ldots, s_n) f_{S_{n+1}|S_1,\ldots,S_n}(s_{n+1} \mid s_1, \ldots, s_n) \\ &= e^{-s_n} \mathbf{1}_{\{0 < s_1 < \cdots < s_n\}} e^{-(s_{n+1}-s_n)} \mathbf{1}_{\{s_n < s_{n+1}\}} \\ &= e^{-s_{n+1}} \mathbf{1}_{\{0 < s_1 < \cdots < s_{n+1}\}}\end{aligned}$$

Because the density of S_{n+1} is $s_{n+1}^{n} e^{-s_{n+1}}/(n!)\mathbf{1}_{\{s_{n+1}>0\}}$,

$$f_{S_1,\dots,S_n|S_{n+1}}(s_1,\dots,s_n \mid s_{n+1}) = \frac{n!}{s_{n+1}^{n}}\mathbf{1}_{\{0<s_1<\cdots<s_n<s_{n+1}\}}.$$

If $T_k = S_k/S_{n+1}$ for $k = 1,\dots,n$, then

$$f_{T_1,\dots,T_k|S_{n+1}}(t_1,\dots,t_n \mid s_{n+1}) = n!\mathbf{1}_{\{0<t_1<\cdots<t_n<1\}}.$$

Since the right-hand side does not depend on s_{n+1}, the vector

$$\left(\frac{S_1}{S_{n+1}}, \frac{S_2}{S_{n+1}}, \dots, \frac{S_1}{S_{n+1}}\right)$$

is uniform over the set

$$\{(x_1,\dots,x_n) \; : \; x_1 < x_2 < \cdots < x_n\}.$$

∎

Solution to 4.2. Let f_k be the expected value of the time until our gambler stops playing. Just as for regular gambler's ruin, the values f_k are related:

$$f_0 = f_n = 0 \quad \text{and} \quad f_k = \frac{p}{2}(1 + f_{k-1}) + \frac{p}{2}(1 + f_{k+1}) + (1-p)(1 + f_k).$$

It is easy to check that setting $f_k = k(n-k)/p$ solves this system of equations. (Note that the answer is just what it should be. If she only bets a fraction p of the time, then it should take a factor of $1/p$ longer to reach her final state.) ∎

Solution to 4.3. Let (X_t) be a fair random walk on the set $\{-n,\dots,n\}$, starting at the state 0 and absorbing at $\pm n$. By Proposition 4.1, the expected time for this walk to be absorbed is $(2n-n)(2n-n) = n^2$.

The walk described in the problem can be viewed as $n - |X_t|$. Hence its expected time to absorption is also n^2. ∎

Solution to 4.5.

$$\sum_{k=1}^{n} \frac{1}{k} \geq \sum_{k=1}^{n} \int_{k}^{k+1} \frac{dt}{t} = \int_{1}^{n+1} \frac{dt}{t} = \log(n+1) \geq \log n, \tag{C.4}$$

and

$$\sum_{k=1}^{n} \frac{1}{k} = 1 + \sum_{k=2}^{n} \frac{1}{k} \leq 1 + \sum_{k=2}^{n} \int_{k-1}^{k} \frac{dt}{t} = 1 + \int_{1}^{n} \frac{dt}{t} = 1 + \log n. \tag{C.5}$$

∎

Solution to 4.6.

$$\begin{aligned}\binom{d}{k+1}P(k+1,k)+\binom{d}{k-1}P(k-1,k) &= \frac{d!}{(k+1)!(d-k-1)!}\frac{k+1}{d} \\ &\quad + \frac{d!}{(k-1)!(d-k+1)!}\frac{d-k+1}{d} \\ &= \binom{d-1}{k-1}+\binom{d}{k} \\ &= \binom{d}{k}.\end{aligned}$$

The last combinatorial identity can be seen by counting the number of size k subsets from d objects which contain a distinguished element and the number which do not contain this distinguished element. ■

Chapter 5

Solution to 5.1.

$$\begin{aligned}\left\|\mu P^t-\pi\right\|_{\mathrm{TV}} &= \frac{1}{2}\sum_{y\in\Omega}|\mu P^t(y)-\pi(y)| \\ &= \frac{1}{2}\sum_{y\in\Omega}\left|\sum_{x\in\Omega}\mu(x)P^t(x,y)-\sum_{x\in\Omega}\mu(x)\pi(y)\right| \\ &\le \frac{1}{2}\sum_{y\in\Omega}\sum_{x\in\Omega}\mu(x)|P^t(x,y)-\pi(y)| \\ &= \sum_{x\in\Omega}\mu(x)\frac{1}{2}\sum_{y\in\Omega}|P^t(x,y)-\pi(y)| \\ &= \sum_{x\in\Omega}\mu(x)\left\|P^t(x,\cdot)-\pi\right\|_{\mathrm{TV}} \\ &\le \max_{x\in\Omega}\left\|P^t(x,\cdot)-\pi\right\|_{\mathrm{TV}}.\end{aligned}$$

Since this holds for any μ, we have

$$\sup_{\mu}\left\|\mu P^t-\pi\right\|_{\mathrm{TV}} \le \max_{x\in\Omega}\left\|P^t(x,\cdot)-\pi\right\|_{\mathrm{TV}} = d(t).$$

The opposite inequality holds, since the set of probabilities on Ω includes the point masses.

Similarly, if α and β are two probabilities on Ω, then

$$\begin{aligned}\|\alpha P - \beta P\|_{\mathrm{TV}} &= \frac{1}{2}\sum_{z\in\Omega}\left|\alpha P(z) - \sum_{w\in\Omega}\beta(w)P(w,z)\right| \\ &\leq \frac{1}{2}\sum_{z\in\Omega}\sum_{w\in\Omega}\beta(w)|\alpha P(z) - P(w,z)| \\ &= \sum_{w\in\Omega}\beta(w)\frac{1}{2}\sum_{z\in\Omega}|\alpha P(z) - P(w,z)| \\ &= \sum_{w\in\Omega}\beta(w)\,\|\alpha P - P(w,\cdot)\|_{\mathrm{TV}} \\ &\leq \max_{w\in\Omega}\|\alpha P - P(w,\cdot)\|_{\mathrm{TV}}\,. \end{aligned} \tag{C.6}$$

{Eq:GetRid1}

Thus, applying with $\alpha = \mu$ and $\beta = \nu$ gives that

{Eq:TVExB1}

$$\|\mu P - \nu P\|_{\mathrm{TV}} \leq \max_{y\in\Omega}\|\mu P - P(y,\cdot)\|_{\mathrm{TV}}\,. \tag{C.7}$$

Applying (C.6) with $\alpha = \delta_y$, where $\delta_y(z) = \mathbf{1}_{\{z=y\}}$, and $\beta = \mu$ shows that

{Eq:TVExB2}

$$\|\mu P - P(y,\cdot)\|_{\mathrm{TV}} = \|P(y,\cdot) - \mu P\|_{\mathrm{TV}} \leq \max_{x\in\Omega}\|P(y,\cdot) - P(x,\cdot)\|_{\mathrm{TV}}\,. \tag{C.8}$$

Combining (C.7) with (C.8) shows that

$$\|\mu P - \nu P\|_{\mathrm{TV}} \leq \max_{x,y\in\Omega}\|P(x,\cdot) - P(y,\cdot)\|_{\mathrm{TV}}\,.$$

■

Solution to 5.3. This is a standard exercise in manipulation of sums and inequalities. Apply Proposition 5.2, expand the matrix multiplication, apply the triangle inequality, switch order of summation, and apply Proposition 5.2 once more:

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

■

Solution to 5.2. Define $A_n = n^{-1}\sum_{k=1}^{n}a_k$. Let $n_k \leq m < n_{k+1}$. Then

$$A_m = \frac{n_k}{m}A_{n_k} + \frac{\sum_{j=n_k+1}^{m}a_j}{m}.$$

Because $n_k/n_{k+1} \le m^{-1}n_k \le 1$, the ratio $m^{-1}n_k$ tends to 1. Thus the first term tends to a. If $|a_j| \le B$, then the absolute value of the second term is bounded by

$$B\frac{n_{k+1}-n_k}{n_k} \to 0.$$

Thus $A_m \to a$. ■

Solution to 5.5. The total variation distance obeys the triangle inequality, so

$$\|P(x,\cdot) - P(y,\cdot)\|_{\mathrm{TV}} \le \|P(x,\cdot) - \pi\|_{\mathrm{TV}} + \|P(y,\cdot) - \pi\|_{\mathrm{TV}}\,.$$

Clearly, for all $x, y \in \Omega$, the right-hand side is bounded above by

$$\max_{x\in\Omega}\|P(x,\cdot) - \pi\|_{\mathrm{TV}} + \max_{y\in\Omega}\|P(y,\cdot) - \pi\|_{\mathrm{TV}} = 2d(t).$$

Thus taking the maximum over pairs $x, y \in \Omega$ completes the solution. ■

Chapter 6

Solution to 6.1. Consider the following coupling of the chain started from x and the chain started from π: run the chains independently until the time τ when they meet, and then run them together. Recall that by aperiodicity and irreducibility, there is some r so that $\alpha := \min_{x,y} P^r(x,y) \ge 0$.

