Contents

1 Set up

Definition. We define a Markov Chain with state space Ω and transition matrix P to be a sequence of random variables $\{X_i\}_{i\geqslant 0}$ and a matrix P such that for all $x, y \in \Omega$, all $t \geqslant 1$, and all events $H_{t-1} = \bigcap_{s=0}^{t-1}\{X_t = x_s\}$ satisfying $P\{H_{t-1} \cap \{X_t = x_s\}\} > 0$, we have $P\{X_{t+1} = y : H_{t-1} \cap \{X_t = x\}\} = P\{X_{t+1} = y : X_t = x\} = P(x, y)$. This property is also often referred to as the Markov Property.

Definition. A random mapping representation of a transition matrix P on a state space Ω is a function $f : \Omega \times \Lambda \to \Omega$, along with a Λ-valued random variable Z satisfying $P\{f(x, Z) = y\} = P(x, y)$.

Why do we care about random mapping representations? It helps to condense the transition probabilities from a large matrix to a much simpler function. This is discussed in the riffle shuffeling REU paper, found here.

Proposition 1. Every transition matrix on a finite space has a random mapping representation.

Proof. The proofs for this proposition all seem the same. Let $\Omega = \{x_i\}_{i\geqslant 1}$ and take $\Lambda = [0, 1]$; our 'auxiliary' random variables – $Z, Z_1, Z_2, \ldots$ – will be chosen uniformly on this interval. Define

$$f(x_j, z) := x_k \text{ when } \sum_{i=1}^{k-1} P(x_j, x_i) \leqslant z \leqslant \sum_{i=1}^{k} P(x_j, x_i).$$

We have

$$P\{f(x_j, Z) = x_k\} = P\left\{\sum_{i=1}^{k-1} P(x_j, x_i) < Z \leqslant \sum_{i=1}^{k} P(x_j, x_i)\right\} = P(x_j, x_k)$$

per definition of uniform random variables. **Q.E.D**

Claim 1. If $Z_1, Z_2, \ldots$ is a sequence of independent random variables, each having the same distribution as Z, and X_0 has distribution μ, then the sequence $(X_0, X_1, \ldots)$ defined by

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

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

Proof. We refer back to the first definition in this section. We need to thus verify the Markov property. For simplicity, instead of using all of the notation, we'll just get down to the nitty gritty. We want to check

$$P\{X_n = y : X_{n-1} = x, X_{n-2} = x_{n-2}, \ldots, X_0 = x_0\}.$$

We use the proper definition;

$$P\{f(X_{n-1}, Z_n) = y : f(X_{n-2}, Z_{n-1}) = x, f(X_{n-3}, Z_{n-2}) = x_{n-2}, \ldots, X_0 = x_0\}.$$

Since the Z_i are independent, it follows then that this is equivalent to

$$P\{f(X_{n-1}, Z_n) = y : f(X_{n-2}, Z_{n-1}) = x\} = P(x, y).$$

For the reason of independence, see the prior proof and this (note the conditional probability definition). **Q.E.D**

Definition. We say that a chain P is irreducible if, for any $x, y \in \Omega$, there exists an integer t such that $P^t(x, y) > 0$. In other words, it is possible to get from any location to any location in Ω in a finite number of steps.

Definition. Let $T(x) := \{t \geqslant 1 : P^t(x, x) > 0\}$ be the set of times when it is possible for the chain to return to it's starting position. The period of a state x is defined to be $\gcd T(x)$.

Irreducibility and aperiodicity will be important when we set up certain convergence theorems.

Proposition 2. If P is irreducible, then $\gcd T(x) = \gcd T(y)$ for all $x, y \in \Omega$.

Proof. We want to establish that $\gcd(T(x)) = a$ divides $\gcd(T(y)) = b$. Let $r, s > 0$ such that $P^r(x, y) > 0$ and $P^s(y, x) > 0$; this is valid since the chain is irreducible. Let $m := r + s$, and let $z \in T(x)$. We have

$$P^{z+m}(y, y) \geqslant P^s(y, x) \cdot P^z(x, x) \cdot P^r(x, y) > 0.$$

Notice that this gives us that for all $z \in T(x)$, $z + m \in T(y)$. In other words, $b|z + m$. However, $b|m$, so we get for free that $b|z$. Thus, for all $z \in T(x)$, $b|z$. However, a is the $\gcd T(x)$, so this means that $a|b$; in othe words, $a \leqslant b$. By a symmetric argument, we get that $b \leqslant a$, and so $b = a$. **Q.E.D**

Remark. This proof is a variation of the proof found in this REU paper. Mine is phrased much better.

Definition. We say that a Markov chain is aperiodic if all states have period 1. We say that a chain is periodic if it is not aperiodic.

Proposition 3. If P is aperiodic and irreducible, then there is an integer such that $P^r(x, y) > 0$ for all $x, y \in \Omega$.

Proof. I will skip over this proof, since it requires some number theoretic results I don't want to go over. **Q.E.D**

Remark. If a chain is irreducible and has period two (e.g. SRW) on a cycle of even length, then the state space Ω can be partitioned into two classes; these classes are generally denoted by even and odd (a sort of parity property). Let P have period two, and suppose x_0 is an even state. The probability distribution of the chain after $2t$ steps $P^{2t}(x_0, \cdot)$ is supported on even states, while the distribution of the chain after $2t + 1$ steps is supported on odd states. It should be clear then that there is no convergence as we let $t \to \infty$.

We can repair this, though. Given an arbitrary transition matrix P, let $Q = \frac{I+P}{2}$. Since $Q(x, x) > 0$ for all $x \in \Omega$, the transition matrix Q is aperiodic. We call Q the lazy version of P.

Definition. Let π be a distribution on Ω satisfying

$$\pi = \pi P.$$

We call a probability π satisfying this property a stationary distribution of the Markov chain. Clearly, if π is a stationary distribution and $\mu_0 = \pi$, then $\mu_t = \pi$ for all $t \geqslant 0$. Note that we can write this elementwise;

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

This sort of formulation is useful for studying Markov chains on graphs.

Definition. For $x \in \Omega$, we define a hitting time for x to be $\tau_x := \min\{t \geqslant 0 : X_t = x\}$. We also define $\tau_x^+ := \min\{t \geqslant 1 : X_t = x\}$. When $X_0 = x$, we call τ_x^+ to be the first return time.

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

Proof. We first need to find $P_x\{\tau_y^+ > kr\}$. We use the following property of irreducible chains: there exists an integer $r > 0$ and a real $0 < \epsilon < 1$ with the following property: for any states $z, w \in \Omega$, there exists a $j \leqslant r$ with $\epsilon < P^j(z,w) \leqslant (1-\epsilon)$. So, the probability of a hitting state y at a time between t and $t+r$ is at least ϵ. Hence, for $k > 0$, we have

$$P_x\{\tau_y^+ > kr\} \leqslant (1-\epsilon) P_x\{\tau_y^+ > (k-1)r\}.$$

By induction, we get

$$P_x\{\tau_y^+ > kr\} \leqslant (1-\epsilon)^k.$$

We also have

$$\mathbb{E}(Y) = \sum_{t \geqslant 0} P\{Y > t\}.$$

Since $P_x\{\tau_y^+ > t\}$ is a decreasing function of t, we find

$$\mathbb{E}_x(\tau_y^+) = \sum_{t \geqslant 0} P_x\{\tau_y^+ > t\} \leqslant \sum_{k \geqslant 0} r \cdot P_x\{\tau_y^+ > kr\} \leqslant r \cdot \sum_{k \geqslant 0} (1-\epsilon)^k < \infty.$$

Q.E.D

Remark. I'm struggling a lot with the first step of this proof.