Fix some state x_0. Then the probability that both chains, starting from say x and y, are not at x_0 after r steps is at most $(1-\alpha)$. If the two chains are not at x_0 after these r steps, the probability that they are not both at x_0 after another r steps is again $(1-\alpha)$. Continuing in this way, we get that $\mathbf{P}\{\tau > kr\} \le (1-\alpha)^k$. This shows that $\mathbf{P}\{\tau < \infty\} = 1$. ■

Solution to 6.2. We show that

$$\mathbf{P}\{\tau_{\mathrm{couple}} > kt_0\} \le (1-\alpha)^k, \tag{C.9}$$ {Eq:CoupleTimeGeo}

from which the conclusion then follows by summing. An *unsuccessful coupling attempt* occurs at trial j if $X_t \ne Y_t$ for all $jt_0 < t \le (j+1)t_0$. Since (X_t, Y_t) is a Markovian coupling, so is (X_{t+jt_0}, Y_{t+jt_0}) for any j, and we can apply the given bound on the probability of not coupling to any length-t_0 segment of the trajectories. Hence the probability of an unsuccessful coupling attempt at trial j is at most $(1-\alpha)$. It follows that the probability that all the first k attempts are unsuccessful is at most $(1-\alpha)^k$. ■

Solution to 6.4. If τ_i is the coupling time of the ith coordinate, we have seen already that $\mathbf{E}(\tau_i) \le n^2/4$, so

$$\mathbf{P}\{\tau_i > dn^2\} \le \frac{\mathbf{E}(\tau_i)}{kdn^2} \le \frac{1}{4}.$$

Suppose that $\mathbf{P}\{\tau_i > (k-1)dn^2\} \le 4^{-(k-1)}$. Then

$$\begin{aligned}\mathbf{P}\{\tau_i > kdn^2\} &= \mathbf{P}\{\tau_i > kdn^2 \mid \tau_i > (k-1)dn^2\}\mathbf{P}\{\tau_i > (k-1)dn^2\}\\ &\le 4^{-1}4^{-(k-1)}\\ &= 4^{-k}.\end{aligned}$$

Letting $G_i = \{\tau_i > kdn^2\}$, we have $\mathbf{P}(G_i) \leq 4^{-1}$. Thus

$$\mathbf{P}\left\{\max_{1\leq i\leq d} \tau_i > kdn^2\right\} \leq \mathbf{P}\left(\bigcup_{i=1}^{d} G_i\right) \leq \sum_{i=1}^{d} \mathbf{P}(G_i) \leq d4^{-k}.$$

Taking $k = (1/2)\log_2(4d)$ makes the right-hand side equal $(1/4)$. Thus

$$t_{\mathrm{mix}} \leq (1/2)[\log_2(4d)]dn^2 = O([d\log_2 d]n^2).$$

■

Chapter 7.

Solution to 7.1. From any ordering of the cards, the shuffle can move to exactly n orderings, each with probability n^{-1}. Furthermore, each ordering of the deck has exactly n possible predecessors. Consequently, because

$$\sum_{k=1}^{n} \frac{1}{n!}\frac{1}{n} = \frac{1}{n!},$$

it follows that $\pi = \pi P$ where π is the uniform distribution. ■

Solution to 7.4. For $x = (x_1, \ldots, x_d), y = (y_1, \ldots, y_d) \in \mathbb{Z}_n^d$, let

$$\phi_{x,y}(z_1, z_2, \ldots, z_d) = (z_1 + y_1 - x_1 \mod n, \ldots, z_d + y_d - x_d \mod n).$$

Clearly, $\phi_{x,y}(x) = y$.

Consider $z = (z_1, z_2, \ldots, z_d)$ and $z' = (z_1, \ldots, z_i + \delta \mod n, \ldots, z_n)$, where $\delta \in \{+1, -1\}$. The only transitions for the chain are of the form $z \to z'$ and $z \to z$. Since $\phi_{x,y}(z)$ and $\phi_{x,y}(z')$ also differ exactly in the ith coordinate by ±1,

$$P(z, z') = \frac{1}{2d} = P(\phi_{x,y}(z), \phi_{x,y}(z')).$$

■

Solution to 7.9. By Exercise 7.8,

$$s(t) = s\left(t_0\frac{t}{t_0}\right) \leq s(t_0)^{t/t_0}.$$

Since $s(t_0) \leq \varepsilon$ by hypothesis, applying Lemma 7.5 finishes the solution. ■

Solution to 7.3. Let $\varepsilon := [2(2n-1)]^{-1}$. Let $\mu(v) = (2n-1)^{-1}$. For $v \neq v^\star$,

$$\begin{aligned}\sum_{w} \mu(w)P(w, v) &= \sum_{\substack{w\,:\,w\sim v\\ w\neq v}} \frac{1}{(2n-1)}\left[\frac{1}{2} - \varepsilon\right]\frac{1}{n-1} + \frac{1}{(2n-1)}\left[\frac{1}{2} + \varepsilon\right] \\ &= \frac{1}{(2n-1)}\left\{(n-1)\left[\frac{1}{2} - \varepsilon\right]\frac{1}{n-1} + \left[\frac{1}{2} + \varepsilon\right]\right\} \\ &= \frac{1}{2n-1}\end{aligned}$$

Also,

$$\sum_w \mu(w)P(w,v^\star) = (2n-2)\frac{1}{2n-1}\left[\frac{1}{2}-\varepsilon\right]\frac{1}{n-1} + \frac{1}{2n-1}\left(\frac{1}{2n-1}\right)$$
$$= \frac{1}{2n-1}$$

■

Solution to 7.10. Following the hint and taking expectations,

$$\mathbf{E}\left(\sum_{t=1}^{\tau} Y_t\right) = \sum_{t=1}^{\infty}\mathbf{E}\left(Y_t\mathbf{1}_{\{\tau\geq t\}}\right). \tag{C.10}$$ {Eq:WE1}

Since the event $\{\tau \geq t\}$ is by assumption independent of Y_t, and $\mathbf{E}(Y_t) = \mathbf{E}(Y_1)$ for all $t \geq 1$, the right-hand side equals

$$\sum_{t=1}^{\infty}\mathbf{E}\left(Y_1\right)\mathbf{P}\{\tau \geq t\}.$$

The conclusion then follows by Exercise 3.12(a).

Now suppose that τ is a stopping time. For each t,

$$\mathbf{1}_{\{\tau\geq t\}} = 1 - \mathbf{1}_{\{\tau\leq t-1\}} = g_t(Y_1, Y_2, \ldots, Y_{t-1})$$

for some function $g_t : \mathbb{R}^t \to \{0,1\}$. Since the sequence (Y_t) is i.i.d. and $\mathbf{1}_{\{\tau\geq t\}}$ is a function of $Y_1, \ldots, Y_{t-1}$, the indicator is independent of Y_t.

■

Solution to 7.11. Let A be the set of vertices in one of the complete graphs making up G. Clearly, $\pi(A) = n/(2n-1) \geq 2^{-1}$.

On the other hand, for $x \notin A$,

$$P^t(x,A) = 1 - (1-\alpha_n)^t \tag{C.11}$$

where

$$\alpha_n = \frac{1}{2}\left[1 - \frac{1}{2(n-1)}\right]\frac{1}{n-1} = \frac{1}{2n}\left[1 + o(1)\right].$$

The total variation distance can be bounded below:

$$\left\|P^t(x,\cdot) - \pi\right\|_{\mathrm{TV}} \geq \pi(A) - P^t(x,A) \geq (1-\alpha_n)^t - \frac{1}{2}. \tag{C.12}$$

Since

$$\log(1-\alpha_n)^t \geq t(-\alpha_n - \alpha_n^2/2),$$

and $-1/4 \geq \log(3/4)$, if $t < [4\alpha_n(1-\alpha_n/2)]^{-1}$, then

$$(1-\alpha_n)^t - \frac{1}{2} \geq \frac{1}{4}.$$

This implies that

$$t_{\mathrm{mix}}(1/4) \geq \frac{n}{2}\left[1 + o(1)\right].$$

■

Solution to 7.6.

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

■

Solution to 7.7. Let ϕ be the function which maps $y \mapsto x$ and preserves P. Then

$$\hat{P}(z, w) = \frac{\pi(w)P(w, z)}{\pi(z)} = \frac{\pi(w)P(\phi(w), \phi(z))}{\pi(z)} = \hat{P}(w, z). \tag{C.13}$$

Note the last equality follows since π is uniform, and so $\pi(x) = \pi(\phi(x))$ for all x. ■

Chapter 14

Solution to 14.2. Notice that

$$\mathbf{P}\{X_1 = x_1\} = \sum_{y_1 \in \Omega} \mathbf{P}\{X_1 = x_1, Y_1 = y_1\}.$$

By condition on the values of X_0 and Y_0, this equals

$$\sum_{y_1 \in \Omega} \sum_{(x_0, y_0) \in \Omega \times \Omega} \mathbf{P}\{X_1 = x_1, Y_1 = y_1 \mid X_0 = x_0, Y_0 = y_0\}\mathbf{P}\{X_0 = x_0, Y_0 = y_0\}.$$

Changing the order of summation, the above is

$$\sum_{(x_0, y_0) \in \Omega \times \Omega} \left[\sum_{y_1 \in \Omega} \mathbf{P}\{X_1 = x_1, Y_1 = y_1 \mid X_0 = x_0, Y_0 = y_0\}\right] \mathbf{P}\{X_0 = x_0, Y_0 = y_0\}.$$

The conditional distribution of (X_1, Y_1) given $X_0 = x_0, Y_0 = y_0$ is a coupling of $P(x_0, \cdot)$ and $P(y_0, \cdot)$, so the inner sum above is $P(x_0, x_1)$. Thus,

{Eq:OS2}

$$\begin{aligned}\mathbf{P}\{X_1 = x\} &= \sum_{(x_0, y_0) \in \Omega \times \Omega} P(x_0, x_1)\mathbf{P}\{X_0 = x_0, Y_0 = y_0\}\\ &= \sum_{x_0 \in \Omega} P(x_0, x_1) \sum_{y_0 \in \Omega} \mathbf{P}\{X_0 = x_0, Y_0 = y_0\}\\ &= \sum_{x_0 \in \Omega} P(x_0, x_1)\mu(x_0) \qquad (C.14)\\ &= (\mu P)(x_1).\end{aligned}$$

The equality in (C.14) follows since (X_0, Y_0) is a coupling of μ and ν.

This shows that X_1 has distribution μP. The argument that Y_1 has distribution νP is similar. ■

Solution to 14.4. If $\operatorname{lip}(f) \leq 1$ and (X, Y) is a coupling of μ and ν attaining the minimum in the definition of Kantorovich distance, then

$$\left| \int f d\mu - \int f d\nu \right| = |\mathbf{E}\left(f(X) - f(Y)\right)| \leq \mathbf{E}\left(\rho(X, Y)\right) = \rho_K(\mu, \nu),$$

where we used $\operatorname{lip}(f) \leq 1$ for the inequality and the fact that (X, Y) is the optimal coupling for the last equality. ∎

Chapter 8

Solution to 8.1. Let $Y_t^i = 2X_t^i - 1$. Since covariance is bilinear, $\operatorname{Cov}(Y_t^i, Y_t^j) = 4\operatorname{Cov}(X_t^i, X_t^j)$ and it is enough to check that the $\operatorname{Cov}(Y_t^i, Y_t^j) \leq 0$.

If the ith coordinate is chosen in the first t steps, the conditional expectation of Y_t^i is 0. Thus

$$\mathbf{E}(Y_t^i) = \left(1 - \frac{1}{n}\right)^t.$$

Similarly,

$$\mathbf{E}(Y_t^i Y_t^j) = \left(1 - \frac{2}{n}\right)^t$$

since we only have a positive contribution if both the coordinates i, j were not chosen in the first t steps. Finally,

$$\begin{aligned}\operatorname{Cov}\left(Y_t^i, Y_t^j\right) &= \mathbf{E}\left(Y_t^i Y_t^j\right) - \mathbf{E}\left(Y_t^i\right)\mathbf{E}\left(Y_t^j\right)\\ &= \left(1 - \frac{2}{n}\right)^t - \left(1 - \frac{1}{n}\right)^{2t}\\ &\leq 0,\end{aligned}$$

because $(1 - 2/n) < (1 - 1/n)^2$.

The variance of the sum $W_t = \sum_{i=1}^n X_t^i$ is

$$\operatorname{Var}(N_t) = \sum_{i=1}^n \operatorname{Var}(X_t^i) + \sum_{i \neq j} \operatorname{Cov}(X_t^i, X_t^j) \leq \sum_{i=1}^n \frac{1}{4}.$$

∎

Solution to 9.1. Suppose that the reflected walk hits c at or before time n. It has probability at least $1/2$ of finishing at time n in $[c, \infty)$. (The probability can be larger than $1/2$ because of the reflecting at 0.) Thus

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

∎

SOLUTION TO 8.2.

$$\begin{aligned}
Q(S,S^c) &= \sum_{x\in S}\sum_{y\in S^c}\pi(x)P(x,y)\\
&= \sum_{y\in S^c}\left[\sum_{x\in\Omega}\pi(x)P(x,y) - \sum_{x\in S^c}\pi(x)P(x,y)\right]\\
&= \sum_{y\in S^c}\sum_{x\in\Omega}\pi(x)P(x,y) - \sum_{x\in S^c}\pi(x)\sum_{y\in S^c}P(x,y)\\
&= \sum_{y\in S^c}\pi(y) - \sum_{x\in S^c}\pi(x)\left[1-\sum_{y\in S}P(x,y)\right]\\
&= \sum_{y\in S^c}\pi(y) - \sum_{x\in S^c}\pi(x) + \sum_{x\in S^c}\sum_{y\in S}\pi(x)P(x,y)\\
&= \sum_{x\in S^c}\sum_{y\in S}\pi(x)P(x,y)\\
&= Q(S^c,S).
\end{aligned}$$

■

SOLUTION TO 8.3. Suppose that a graph G has vertex set V and diameter ρ.

Let $D_k = \{v : d(v,x_0) = k\}$ be all vertices at distance exactly k from x_0. If $v \in D_k$, then $\{v,w\}$ is an edge for some $w \in D_{k-1}$. (Take w to be the vertex connected to v in the minimal path from x_0 to v. Since there is a path from w to x_0 of length $k-1$, it must be that $d(w,x_0) \leq k-1$. If $d(w,x_0) < k-1$, then there is a path from x_0 to v of length strictly smaller than k and $d(x_0,v) \leq k-1$. Therefore, $w \in D_{k-1}$.) It follows that the set of vertices connected by edges to vertices in D_{k-1} contains D_k, so $|D_k| \leq \Delta|D_{k-1}|$. By induction, $|D_k| \leq \Delta^k$, and provided $\Delta \geq 2$,

$$|V| \leq \sum_{k=0}^{\rho}\Delta^k \leq \Delta^{\rho+1}.$$

Taking logarithms shows that $\log|V|/\log\Delta \leq \rho+1$. ■

SOLUTION TO 8.4. Write $\{v_1,\ldots,v_n\}$ be the vertices of the graph, and let (X_t) be the Markov chain started with the initial configuration $\vec{q}$ in which every vertex has color q.

Let $N : \Omega \to \{0,1,\ldots,n\}$ be the number of sites in the configuration x colored with q. That is,

$$N(x) = \sum_{i=1}^{n}\mathbf{1}_{\{x(v_i)=q\}}. \tag{C.15}$$

We write N_t for $N(X_t)$.

We compare the mean and variance of the random variable N under the uniform measure π and under the measure $P^t(\vec{q},\cdot)$. (Note that the distribution of $N(X_t)$ equals the distribution of N under $P^t(\vec{q},\cdot)$.)

The distribution of N under the stationary measure π is Binomial with parameters n and $1/q$, implying

$$E_\pi(N) = \frac{n}{q}, \quad \mathrm{Var}_\pi(N) = n\frac{1}{q}\left(1 - \frac{1}{q}\right) \leq \frac{n}{4}.$$

Let $X_i(t) = \mathbf{1}_{\{X_t(v_i)=q\}}$, the indicator that vertex v_i has color q. Since $X_i(t) = 0$ if and only if vertex v_i has been updated at least once by time t and the latest of these updates is *not* to color q, we have

$$\mathbf{E}_{\vec{q}}(X_i(t)) = 1 - \left[1 - \left(1 - \frac{1}{n}\right)^t\right]\frac{q-1}{q} = \frac{1}{q} + \frac{q-1}{q}\left(1 - \frac{1}{n}\right)^t,$$

and

$$\mathbf{E}_{\vec{q}}(N_t) = \frac{n}{q} + \frac{n(q-1)}{q}\left(1 - \frac{1}{n}\right)^t.$$

Consequently,

$$\mathbf{E}_{\vec{q}}(N_t) - E_\pi(N) = \left(\frac{q-1}{q}\right)n\left(1 - \frac{1}{n}\right)^t.$$

The random variables $\{X_i(t)\}$ are negatively correlated; check that $Y_i = qX_i - (q-1)$ are negatively correlated as in the solution to Exercise 8.1. Thus,

$$\sigma^2 := \max\{\mathrm{Var}_{\vec{q}}(N_t), \mathrm{Var}_\pi(N)\} \leq \frac{n}{4},$$

and

$$\left|E_\pi(N) - \mathbf{E}_{\vec{q}}(N(X_t))\right| = \frac{n}{2}\left(1 - \frac{1}{n}\right)^t \geq \sigma\frac{2(q-1)}{q}\sqrt{n}\left(1 - \frac{1}{n}\right)^t.$$

Letting $r(t) = [2(q-1)/q]\sqrt{n}(1 - n^{-1})^t$,

$$\begin{aligned}\log(r^2(t)) &= 2t\log(1 - n^{-1}) + \frac{2(q-1)}{q}\log n \\ &\geq 2t\left(-\frac{1}{n} - \frac{1}{2n^2}\right) + \frac{2(q-1)}{q}\log n, \qquad \text{(C.16)}\end{aligned}$$

{Eq:RtHC2}

where the inequality follows from $\log(1-x) \geq -x - x^2/2$, for $x \geq 0$. As in the proof of Proposition 8.8, it is possible to find a $c(q)$ so that for $t \leq (1/2)n\log n - c(q)n$, the inequality $r^2(t) \geq 32/3$ holds. By Corollary **??**, $t_{\mathrm{mix}} \geq (1/2)n\log n - c(q)n$. ■

Chapter 9

Solution to 9.2. False! Consider, for example, the distribution that assigns weight $1/2$ each to the identity and to the permutation that lists the elements of $[n]$ in reverse order. ■

Solution to 9.3. False! Consider, for example, the distribution that puts weight $1/n$ on all the cyclic shifts of a sorted deck: $123\ldots n, 23\ldots n1, \ldots, n12\ldots n{-}1$. ■

Solution to 9.6. By Cauchy-Schwarz, for any permutation $\sigma \in \mathcal{S}_n$ we have

$$\phi_\sigma = \sum_{k\in[n]} \phi(k)\phi(\sigma(k)) \leq \left(\sum_{k\in[n]} \phi(k)^2\right)^{1/2} \left(\sum_{k\in[n]} \phi(\sigma(k))^2\right)^{1/2} = \phi(\mathrm{id}).$$

■

Solution to 9.7. By the half-angle identity $\cos\theta = (\cos(2\theta) - 1)/2$, we have

$$\sum_{k\in[n]} \cos^2\left(\frac{(2k-1)\pi}{2n}\right) = \frac{1}{2}\sum_{k\in[n]}\left(\cos\left(\frac{(2k-1)\pi}{n}\right) + 1\right).$$

Now,

$$\sum_{k\in[n]} \cos\left(\frac{(2k-1)\pi}{n}\right) = \mathrm{Re}\left(e^{-\pi/n}\sum_{k\in[n]} e^{2k\pi/n}\right) = 0,$$

since the sum of the n-th roots of unity is 0. Hence

$$\sum_{k\in[n]} \cos^2\left(\frac{(2k-1)\pi}{2n}\right) = \frac{n}{2}.$$

■

Solution to 9.8. (a) Just as assigning t independent bits is the same as assigning a number chosen uniformly from $\{0, \ldots, 2^n - 1\}$ (as we implicitly argued in the proof of Proposition 9.6), assigning a digit in base a, and then a digit in base b, is the same as assigning a digit in base ab.