Proposition 4. Let P be the transition matrix of an irreducible Markov chain. Then

(i) There exists a probabilty distribution π on Q such that $\pi = \pi P$ and $\pi(x) > 0$ for all $x \in \Omega$, and moreover;

(ii) $\pi(x) = \frac{1}{\mathbb{E}_x(\tau_x^+)}$.

Proof. See the proof in the notes. **Q.E.D**

Definition. A stopping time τ for (X_t) is a $\{0, 1, \ldots\} \cup \{\infty\}$ valued random varaible such that, for each t, the event $\{\tau = t\}$ is determined by $X_0, \ldots, X_t$.

Definition. The strong Markov property is defined by

$$P_{x_0}\{(X_{\tau+1}, X_{\tau+2}, \ldots, X_l) \in A : \tau = k \text{ and } (X_1, \ldots, X_k)$$
$$= (x_1, \ldots, x_k)\} = P_{x_k}\{(X_1, \ldots, X_l) \in A\}$$

for any $A \subset \Omega^l$ and τ a stopping time.

Definition. Suppose that $\{X_i\}$ is an irreducible and positive recurrent chain, which is started at it's unique invariant distribution π. Recall that this means that π is the p.m.f. Now suppose that for every n, $X_0, X_1, \ldots, X_n$ have the same joint p.m.f as their time-reversal $X_n, X_{n-1}, \ldots, X_0$. Then we call the chain reversible – sometimes it is, equivalently, also said that it's invariant distribution π is reversible. A good heuristic is that the recorded simulation of a reversible chain looks the same if the 'movie' is run backwards.

Theorem. A Markov chain with invariant measure π is reversible if and only if

$$\pi_i P_{ij} = \pi_j P_{ji}$$

for all states i, j.

This leads to an interesting result:

Proposition 5. Reversibility implies invariance; in other words, if the probability mass function π_i satisfies the condition in the previous theorem, then it is invariant.

Remark. The above definition and condition were retrieved from here.

Here is an examples of reversibility in action.

Example 1. We will explore the random walk on weighted graphs. Assume that every undirected edge between vertices i and j in a complete graph has a weight $w_{ij} = w_{ji}$; we think of edges with 0 weight as not present at all. When in i, the walker goes to j with probability proportional to w_{ij} so that

$$P_{ij} = \frac{w_{ij}}{\sum_k w_{ik}}.$$

Let

$$s = \sum_i \sum_k w_{ik}$$

and let

$$\pi_i = \frac{\sum_k w_{ik}}{s}.$$

Then we see

$$\pi_i P_{ij} = \frac{\sum_k w_{ik}}{s} \frac{w_{ij}}{\sum_k w_{ik}} = \frac{w_{ij}}{s}$$

$$= \frac{w_{ji}}{s} = \frac{\sum_k w_{jk}}{s} \frac{w_{ji}}{\sum_k w_{kj}} = \pi_j P_{ji}$$

implying reversibility.

Definition. The detailed balance equations are defined as follows: Let π be a probability on Ω. Then we say π satisfies the detailed balance equations if

$$\pi(x)P(x,y) = \pi(y)P(y,x)$$

for all $x, y \in \Omega$. In other words, if it is reversible.

Proposition 6. Let P be the transition matrix of a Markov chain with state space Ω. Any distribution π satisfying the detailed balance equations is stationary for P.

Proof. Assume π satisfies

$$\pi(x)P(x,y) = \pi(y)P(y,x)$$

for all $x, y \in \Omega$. Then we get

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

since P is stochastic. **Q.E.D**

Why is reversibility/detailed balance equations important? It's often the easiest way to find the stationary distribution.