(b) To perform a forwards a-shuffle, divide the deck into a multinomially-distributed stacks, then uniformly choose an arrangement from all possible permutations that preserve the relative order within each stack. The resulting deck has at most a rising sequences, and there are a^n ways to divide, then riffle together (some of which can lead to identical permutations).

Given a permutation π with $r \leq a$ rising sequences, we need to count the number of ways it could possibly arise from a deck divided into a parts. Each rising sequence is a union of stacks, so the rising sequences together determine the positions of $r - 1$ out of the $a - 1$ dividers between stacks. The remaining $a - r$ dividers can be placed in any of the $n + 1$ possible positions, repetition allowed, irrespective of the positions of the $r - 1$ dividers already determined.

For example: set $a = 5$ and let $\pi \in \mathcal{S}_9$ be 152738946. The rising sequences are $(1, 2, 3, 4)$, $(5, 6)$, and $(7, 8, 9)$, so there must be packet divisions between 4 and 5, and between 6 and 7, and two additional dividers must be placed.

This is a standard choosing-with-repetition scenario. We can imagine building a row of length $n + (a - r)$ objects, of which n are numbers and $a - r$ are dividers. There are $\binom{n+a-r}{n}$ such rows.

Since each (division, riffle) pair has probability $1/a^n$, the probability that π arises from an a-shuffle is exactly $\binom{n+a-r}{n}/a^n$.

■

Solution to 10.3. Solution for (a): Let X_t be the number of umbrellas at home after t one-way trips, so $X_0 = k$, and write $Y_t = 2X_t + t \mod 2$. Then $Y_0 = 2k$ and Y_t evolves as simple random walk on the integers. Part (i) reduces to the mean time for Y_t to hit $\{1, 2n\}$, which is $(2k-1)(2n-2k)$. Part (ii) is a bit tricky: it reduces to the mean time for Y_t to hit $\{-1, 2n+2\}$, which is $(2k+1)(2n+2-2k)$. (Consider one extra tattered umbrella at each location, used as a last resort.)

Part (b) requires writing out the linear equations. ■

Solution to 10.5. Using the series law, $\mathcal{R}(a \leftrightarrow x) = x$ and $\mathcal{R}(a \leftrightarrow n) = n$. ■

Solution to 10.4. Let τ_A be the first time the walk visits a vertex in A. Check that $g(x) = \mathbf{E}_x(h(X_{\tau_A}))$ is harmonic for $x \in V \setminus A$. Uniqueness follows by extending Proposition 10.1. ■

Solution to 10.8. Let W_1 b e a voltage function for the unit current flow from x to y so that $W_1(x) = \mathcal{R}(x \leftrightarrow y)$ and $W_1(y) = 0$. Let W_2 be a voltage function for the unit current flow from y to z so that $W_2(y) = \mathcal{R}(y \leftrightarrow z)$ and $W_2(z) = 0$. By harmonicity (the maximum principle) at all vertices v we have

$$0 \le W_1(v) \le \mathcal{R}(x \leftrightarrow y) \tag{C.17}$$ {Eq:RMSt1}

$$0 \le W_1(v) \le \mathcal{R}(y \leftrightarrow z) \tag{C.18}$$ {Eq:RMSt2}

Recall the hint. Thus $W_3 = W_1 + W_2$ is a voltage function for the unit current flow from x to z and

$$\mathcal{R}(x \leftrightarrow z) = W_3(x) - W_3(z) = \mathcal{R}(x \leftrightarrow y) + W_2(x) - W_1(z). \tag{C.19}$$ {Eq:RMSt3}

Applying (C.18) gives $W_2(x) \le \mathcal{R}(y \leftrightarrow z)$ and (C.17) gives $W_1(z) \ge 0$ so finally by (C.19) we get the triangle inequality. ■

Solution to 11.3.

(a) Use the fact that, since the B_j's partition B, $\mathbf{E}(Y \mid B) = \sum_j \mathbf{P}(B_j)\mathbf{E}(Y \mid B_j)$.

(b) Many examples are possible; a small one is $\Omega = B = \{1,2,3\}$, $Y = \mathbf{1}_{\{1,3\}}$, $B_1 = \{1,2\}$, $B_2 = \{2,3\}$, $M = 1/2$.

■

Solution to 11.4.

(a) Let σ be a uniform random permutation of the elements of A. Let T_k be the first time at which all of $\sigma(1), \sigma(2), \ldots, \sigma(k)$ have been visited, and let $L_k = X_{T_k}$.

With probability $1/|A|$, $\sigma(1) = x$ and $T_1 = 0$. Otherwise, the walk must proceed from x to $\sigma(1)$. Thus

$$\begin{aligned} \mathbf{E}_x(T_1) &\ge \frac{1}{|A|}0 + \frac{|A|-1}{|A|}T^A_{\text{min}} \\ &= \left(1 - \frac{1}{|A|}\right)T^A_{\text{min}}. \end{aligned}$$

For $2 \le k \le |A|$ and $r, s \in A$, define

$$B_k(r,s) = \{\sigma(k-1) = r, \sigma(k) = L_k = s\},$$

so that

$$\mathbf{E}_x(T_k - T_{k-1} \mid B_k(r, s)) = \mathbf{E}_r\tau_s.$$

Then

$$B_k = \bigcup_{r,s\in A} B_k(r, s)$$

is the event that $L_k = \sigma(k)$. By (an obvious corollary to) Exercise 11.3,

$$\mathbf{E}_x(T_k - T_{k-1} \mid B_k^c) = 0 \quad \text{and} \quad \mathbf{E}_x(T_k - T_{k-1}|B_k) \geq T^A_{\min}.$$

By symmetry, $\mathbf{P}(B_k) = 1/k$, so $\mathbf{E}_x(T_k - T_{k-1}) \geq (1/k)T^A_{\min}$. Adding all these bounds gives the final result (note how the negative portion of the first term cancels out the last term).

(b) Clearly $\mathbf{E}_x(C) \geq \mathbf{E}_x(C_A)$ for every $A \subseteq \mathcal{X}$.

■

Solution to 11.9.

(a) An edge is defined by which coordinate flips. There are m coordinates to choose and then 2^{m-1} possibilities for assigning values to the other coordinates.

(b) There are $\binom{m}{k}$ nodes of weight k.

■

Solution to 11.10. Observe that $h_m(k)$ is the mean hitting time from k to 0 in G_k, which implies that $h_m(k)$ is monotone increasing in k. (This is intuitively clear but harder to prove directly on the cube.) The expected return time from o to itself in the hypercube equals 2^m but considering the first step it also equals $1 + h_m(1)$. Thus

{Eq:CHStar}
$$h_m(1) = 2^m - 1. \tag{C.20}$$

To compute $h_m(m)$ use symmetry and the commute time identity. The effective resistance between 0 and m in G_m is $\mathcal{R}(0 \leftrightarrow m) = \sum_{k=1}^{m}[k\binom{m}{k}]^{-1}$. In this sum all but the first and last terms are negligible: The sum of the other terms is at most $4/m^2$ (check!). Thus

$$2h_m(m) = 2\mathcal{R}(0 \leftrightarrow m)|\text{edges}(G_m)| \leq 2\left(\frac{2}{m} + \frac{4}{m^2}\right)(m2^{m-1}),$$

so

{Eq:CH2Star}
$$h_m(m) \leq 2^m(1 + 2/m). \tag{C.21}$$

(C.20) together with (C.21) and monotonicity concludes the proof. ■

Solution to 11.12. By Lemma 11.8,

$$\begin{aligned} 2\mathbf{E}_a(\tau_{bca}) &= [\mathbf{E}_a(\tau_b) + \mathbf{E}_b(\tau_c) + \mathbf{E}_c(\tau_a)] + [\mathbf{E}_a(\tau_c) + \mathbf{E}_c(\tau_b) + \mathbf{E}_b(\tau_a)] \\ &= [\mathbf{E}_a(\tau_b) + \mathbf{E}_b(\tau_a)] + [\mathbf{E}_b(\tau_c) + \mathbf{E}_c(\tau_b)] + [\mathbf{E}_c(\tau_a) + \mathbf{E}_a(\tau_c)]\,. \end{aligned}$$

Then the conclusion follows from Proposition 11.6. ■

Solution to 11.13. Taking expectations in (11.33) yields

$$\mathbf{E}_x(\tau_a) + \mathbf{E}_a(\tau_z) = \mathbf{E}_x(\tau_z) + \mathbf{P}_x\{\tau_z < \tau_a\}\left[\mathbf{E}_z(\tau_a) + \mathbf{E}_a(\tau_z)\right],$$

which shows that

{Eq:HT2}

$$\mathbf{P}_x\{\tau_z < \tau_a\} = \frac{\mathbf{E}_x(\tau_a) + \mathbf{E}_a(\tau_z) - \mathbf{E}_x(\tau_z)}{\mathbf{E}_z(\tau_a) + \mathbf{E}_a(\tau_z)}, \tag{C.22}$$

without assuming reversibility.