Proposition 7. Let (X_t) be an irreducible Markov chain with transition matrix P and stationary distribution π. Write $(\bar{X}_t)$ for the time-reversed chain with transition matrix $\bar{P}$. Then π is stationary for $\bar{P}$, and for any $x_0, \ldots, x_t \in \Omega$ we have

$$P_\pi\{X_0 = x_0, \ldots, X_t = x_t\} = P_\pi\{\bar{X}_0 = x_t, \ldots, \bar{X}_t = x_0\}.$$

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

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

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

$$P_\pi\{X_0 = x_0, \ldots, X_n = x_n\} = \pi(x_0)P(x_0,x_1)P(x_1,x_2) \cdot P(x_{n-1},x_n)$$

$$= \pi(x_n)\bar{P}(x_n,x_{n-1}) \cdots \bar{P}(x_2,x_1)\bar{P}(x_1,x_0)$$

$$= P_\pi\{\bar{X}_0 = x_n, \ldots, \bar{X}_n = x_0\},$$

since $P(x_{i-1},x_i) = \pi(x_i)\bar{P}(x_i,x_{i-1})/\pi(x_{i-1})$ for each i. **Q.E.D**

Definition. Given $x, y \in \Omega$, we say that y is accessible from x and write $x \to y$ if there exxists an $r > 0$ such that $P^r(x, y) > 0$. That is, $x \to y$ if it is possible for the chain to move from x to y in a finite number of steps.

We first need to discuss the Chapman-Kolmogorov equation.

Proposition 8. (Chapman-Kolmogorov equation) We have

$$P_{ij}^{n+m} = \sum_{l \in \Omega} P_{il}^n P_{lj}^m.$$

Proof. Proof omitted. **Q.E.D**

Claim 2. Accessibility is transitive; that is to say, if $x \to y$, $y \to z$, then $x \to z$.

Proof. We use the Kolmogorov-Chapman equation. If $x \to y$, then we have, for some $r_1 > 0$, then $P^{r_1}(x, y) > 0$. Since $y \to z$, we have for some $r_2 > 0$ that $P^{r_2}(y, z) > 0$. Then $r = r_1 + r_2$, and so $P^r(x, z) = \sum_{l \in \Omega} P^{r_1}(x, l) P^{r_2}(l, z) \geqslant P^{r_1}(x, y) P^{r_2}(y, z) > 0$. **Q.E.D**

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

Definition. 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 communicating classes. For $x \in \Omega$, the communicating class of x is denoted by $[x]$.

Claim 3. $\leftrightarrow$ is an equivalence class.

Proof. Recall that the definition of equivalence class requires three things: reflexivity, symmetry, and transitivity. We go through each. For reflexivity, we have clearly that $x \leftrightarrow x$. For symmetry, it's clear that $x \leftrightarrow y$ implies $y \leftrightarrow x$. The tricky one is transitivity, but by the prior claim it's clear that if $x \leftrightarrow y$, $y \leftrightarrow z$, then $x \leftrightarrow z$. **Q.E.D**

This leads us to the following proposition.

Proposition 9. If x is an essential state, and $x \to y$, then y is essential.

Proof. If $y \to z$, then $x \to z$. Because x is essnetial, $z \to x$, and so using transitivity $z \to y$. Hence, $y \leftrightarrow z$. **Q.E.D**

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

Remark. If $[x] = \{x\}$ and x is inessential, then we see that once the chain leaves x it never returns. Likewise, if $[x] = \{x\}$ and x is essential, we see that the chain never leaves x once it first visits x.

Definition. If $[x] = \{x\}$ and x is essential, then x is absorbing.

Proposition 10. Every finite chain has at least one essential class.

Proof. Proof omitted for now. **Q.E.D**

Proposition 11. If π is stationary for the finite transition matrix P, then $\pi(y_0) = 0$ for all inessential states y_0.

Proof. Let C be an essential communicating class. Then

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

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

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

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

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

Since π is invariant, $\pi P(C) = \pi(C)$. In view of the prior equation, we must have $\pi(y)P(y,z) = 0$ for all $y \notin C$ and $z \in C$.

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

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

This implies in particular that $\pi(y)P(y, y_k) = 0$ for all y, and $\pi(y_{k-1}) = 0$. By induction, we find that $\pi(y_0) = 0$. **Q.E.D**

Proposition 12. The stationary distribution π for a transition matrix P is unique if and only if there is a unique essential communicating class.

Proof. Proof omitted. **Q.E.D**

Exercises

Exercise 1. Let G be a connected graph. Show that a random walk on G is irreducible if and only if G is connected.

Proof. We prove the forward direction. Since G is irreducible, we have for some $r > 0$ that $P^r(x,y) > 0$ for all $x, y \in G$. However, this means that we can construct a series of edges such that $x \to y$. Since this applies for all $x, y \in G$, then we have that G is connected. We prove the converse direction. Since G is connected, we have that there is a sequence of edges such that $x \to y$ for all $x, y \in G$. Say that there are r edges on this path. Then we have that $P^r(x,y) > 0$ clearly. Hence, the result follows. **Q.E.D**

Exercise 2. Let P be an irreducible matrix of period b. Show that Ω can be partitioned into b sets $C_1, C_2, \ldots, C_b$ in such a way that $P(x, y) > 0$ only if $x \in C_i$ and $y \in C_{i+1}$.

Proof. We partition our graph based on the fact that the period is b. So any element which is b away from our current element, when arranged in a cyclic graph, is placed in the same class C_1. Go to the next element and repeat. We then get the corresponding classes we need. Now, notice that since this is a digraph (otherwise period properties are broken), we get that if $x \in C_i$ and $y \notin C_{i+1}$, then we cannot have $P(x, y) > 0$; otherwise, we get that the period will no longer be b. **Q.E.D**

2 Examples of Markov Chains

2.1 Gambler's ruin

The way the gambler's ruin works is simple; say we have some coin with probability p for heads and $1-p$ for tails (not necessarily fair). If the coin lands heads, the person gets a dollar, and if it lands tails they lose a dollar. If the person reaches a dollar amount, say n, then they will stop playing the game. If they have no money, they must stop playing the game.

In essence, this is just a simple random walk on the integers modulo $n+1$ with some boundary conditions (you're stuck once you hit 0 and once you hit n).

Definition. We say that 0 and n in the prior example are absorbing states.

Claim 4. The above set up gives us a Markov chain.

Proof. The sketch of the proof is that the chain does not rely on any further information beyond what just happened. Prior information does not influence the future information. **Q.E.D**

This leads us to the following proposition. In this proposition, we assume $p = \frac{1}{2}$.

Proposition 13. Assume that a gambler making fair unit bets on coin flips will abandon the game when they reach the absorbing states. Let X_t be the gambler's fortune at time t, and let τ be the time required to be absorbed at one of 0 or n. Assume that $X_0 = k$ (here, k denotes their starting dollar amount), where $0 \leqslant k \leqslant n$. Then

$$P_k\{X_\tau = n\} = k/n$$

and

$$\mathbb{E}_k(\tau) = k(n-k).$$

Proof. We set up a system of equations. Let $\{p_i\}_{0\leqslant i\leqslant n}$ be the probabilities such that the gambler reaches a fortune of n before reaching a fortune of 0 when starting at i. Then clearly $p_0 = 0$ and $p_n = 1$. For times inbetween those, we have that (since there is a $\frac{1}{2}$ chance of going either direction)

$$p_k = \frac{1}{2}p_{k-1} + \frac{1}{2}p_{k+1}.$$

We'll try and see if there's a pattern, now. We find

$$p_1 = \frac{1}{2}p_2.$$

Likewise,

$$p_2 = \frac{1}{2}p_3 + \frac{1}{2}p_1 = \frac{1}{2}p_3 + \frac{1}{4}p_2.$$

This is equivalent to

$$\frac{3}{4}p_2 = \frac{1}{2}p_3$$

or

$$p_2 = \frac{2}{3}p_3.$$

Again, we have

$$p_3 = \frac{1}{2}p_2 + \frac{1}{2}p_4.$$

Substituting this in gives us

$$p_3 = \frac{1}{3}p_3 + \frac{1}{2}p_4 \leftrightarrow \frac{2}{3}p_3 = \frac{1}{2}p_4 \leftrightarrow p_3 = \frac{3}{4}p_4.$$