In the reversible case, the cycle identity (Lemma 11.8) yields

$$\mathbf{E}_x(\tau_a) + \mathbf{E}_a(\tau_z) - \mathbf{E}_x(\tau_z) = \mathbf{E}_a(\tau_x) + \mathbf{E}_z(\tau_a) - \mathbf{E}_z(\tau_x). \tag{C.23}$$

{Eq:HT3}

Adding the two sides of (C.23) together establishes that

$$\begin{aligned}&\mathbf{E}_x(\tau_a) + \mathbf{E}_a(\tau_z) - \mathbf{E}_z(\tau_z)\\ &\qquad = \frac{1}{2}\left\{[\mathbf{E}_x(\tau_a) + \mathbf{E}_a(\tau_x)] + [\mathbf{E}_a(\tau_z) + \mathbf{E}_z(\tau_a)] - [\mathbf{E}_x(\tau_z) + \mathbf{E}_z(\tau_x)]\right\}.\end{aligned}$$

Let $c_G = \sum_{x \in V} c(x) = 2\sum_e c(e)$, as usual. Then by the commute time formula (Proposition 11.6), the denominator in (C.22) is $c_G \mathcal{R}(a \leftrightarrow z)$ and the numerator is $(1/2)c_G\left[\mathcal{R}(x \leftrightarrow a) + \mathcal{R}(a \leftrightarrow z) - \mathcal{R}(z \leftrightarrow x)\right]$. ■

Solution to 11.15.

$$\begin{aligned}\sum_{k=0}^{\infty} c_k s^k &= \sum_{k=0}^{\infty}\sum_{j=0}^{k} a_j b_{k-j} s^k\\ &= \sum_{k=0}^{\infty}\sum_{j=0}^{\infty} a_j s^j b_{k-j} s^{k-j}\mathbf{1}_{\{k \geq j\}}\\ &= \sum_{j=0}^{\infty}\sum_{k=0}^{\infty} a^j s^j b_{k-j} s^{k-j}\mathbf{1}_{\{k \geq j\}}\\ &= \sum_{j=0}^{\infty} a^j s^j \sum_{k=0}^{\infty} b_{k-j} s^{k-j}\mathbf{1}_{\{k \geq j\}}\\ &= \sum_{j=0}^{\infty} a^j s^j \sum_{\ell=0}^{\infty} b_\ell s^\ell\\ &= A(s)B(s).\end{aligned}$$

The penultimate equality follows from letting $\ell = k - j$. The reader should check that the change of the order of summation is justified. ■

Chapter 12

Solution to 12.1.

(a) For any function f,

$$\|Pf\|_\infty = \max_{x \in \Omega}\left|\sum_{y \in \Omega} P(x,y) f(y)\right| \leq \|f\|_\infty.$$

If $P\varphi = \lambda\varphi$, then $\|Pf\|_\infty = |\lambda|\,\|f\|_\infty \le \|f\|_\infty$. This implies that $|\lambda| \le 1$.

(b) By the Convergence Theorem, $\lim_{t\to\infty} P^t = \pi$ (pointwise), where by an abuse of notation π denotes the matrix with all rows equal to the vector π.

Suppose that λ is an eigenvalue satisfying $|\lambda| = 1$ and with corresponding eigenvalue φ. Then $\lambda^t\varphi = P^t\varphi \to \pi\varphi$. If $\lambda = -1$ then the left-hand side does not converge to anything, a contradiction. Therefore we can assume that $\lambda = 1$, in which case $\varphi = \pi\varphi$. Writing out this vector equality,

$$\varphi(x) = \sum_{y\in\Omega} \pi(y)\varphi(y) \quad \text{for all } x \in \Omega.$$

In particular, φ does not depend on x, and must be constant. In summary, any eigenvector with $|\lambda| = 1$ is a multiple of $\mathbf{1}$, showing that $|\lambda_j| < 1$ for $j = 2, \dots, |\Omega|$.

■

Solution to 12.2. Let f be an eigenvector of P with eigenvalue μ. Then

$$\mu f = \tilde{P} f = \frac{Pf + f}{2}.$$

Rearranging shows that $(2\mu - 1)$ is an eigenvalue of P. Thus $2\mu - 1 \ge -1$, or equivalently, $\mu \ge 0$. ■

Solution to 12.4. According to (12.4),

$$\frac{P^{2t+2}(x,x)}{\pi(x)} = \sum_{j=1}^{|\Omega|} f_j(x)^2 \lambda_j^{2t+2}.$$

Since $\lambda_j^2 \le 1$ for all j, the right-hand side is bounded above by $\sum_{j=1}^{|\Omega|} f_j(x)^2\lambda_j^{2t}$, which equals $P^{2t}(x,x)/\pi(x)$. ■

Solution to 12.6. A computation verifies the claim:

$$\begin{aligned}(P_1 \otimes P_2)(\phi \otimes \psi)(x,y) &= \sum_{(z,w)\in\Omega_1\times\Omega_2} P_1(x,z)P_2(y,w)\phi(z)\psi(w) \\ &= \sum_{z\in\Omega_1} P_1(x,z)\phi(z) \sum_{w\in\Omega_2} P_2(y,w)\psi(w) \\ &= [P_1\phi(x)]\,[P_2\psi(y)] \\ &= \lambda\mu\phi(x)\psi(y) \\ &= \lambda\mu(\phi\otimes\psi)(x,y).\end{aligned}$$

That is, the product $\lambda\mu$ is an eigenvalue of the eigenfunction $\phi\otimes\psi$. ■

Chapter 13

Solution to 13.3. We bound $\binom{n}{\delta k} \leq n^{\delta k}/(\delta k)!$, similarly $\binom{(1+\delta)k}{\delta k}$ and $\binom{n}{k} \geq n^k/k^k$. This gives

$$\sum_{k=1}^{n/2} \frac{\binom{n}{\delta k}\binom{(1+\delta)k}{\delta k}^2}{\binom{n}{k}} \leq \sum_{k=1}^{n/2} \frac{n^{\delta k}((1+\delta)k)^{2\delta k}k^k}{(\delta k)!^3 n^k}.$$

Recall that for any integer ℓ we have $\ell! > (\ell/e)^\ell$, and bound $(\delta k)!$ by this. We get

$$\sum_{k=1}^{n/2} \frac{\binom{n}{\delta k}\binom{(1+\delta)k}{\delta k}^2}{\binom{n}{k}} \leq \sum_{k=1}^{\log n} \left(\frac{\log n}{n}\right)^{(1-\delta)k} \left[\frac{e^3(1+\delta)^2}{\delta^3}\right]^{\delta k} + \sum_{k=\log n}^{n/2} \left(\frac{k}{n}\right)^{(1-\delta)k} \left[\frac{e^3(1+\delta)^2}{\delta^3}\right]^{\delta k}.$$

The first sum clearly tends to 0 as n tends to ∞, for any $\delta \in (0,1)$, and since $k/n \leq 1/2$ and

$$(1/2)^{(1-\delta)} \left[\frac{e^3(1+\delta)^2}{\delta^3}\right]^{\delta} < 0.9$$

for $\delta < 0.01$, for any such δ the second sum tends to 0 as n tends to ∞. ■

Chapter 15

Solution to 15.2. Note that

$$\tanh'(\beta) = \frac{1}{\cosh^2(\beta)} = \frac{1}{1+\sinh^2(\beta)}.$$

Thus, $\tanh'(0) = 1$ and the derivative $\tanh'(\beta) \leq 1$ for $\beta > 0$, so $\tanh(\beta) \leq \beta$ for all $\beta > 0$. ■

Chapter 18

Solution to 18.1. We can write $X_t = x + \sum_{s=1}^{t} Y_s$, where $x \in \Omega$ and $(Y_s)_{s=1}^{\infty}$ is an i.i.d. sequence of $\{-1,1\}$-valued random variables satisfying

$$\mathbf{P}\{Y_s = +1\} = p,$$
$$\mathbf{P}\{Y_s = -1\} = q.$$

By the Strong Law, $\mathbf{P}_0\{\lim_{t\to\infty} t^{-1}X_t = (p-q)\} = 1$. In particular,

$$\mathbf{P}_0\{X_t > (p-q)t/2 \text{ for } t \text{ sufficiently large}\} = 1.$$

That is, with probability one, there are only finitely many visits of the walker to 0. Since the number of visits to 0 is a geometric random variable with parameter $\mathbf{P}_0\{\tau_0^+ = \infty\}$ (see the proof of Proposition 18.3 below), this probability must be positive. ■

Solution to 18.2. Suppose that $\pi(v) = 0$. Since $\pi = \pi P$,

$$0 = \pi(v) = \sum_{u \in X} \pi(u)P(u,v).$$

Since all the terms on the right-hand side are non-negative, each is zero. That is, if $P(u,v) > 0$, it must be that $\pi(u) = 0$.

Suppose that there is some $y \in \Omega$ so that $\pi(y) = 0$. By irreducibility, for any $x \in \Omega$, there is a sequence $u_0, \ldots, u_t$ so that $u_0 = x$, $u_t = y$, and each $P(u_{i-1}, u_i) > 0$

for $i = 1, \dots, t$. Then by induction it is easy to see that $\pi(u_i) = 0$ for each of $i = 0, 1, 2, \dots, t$. Thus $\pi(x) = 0$ for all $x \in \Omega$, and π is not a probability distribution. ∎

Solution to 18.4. If the original graph is regarded as a network with conductances $c(e) = 1$ for all e, then the subgraph is also a network, but with $c(e) = 0$ for all edges which are omitted. By Rayleigh's Monotonicity Law, the effective resistance from a fixed vertex v to ∞ is not smaller in the subgraph than for the original graph. This together with Proposition 18.6 shows that the subgraph must be recurrent. ∎

Solution to 18.5. Define

$$A_{x,y} = \{t \ : \ P^t(x,y) > 0\}.$$

By aperiodicity, g.c.d.$(A_{x,x}) = 1$. Since $A_{x,x}$ is closed under addition, there is some t_x so that $t \in A_{x,x}$ for $t \geq t_x$. Also, by irreducibility, there is some s so that $P^s(x,y) > 0$. Since

$$P^{t+s}(x,y) \geq P^t(x,x)P^s(x,y),$$

if $t \geq t_x$ then $t + s \in A_{y,x}$. That is, there exists $t_{x,y}$ so that if $t \geq t_{x,y}$ then $t \in A_{x,y}$.

Let $t_0 = \max\{t_{x,z}, \, t_{y,w}\}$. If $t \geq t_0$ then $P^t(x,z) > 0$ and $P^t(y,w) > 0$. In particular,

$$P^{t_0}((x,y), \, (z,w)) = P^{t_0}(x,z)P^{t_0}(y,w) > 0.$$

∎

Solution to 18.6. (X_t) is a nearest-neighbor random walk on $\mathbb{Z}^+$ which increases by 1 with probability α and decreases by 1 with probability $\beta = 1 - \alpha$. When the walker is at 0, instead of decreasing with probability β, it remains at 0. Thus if $\alpha < \beta$, then the chain is a downwardly biased random walk on $\mathbb{Z}^+$, which was shown in Example 18.15 to be positive recurrent.

If $\alpha = \beta$, this is an unbiased random walk on $\mathbb{Z}^+$. This is null recurrent for the same reason that the simple random walk on $\mathbb{Z}$ is null recurrent, shown in Example 18.10.

Consider the network with $V = \mathbb{Z}^+$, and with $c(k, k+1) = r^k$. If $r = p/(1-p)$, then the random walk on the network corresponds to a nearest-neighbor random walk which moves "up" with probability p. The effective resistence from 0 to n is

$$\mathcal{R}(0 \leftrightarrow n) = \sum_{k=1}^{n} r^{-k}.$$

If $p > 1/2$ then $r > 1$ and the right-hand side converges to a finite number, so $\mathcal{R}(0 \leftrightarrow \infty) < \infty$. By Proposition 18.6 this walk is transient. The FIFO queue of this problem is an upwardly biased random walk when $\alpha > \beta$, and thus it is transient as well. ∎

Solution to 18.7. Let $r = \alpha/\beta$. Then $\pi(k) = (1-r)r^k$ for all $k \geq 0$, that is, π is the geometric with probability r shifted by 1 to the left. Thus $E_\pi(X+1) = 1/(1-r) = \beta/(\beta - \alpha)$. Since $\mathbf{E}(T \mid X \text{ before arrival}) = (1+X)/\beta$, we conclude that $\mathbf{E}_\pi(T) = 1/(\beta - \alpha)$. ∎

Solution to 18.9. Suppose that $\mu = \mu P$, so that for all k,

$$\mu(k) = \frac{\mu(k-1) + \mu(k+1)}{2}.$$

The difference sequence $d(k) = \mu(k) - \mu(k-1)$ is easily seen to be constant, and hence μ is not bounded. ■

Chapter 17

Solution to 17.6. The distribution of a sum of n independent exponential random variables with rate λ has a Gamma distribution with parameters n and λ, so S_k has density

$$f_k(s) = \frac{s^{k-1}e^{-s}}{(k-1)!}.$$

Since S_k and X_{k+1} are independent,

$$\begin{aligned}\mathbf{P}\{S_k \le t < S_k + X_{k+1}\} &= \int_0^t \frac{s^{k-1}e^{-s}}{(k-1)!}\int_{t-s}^{\infty} e^{-x}dxds \\ &= \int_0^t \frac{s^{k-1}}{(k-1)!}e^{-t}ds \\ &= \frac{t^k e^{-t}}{k!}\end{aligned}$$

■

Solution to 17.1. Let $g(y,u)$ be the joint density of (Y, U_Y). Then

$$\begin{aligned}f_{Y,U}(y,u) &= f_Y(y)f_{U_Y|Y}(u|\,y) \\ &= g(y)\mathbf{1}\{g(y) > 0\}\frac{\mathbf{1}\{0 \le u \le Cg(y)\}}{Cg(y)} = \frac{1}{C}\mathbf{1}\{g(y) > 0, u \le Cg(y)\}. \quad \text{(C.24)}\end{aligned}$$ {Eq:UniformJoint}

This is the density for a point (Y, U) drawn from the region under the graph of the function g.

Conversely, let (Y, U) be a uniform point from the region under the graph of the function g. Its density is the right-hand side of (C.24). The marginal density of Y is

$$f_Y(y) = \int_{-\infty}^{\infty}\frac{1}{C}\mathbf{1}\{g(y) > 0, u \le Cg(y)\}du = \mathbf{1}\{g(y) > 0\}\frac{1}{C}Cg(y) = g(y). \quad \text{(C.25)}$$

■

Solution to 17.4. Let R be any region of TA. First, note that since $\text{rank}(T) = d$, by the Rank Theorem, T is one-to-one. Consequently, $TT^{-1}R = R$, and

$$\text{Volume}_d(R) = \text{Volume}_d(TT^{-1}R) = \sqrt{\det(T^tT)}\,\text{Volume}(T^{-1}R),$$

so that $\text{Volume}(T^{-1}R) = \text{Volume}_d(R)/\sqrt{\det(T^tT)}$. To find the distribution of Y, we compute

$$\mathbf{P}\{Y \in R\} = \mathbf{P}\{TX \in R\} = \mathbf{P}\{X \in T^{-1}R\}. \quad \text{(C.26)}$$

Since X is uniform, the right-hand side is

$$\frac{\text{Volume}(T^{-1}R)}{\text{Volume}(A)} = \frac{\text{Volume}_d(R)}{\sqrt{\det(T^tT)}\,\text{Volume}(A)} = \frac{\text{Volume}_d(R)}{\text{Volume}_d(TA)}. \tag{C.27}$$

■

Chapter 19

Solution to 19.1. Let (X_t) be simple random walk on $\mathbb{Z}$.

$$\begin{aligned} M_{t+1} - M_t &= (X_t + \Delta X_t)^3 - 3(t+1)(X_t + \Delta X_t) - X_t^3 + 3tX_t \\ &= 3X_t^2(\Delta X_t) + 3X_t(\Delta X_t)^2 + (\Delta X_t)^3 - 3t(\Delta X_t) - 3X_t - \Delta X_t \end{aligned}$$

Note that $(\Delta X_t)^2 = 1$, so

$$M_{t+1} - M_t = (\Delta X_t)(3X_t^2 - 3t),$$

and

$$\mathbf{E}_k\left(M_{t+1} - M_t \mid X_t\right) = (3X_t^2 - 3t)\mathbf{E}_k(\Delta X_t \mid X_t) = 0.$$

Using the Optional Stopping Theorem,

$$\begin{aligned} k^3 &= \mathbf{E}_k(M_\tau) \\ &= \mathbf{E}_k\left[\left(X_\tau^3 - 3\tau X_\tau\right)\mathbf{1}_{\{X_\tau = n\}}\right] \\ &= n^3\mathbf{P}_k\{X_\tau = n\} - 3n\mathbf{E}_k\left(\tau\mathbf{1}_{\{X_\tau=n\}}\right) \end{aligned}$$

Dividing through by $kn^{-1} = \mathbf{P}_k\{X_\tau = n\}$ shows that

$$nk^2 = n^3 - 3n\mathbf{E}_k\left(\tau \mid X_\tau = n\right).$$

Rearranging,

$$\mathbf{E}_k\left(\tau \mid X_\tau = n\right) = \frac{n^2 - k^2}{3}.$$

The careful reader will notice that we have used the Optional Stopping Theorem without verifying its hypotheses! The application can be justified by applying it to $\tau \wedge B$, and then letting $B \to \infty$ and appealing to the Dominated Convergence Theorem. ■

Bibliography

Aldous, D. 1983. *Random walks on finite groups and rapidly mixing Markov chains*, Seminar on probability, XVII, Lecture Notes in Math., vol. 986, Springer, Berlin, pp. 243–297.

Aldous, D. 1995. *On simulating a Markov chain stationary distribution when transition probabilities are unknown* (D. Aldous, P. Diaconis, J. Spencer, and J. M. Steele, eds.), IMA Volumes in Mathematics and its Applications, vol. 72, Springer-Verlag.

Aldous, D. and P. Diaconis. 1986. *Shuffling cards and stopping times*, Amer. Math. Monthly **93**, no. 5, 333–348.