If you notice the pattern, we have that $p_k = \frac{k}{k+1}p_{k+1}$. It is a simple induction argument to show that this holds. Moreover, since we have that $p_n = 1$, we get

$$p_{n-1} = \frac{n-1}{n}.$$

Moving down the line, we get

$$p_{n-2} = \left(\frac{n-2}{n-1}\right)\left(\frac{n-1}{n}\right) = \frac{n-2}{n}.$$

Again, by another induction argument, we find that we have

$$p_k = \frac{k}{n}$$

for all $0 \leqslant k \leqslant n$ integer. This establishes the first claim.

For the second claim, we let f_k denote the expected time to be absorbed (this is either at 0 or n). It is self-evident that $f_0 = f_n = 0$, since the walk doesn't have to move in either direction to get absorbed (it is already absorbed). For all states inbetween, we get

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

The reasoning is outlined in the book. We now need to solve our system. For f_1, we have

$$f_1 = \frac{1}{2}(1 + f_2) + \frac{1}{2} = 1 + \frac{f_2}{2}.$$

Moving down the line, we get

$$f_2 = \frac{1}{2}(1 + f_1) + \frac{1}{2}(1 + f_3) = 1 + \frac{f_2}{4} + \frac{1}{2}(1 + f_3).$$

In other words,

$$f_2 = 2 + \frac{2}{3}f_3.$$

Continuing down the line, we find

$$f_k = k + \frac{k}{k+1} f_{k+1}$$

by a simple induction argument. At $n-1$, we recall $f_n = 0$, and so we have

$$f_{n-1} = n-1$$

Moving down the line, we find

$$f_{n-2} = n-2 + \frac{n-2}{n-1} \cdot n-1 = n-2+n-2 = 2(n-2) = (n-2)\cdot(n-(n-2)).$$

For fun, we have

$$f_{n-3} = n-3 + \frac{n-3}{n-2} \cdot 2(n-2) = n-3+2(n-3) = 3(n-3) = (n-(n-3))\cdot(n-3).$$

By an induction argument again, we find

$$f_k = k(n-k)$$

which is the result we desired. **Q.E.D**

Question 1. Can we do this with $p \neq \frac{1}{2}$ and still get a nice result?

2.2 Coupon Collecting

A company decides to issue n different type of coupons, and some collector desires to have each type of coupon. We suppose that the probability of acquiring each coupon is equally likely among the n types. How many coupons must they collect in order to get the n types?

Let X_t denote the number of different types represented among the collector's first t coupons. We clearly have that $X_0 = 0$. When we've reached k different types, we're missing $n-k$ types of coupons, and so we have

$$P\{X_{t+1} = k+1 : X_t = k\} = \frac{n-k}{n}.$$

Likewise,

$$P\{X_{t+1} = k : X_t = k\} = 1 - P\{X_{t+1} = k+1 : X_t = k\} = \frac{k}{n}.$$

Every trajectory is non-decreasing in this chain. The states n is an absorbing state.

Claim 5. The set up above is a Markov chain.

Proof. Again, this is rather a pseudo-proof than a real proof. It is really self-evident that the information about the next step only relies on the current step, and not on any prior information (it does not matter the way in which we reach the point). Hence, it is a Markov chain. **Q.E.D**

This leads us to the following proposition.