______. 1987. *Strong uniform times and finite random walks*, Adv. in Appl. Math. **8**, no. 1, 69–97.

______. 2002. *The asymmetric one-dimensional constrained Ising model: rigorous results*, J. Statist. Phys. **107**, no. 5-6, 945–975.

Aldous, D. and J. Fill. *Reversible Markov chains and random walks on graphs*, in progress. Manuscript available at `http://www.stat.berkeley.edu/~aldous/RWG/book.html`.

Alon, N. 1986. *Eigenvalues and expanders*, Combinatorica **6**, no. 2, 83–96.

Alon, N. and V. D. Milman. 1985. λ_1, *isoperimetric inequalities for graphs, and superconcentrators*, J. Combin. Theory Ser. B **38**, no. 1, 73–88.

Archer, A. F. 1999. *A modern treatment of the 15 puzzle*, Amer. Math. Monthly **106**, no. 9, 793–799.

Artin, M. 1991. *Algebra*, Prentice Hall Inc., Englewood Cliffs, NJ.

Asmussen, S., P. Glynn, and H. Thorisson. 1992. *Stationary detection in the initial transient problem*, ACM Transactions on Modeling and Computer Simulation **2**, 130–157.

Baxter, R. J. 1982. *Exactly Solved Models in Statistical Mechanics*, Academic Press.

Bayer, D. and P. Diaconis. 1992. *Trailing the dovetail shuffle to its lair*, Ann. Appl. Probab. **2**, no. 2, 294–313.

Benjamin, A. T. and J. J. Quinn. 2003. *Proofs that really count: The art of combinatorial proof*, Dolciani Mathematical Expositions, vol. 27, Math. Assoc. Amer., Washington, D. C.

Billingsley, P. 1995. *Probability and measure*, 3rd ed., Wiley Series in Probability and Mathematical Statistics, John Wiley & Sons Inc., New York.

Borovkov, A. A. and S. G. Foss. 1992. *Stochastically recursive sequences and their generalizations*, Siberian Advances in Mathematics **2**, 16–81.

Bubley, R. and M. Dyer. 1997. *Path coupling: A technique for proving rapid mixing in Markov chains*, Proceedings of the 38th Annual Symposium on Foundations of Computer Science, pp. 223–231.

Cerf, R. and A. Pisztora. 2000. *On the Wulff crystal in the Ising model*, Ann. Probab. **28**, no. 3, 947–1017.

Cesi, F., G. Guadagni, F. Martinelli, and R. H. Schonmann. 1996. *On the two-dimensional stochastic Ising model in the phase coexistence region near the critical point*, J. Statist. Phys. **85**, no. 1-2, 55–102.

Chandra, A. K., P. Raghavan, W. L. Ruzzo, R. Smolensky, and P. Tiwari. 1996/97. *The electrical resistance of a graph captures its commute and cover times*, Comput. Complexity **6**, no. 4, 312–340. Extended abstract originally published in *Proc. 21st ACM Symp. Theory of Computing* (1989) 574–586.

Chayes, J. T., L. Chayes, and R. H. Schonmann. 1987. *Exponential decay of connectivities in the two-dimensional Ising model*, J. Statist. Phys. **49**, no. 3-4, 433–445.

Cheeger, J. 1970. *A lower bound for the smallest eigenvalue of the Laplacian*, Problems in analysis (Papers dedicated to Salomon Bochner, 1969), Princeton Univ. Press, Princeton, pp. 195–199.

Chen, F., L. Lovász, and I. Pak. 1998. *Lifting Markov chains to speed up mixing*. Unpublished.

Chen, M.-F. 1998. *Trilogy of couplings and general formulas for lower bound of spectral gap*, Probability towards 2000 (New York, 1995), Lecture Notes in Statist., vol. 128, Springer, New York, pp. 123–136.

Chyakanavichyus, V. and P. Vaĭtkus. 2001. *Centered Poisson approximation by the Stein method*, Liet. Mat. Rink. **41**, no. 4, 409–423 (Russian, with Russian and Lithuanian summaries); English transl.,. 2001, Lithuanian Math. J. **41**, no. 4, 319–329.

Dembo, A., Y. Peres, J. Rosen, and O. Zeitouni. 2004. *Cover times for Brownian motion and random walk in two dimensions*, Ann. Math. **160**, 433–464.

Devroye, L. 1986. *Nonuniform random variate generation*, Springer-Verlag, New York.

Diaconis, P. 1988. *Group Representations in Probability and Statistics*, Lecture Notes - Monograph Series, vol. 11, Inst. Math. Stat., Hayward, CA.

Diaconis, P. 2003. *Mathematical developments from the analysis of riffle shuffling*, Groups, combinatorics & geometry (Durham, 2001), World Sci. Publ., River Edge, NJ, pp. 73–97.

Diaconis, P. and D. Freedman. 1999. *Iterated random functions*, SIAM Review **41**, 45–76.

Diaconis, P., M. McGrath, and J. Pitman. 1995. *Riffle shuffles, cycles, and descents*, Combinatorica **15**, no. 1, 11–29.

Diaconis, P. and L. Saloff-Coste. 1996. *Nash inequalities for finite Markov chains*, J. Theoret. Probab. **9**, no. 2, 459–510.

Diaconis, P. and M. Shahshahani. 1981. *Generating a random permutation with random transpositions*, Z. Wahrsch. Verw. Gebiete **57**, no. 2, 159–179. MR **626813 (82h:**60024)

Diaconis, P. and D. Stroock. 1991. *Geometric bounds for eigenvalues of Markov chains*, Ann. Appl. Probab. **1**, no. 1, 36–61.

Doyle, P. G. and E. J. Snell. 1984. *Random walks and electrical networks*, Carus Math. Monographs, vol. 22, Math. Assoc. Amer., Washington, D. C.

Dobrushin, R., R. Kotecký, and S. Shlosman. 1992. *Wulff construction. A global shape from local interaction*, Translations of Mathematical Monographs, vol. 104, American Mathematical Society, Providence, RI. Translated from the Russian by the authors.

Dyer, M., C. Greenhill, and M. Molloy. 2002. *Very rapid mixing of the Glauber dynamics for proper colorings on bounded-degree graphs*, Random Structures Algorithms **20**, no. 1, 98–114.

Elias, P. 1972. *The efficient construction of an unbiased random sequence*, Ann. Math. Statist. **43**, 865–870.

Feller, W. 1968. *An introduction to probability theory and its applications*, third edition, Vol. 1, Wiley, New York.

Fill, J. A. 1991. *Eigenvalue bounds on convergence to stationarity for nonreversible Markov chains, with an application to the exclusion process*, Ann. Appl. Probab. **1**, no. 1, 62–87.

Fill, J. A. 1998. *An interruptible algorithm for perfect sampling via Markov chains*, Annals of Applied Probability **8**, 131–162.

Fill, J. A. and M. Huber. 2000. *The randomness recycler: A new technique for perfect sampling*, 41st Annual Symposium on Foundations of Computer Science, pp. 503–511.

Fill, J. A., M. Machida, D. J. Murdoch, and J. S. Rosenthal. 2000. *Extension of Fill's perfect rejection sampling algorithm to general chains*, Random Structure and Algorithms **17**, 290–316.

Graham, R. L., D. E. Knuth, and O. Patashnik. 1994. *Concrete mathematics: A foundation for computer science*, second edition, Addison Wesley, Reading, Massachusetts.

Häggström, O. and J. Jonasson. 1997. *Rates of convergence for lamplighter processes*, Stochastic Process. Appl. **67**, no. 2, 227–249. MR **1449833 (98j:**60097)

Häggström, O. and K. Nelander. 1998. *Exact sampling from anti-monotone systems*, Statist. Neerlandica **52**, no. 3, 360–380.

Hajek, B. 1988. *Cooling schedules for optimal annealing*, Math. Oper. Res. **13**, no. 2, 311–329.

Hayes, T. P. and A. Sinclair. 2005. *A general lower bound for mixing of single-site dynamics on graph*, available at arXiv:math.PR/0507517.

Herstein, I. N. 1975. *Topics in algebra*, 2nd ed., John Wiley and Sons, New York.

Horn, R. A. and C. R. Johnson. 1990. *Matrix analysis*, Cambridge University Press, Cambridge.

Huber, M. 1998. *Exact sampling and approximate counting techniques*, Proceedings of the 30th Annual ACM Symposium on the Theory of Computing, pp. 31–40.

Ioffe, D. 1995. *Exact large deviation bounds up to T_c for the Ising model in two dimensions*, Probab. Theory Related Fields **102**, no. 3, 313–330.

Ising, E. 1925. *Beitrag zur theorie der ferromagnetismus*, Zeitschrift Fur Physik **31**, 253–258.

Jerrum, M. R. 1995. *A very simple algorithm for estimating the number of k-colorings of a low-degree graph*, Random Structures Algorithms **7**, no. 2, 157–165.

Jerrum, M. R. and A. J. Sinclair. 1989. *Approximating the permanent*, SIAM Journal on Computing **18**, 1149–1178.

Jerrum, M. and A. Sinclair. 1996. *The Markov chain Monte Carlo method: an approach to approximate counting and integration*, Approximation Algorithms for NP-hard Problems.

Kantorovich, L. V. 1942. *On the translocation of masses*, C. R. (Doklady) Acad. Sci. URSS (N.S.) **37**, 199–201.

Kantorovich, L. V. and G. S. Rubinstein. 1958. *On a space of completely additive functions*, Vestnik Leningrad. Univ. **13**, no. 7, 52–59 (Russian, with English summary).

Karlin, S. and H. M. Taylor. 1975. *A first course in stochastic processes*, 2nd ed., Academic Press, New York.

Kasteleyn, P. W. 1961. *The statistics of dimers on a lattice I. The number of dimer arrangements on a quadratic lattice*, Physica **27**, no. 12, 1209–1225.

Kolata, G. January 9, 1990. *In shuffling cards, 7 is winning number*, New York Times, C1.

Knuth, D. 1997. *The art of computer programming*, third edition, Vol. 2: Seminumerical Algorithms, Addison-Wesley, Reading, Massachusetts.

Kobe, S. 1997. *Ernst Ising—physicist and teacher*, J. Statist. Phys. **88**, no. 3-4, 991–995.

Lawler, G. and A. Sokal. 1988. *Bounds on the L^2 spectrum for Markov chains and Markov processes: a generalization of Cheeger's inequality*, Trans. Amer. Math. Soc. **309**, 557–580.

Letac, G. 1986. *A contraction principle for certain Markov chains and its applications*, Contemporary Mathematics **50**, 263–273.

Li, S.-Y. R. 1980. *A martingale approach to the study of occurrence of sequence patterns in repeated experiments*, Ann. Probab. **8**, no. 6, 1171–1176.

Lindvall, T. 2002. *Lectures on the coupling method*, Dover, Mineola, New York.

Lovász, L. 1993. *Random walks on graphs: a survey*, Combinatorics, Paul Erdős is Eighty, pp. 1–46.

Lovász, L. and P. Winkler. 1993. *On the last new vertex visited by a random walk*, J. Graph Theory **17**, 593–596.

Lovász, L. and P. Winkler. 1995. *Exact mixing in an unknown Markov chain*, Electronic Journal of Combinatorics **2**. paper #R15.

Lovász, L. and P. Winkler. 1998. *Mixing times*, Microsurveys in discrete probability (Princeton, NJ, 1997), DIMACS Ser. Discrete Math. Theoret. Comput. Sci., vol. 41, Amer. Math. Soc., Providence, RI, pp. 85–133.

Loynes, R. M. 1962. *The stability of a queue with non-independent inter-arrival and service times*, Proceedings of the Cambridge Philosophical Society **58**, 497–520.

Luby, M. and E. Vigoda. 1995. *Approximately counting up to four*, Proceedings of the Twenty-Ninth Annual ACM Symposium on Theory of Computing, pp. 150–159. extended abstract.

Luby, M and E Vigoda. 1999. *Fast convergence of the Glauber dynamics for sampling independent sets*, Random Structures and Algorithms **15**, no. 3-4, 229–241.

Lyons, T. 1983. *A simple criterion for transience of a reversible Markov chain*, Ann. Probab. **11**, no. 2, 393–402.

Madras, N. and G. Slade. 1993. *The self-avoiding walk*, Birkhäuser, Boston.

Mann, B. 1994. *How many times should you shuffle a deck of cards?*, UMAP J. **15**, no. 4, 303–332.

Matthews, P. 1988a. *Covering problems for Markov chains*, Ann. Probab. **16**, 1215–1228.

Matthews, P. 1988b. *A strong uniform time for random transpositions*, J. Theoret. Probab. **1**, no. 4, 411–423.

Mihail, M. 1989. *Conductance and convergence of Markov chains - A combinatorial treatment of expanders*, Proceedings of the 30th Annual Conference on Foundations of Computer Science, 1989, pp. 526–531.

Mironov, I. 2002. *(Not so) random shuffles of RC4*, Advances in cryptology—CRYPTO 2002, Lecture Notes in Comput. Sci., vol. 2442, Springer, Berlin, pp. 304–319.

Mossel, E., Y. Peres, and A. Sinclair. 2004. *Shuffling by semi-random transpositions*, Proceedings of the 45th Annual IEEE Symposium on Foundations of Computer Science (FOCS'04) October 17 - 19, 2004, Rome, Italy, pp. 572–581.

Nash-Williams, C. St. J. A. 1959. *Random walks and electric currents in networks*, Proc. Cambridge Philos. Soc. **55**, 181–194.

von Neumann, J. 1951. *Various techniques used in connection with random digits*, National Bureau of Standards Applied Mathematics Series **12**, 36–38.

Peres, Y. 1992. *Iterating von Neumann's procedure for extracting random bits*, Ann. Stat. **20**, no. 1, 590–597.

______. 1999. *Probability on trees: an introductory climb*, Lectures on Probability Theory and Statistics, Ecole d'Ete de Probabilites de Saint-Flour XXVII - 1997, pp. 193–280.

Peres, Y. and D. Revelle. 2004. *Mixing times for random walks on finite lamplighter groups*, Electron. J. Probab. **9**, no. 26, 825–845 (electronic). MR **2110019 (2005m:**60007)

Pinsker, M. S. 1973. *On the complexity of a concentrator*, Proc. 7th Int. Teletraffic Conf., Stockholm, Sweden, pp. 318/1–318/4.

Pisztora, A. 1996. *Surface order large deviations for Ising, Potts and percolation models*, Probab. Theory and Related Fields **104**, no. 4, 427–466.

Propp, J. and D. Wilson. 1996. *Exact sampling with coupled Markov chains and applications to statistical mechanics*, Random Structure and Algorithms **9**, 223–252.

______. 1998. *How to get a perfectly random sample from a generic Markov chain and generate a random spanning tree of a directed graph*, Journal of Algorithms (SODA '96 special issue) **27**, no. 2, 170–217.

Quastel, J. 1992. *Diffusion of color in the simple exclusion process*, Comm. Pure Appl. Math. **45**, no. 6, 623–679.

Randall, D. 2006. *Slow mixing of Glauber dynamics via topological obstructions*, SODA (2006). Available at `http://www.math.gatech.edu/~randall/reprints.html`.

Randall, D. and A. Sinclair. 2000. *Self-testing algorithms for self-avoiding walks*, Journal of Mathematical Physics **41**, no. 3, 1570–1584.

Randall, D. and P. Tetali. 2000. *Analyzing Glauber dynamics by comparison of Markov chains*, J. Math. Phys. **41**, no. 3, 1598–1615. Probabilistic techniques in equilibrium and nonequilibrium statistical physics.

Röllin, A. 2006. *Translated Poisson approximation using exchangeable pair couplings*, available at arxiv:math.PR/0607781.

Saloff-Coste, L. 1997. *Lectures on finite Markov chains*, Lectures on Probability Theory and Statistics, Ecole d'Ete de Probabilites de Saint-Flour XXVI - 1996, pp. 301–413.

Scarabotti, F. and F. Tolli. 2007. *Harmonic analysis of finite lamplighter random walks*, available at arXiv:math.PR/0701603.

Schonmann, R. H. 1987. *Second order large deviation estimates for ferromagnetic systems in the phase coexistence region*, Comm. Math. Phys. **112**, no. 3, 409–422.

Seneta, E. 2006. *Non-negative matrices and Markov chains*, Springer Series in Statistics, Springer, New York. Revised reprint of the second (1981) edition [Springer-Verlag, New York].

Sinclair, A. 1993. *Algorithms for random generation and counting*, Progress in Theoretical Computer Science, Birkhäuser Boston Inc., Boston, MA. A Markov chain approach.

Sinclair, A. and M. Jerrum. 1989. *Approximate counting, uniform generation and rapidly mixing Markov chains*, Inform. and Comput. **82**, no. 1, 93–133.

Spitzer, F. 1976. *Principles of random walks*, 2nd ed., Springer-Verlag, New York. Graduate Texts in Mathematics, Vol. 34.

Stanley, R. P. 1986. *Enumerative combinatorics*, Vol. 1, Wadsworth & Brooks/Cole, Belmont, California.

Thomas, L. E. 1989. *Bound on the mass gap for finite volume stochastic Ising models at low temperature*, Comm. Math. Phys. **126**, no. 1, 1–11.

Thorisson, H. 1988. *Backward limits*, Annals of Probability **16**, 914–924.

Thorp, E. O. 1965. *Elementary Problem E1763*, Amer. Math. Monthly **72**, no. 2, 183.

van Zuylen, A. and F. Schalekamp. 2004. *The Achilles' heel of the GSR shuffle. A note on new age solitaire*, Probab. Engrg. Inform. Sci. **18**, no. 3, 315–328.

Vershik, A. M. 2004. *The Kantorovich metric: the initial history and little-known applications*, Zap. Nauchn. Sem. S.-Peterburg. Otdel. Mat. Inst. Steklov. (POMI) **312**, no. Teor. Predst. Din. Sist. Komb. i Algoritm. Metody. 11, 69–85, 311 (Russian, with English and Russian summaries); English transl.,. 2004, J. Math. Sci. (N. Y.) **133**, no. 4, 1410–1417, available at `arxiv:math.FA/0503035`.

Vigoda, E. 2000. *Improved bounds for sampling colorings*, J. Math. Phys. **41**, no. 3, 1555–1569.

Wilf, H. S. 1989. *The editor's corner: The white screen problem*, Amer. Math. Monthly **96**, 704–707.

Wilson, D. B. 2000. *How to couple from the past using a read-once source of randomness*, Random Structures and Algorithms **16**, 85–113.

Wilson, D. B. 2004. *Mixing times of Lozenge tiling and card shuffling Markov chains*, Ann. Appl. Probab. **14**, no. 1, 274–325.

Zuckerman, D. 1992. *A technique for lower bounding the cover time*, SIAM J. Discrete Math. **5**, 81–87.