Proposition 14. Consider a collector attempting to collect a complete set of coupons. Assume that each new coupon is chosen uniformly and independently from the set of n possible types, and let τ be the (random) number of coupons collected when the set first contains every type (when we've completed our run). Then

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

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

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

(The explanation for the difference; we want to figure out how long it took from τ_i to τ_{i+1}, since we've already counted τ_i. To do so, we take the difference between the two.) Furthermore, $\tau_k - \tau_{k-1}$ is an easy random variable to compute; it is a geometric random variable with success probability

$$\frac{n-k+1}{n};$$

after collecting τ_{k-1} coupons, there are $n-k+1$ types missing from the collection. Each subsequent coupons drawn has the same probability of being a type not already collected, until a new type is finally drawn. Hence, we get (by the linearity of expectation)

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

Q.E.D

We can further improve these bounds, but this is omitted for the time being.

2.3 The Hypercube and the Ehrenfest Urn Model

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

Example 2. One quick example would be to examine the 3-dimensional hypercube. We have $\{0,1,1\}$ and $\{0,1,0\}$ to be differing only by one coordinate – the last – and so they would be connected by an edge. There is a visual of this in the book.

The simple random walk on the hypercube moves from a vertex $(x_1, \ldots, x_n)$ by choosing some coordinate $j \in \{1, 2, \ldots, n\}$ uniformly at random and setting the new state equal to $(x_1, \ldots, x_{j-1}, 1 - x_j, x_{j+1}, \ldots, x_n)$. That is, the bit at the walk's chosen coordinate is flipped.

Unfortunately, however, this is is periodic. To resolve this issue, we introduce the lazy random walk (see pg. 3 for more information). The lazy random walk has a probability of 0.5 of remaining at the same location. You can think of this as selecting a coordinate uniformly at random and refreshing it.

We now consider the Ehrenfest Model. Suppose n balls are distributed among two urns, denoted by A and B. At each move, a ball is selected uniformly at random and transferred from its current urn to the other urn. If X_t is the number of balls in urn A at time t, then the transition matrix for (X_t) is

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

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

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

Definition. Define the Hamming weight $W(x)$ of a vector $x := (x_1, \ldots, x_n) \in \{0,1\}^n$ to be its number of coordinates with value 1:

$$W(x) = \sum_{j=1}^{n} x_j.$$

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

This leads us to the concept of projections of chains. The Ehrenfest urn is a projection, which we define in this section, of hte simple random walk on the hypercube.

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

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

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

$$P(x, [y]) = P(x', [y])$$

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

Definition. The process of constructing a new chain by taking equivalence classes for an equivalence relation compatible with the transition matrix is called projection, or sometimes lumping.

As a final remark, we notice that the Ehrenfest urn is reversible.

Claim 6. The Ehrenfest urn is reversible.

Proof. The invariant measure puts each ball at random into one of the two urns, as switching any ball between the two urns does not alter this assignment. Thus $\pi \sim \text{Bin}(n, \frac{1}{2})$ (a more formal proof will be explored later). In other words,

$$\pi_i = \binom{n}{i} \frac{1}{2^n}.$$

Checking for both $j = i \pm 1$ on Maple, we see the calculations come out as desired (they are equal). Hence, the chain is reversible. **Q.E.D**

2.4 The Polya Urn Model

The Polya urn is an urn containing two balls, one black and one white. From this point on, we choose a ball at random, take the ball out of the urn, and then return the ball along with another of the same color. We can force this into a Markov chain in the following way; if there are j black balls in the urn after k balls have been added (so that there are $k+2$ balls total in the urn), then the probability that another black ball is added is $\frac{j}{k+2}$. The sequence of ordered pairs listing the number of black and white balls is a Markov chain with state space $\{1, 2, \ldots\}^2$.

Lemma. Let B_k be the number of black balls in Polya's urn after the addition of k balls. The distribution of B_k is uniform on $\{1, 2, \ldots, k+1\}$.

Proof. There are many subparts to this proof. Let $\{U_i\}_{i \leqslant n}$ be i.i.d. random variables, each uniform on the interval $[0, 1]$. Let

$$L_k := |\{j \in \{0, 1, \ldots, k\} : U_j \leqslant U_0\}|.$$

The event $\{L_k = j : L_{k+1} = j + 1\}$ occurs if and only if U_0 is the $(j + 1$-st smallest and U_{k+1} is the smallest among $\{U_0, U_1, \ldots, U_{k+1}\}$.

Claim 7. There are $j(k!)$ orderings of $\{U_0, \ldots, U_{k+1}\}$ given $\{L_k = j, L_{k+1} = j + 1\}$.

Proof. First, we examine the first $U_0, \ldots, U_k$ (ignore U_{k+1} for now). Place U_0 in the $j + 1$ place, and then shuffle the remaining into the rest of the spaces; this gives us $k!$ ways of arranging the U_i. Now, we need to place U_{k+1} in any of the first j places, this gives us the j. Hence, we have $j(k!)$ ways of arranging this. **Q.E.D**

Since there are $j(k!)$ orderings of $\{U_0, \ldots, U_{k+1}\}$ making up this event, and since all $(k + 2)!$ orderings are equally likely, we use some basic discrete probability do find

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

This leads us to our next claim.

Claim 8. We have $P\{L_k = j\} = \frac{1}{k+1}$.

Proof. Going back to our first claim, we notice again that there are $k!$ ways of arranging things such that we get $L_k = j$. We also note that there are $(k + 1)!$ ways of arranging things without taking into consideration $L_k = j$. Therefore, using basic discrete probability, we again get

$$P\{L_k = j\} = \frac{k!}{(k + 1)!} = \frac{1}{k + 1}.$$

Q.E.D

Claim 9. Combining Claim 4 and Claim 5, we find

$$P\{L_{k+1} = j + 1 | L_k = j\} = \frac{j}{k + 2}.$$

Claim 10. Using Claim 6, we have

$$P\{L_{k+1} = j | L_k = j\} = \frac{k + 2 - j}{k + 2}.$$

Proof. Now, notice that we have that, given $L_k = j$, we must have $L_{k+1} = j + 1$ or $L_{k+1} = j$. Recall that $P\{L_{k+1} = \cdot | L_k = j\}$ forms a probability measure, and so using the prior fact we find

$$P\{L_{k+1} = j | L_k = j\} = 1 - P\{L_{k+1} = j + 1 | L_k = j\}.$$

This gives the above formula. **Q.E.D**

Claim 11. We have that $\{L_i\}_{i=1}^n$ and $\{B_i\}_{i=1}^n$ share the same distribution and transition probabilities. In particular, B_k and L_k have the same distribution.

Proof. Claim 7 gives us the latter part of this claim. For the former, we need to show that L_1 and B_1 have the same distribution. But this is clear. B_1 is the number of black balls after the addition of one ball, which is $P\{B_1 = 1\} = 1/3$, $P\{B_1 = 2\} = 2/3$, and likewise $P\{L_1 = 1\} = 1/3$, $P\{L_2 = 2\} = 2/3$. Combining the two facts gives us the final part of the claim. **Q.E.D**

Since the position of U_0 is uniform among the $k + 1$ possible positions, it follows that L_k is uniform on $\{1, \ldots, L_{k+1}\}$. Thus, we have the B_k is uniform on $\{1, \ldots, k + 1\}$ as desired. **Q.E.D**

Remark. The book claims that the prior lemma can be proven via showing $P\{B_k = j\} = 1/(k + 1)$ for all $j = 1, \ldots, k + 1$ using induction on k. Maybe as an exercise prove it this way.

2.5 Birth-and-Death Chains

Definition. A birth-and-death chain has state space $\Omega = \{0, 1, 2, \ldots n\}$. In one step, the state can either increase or decrease by at most 1. The current state can be thought of as the size of some population; in a single step of the chain, there can be at most one birth or death. The transition probabilities can be specified by p_k, r_k, and q_k where $k \in [0, \ldots, n]$ and $p_k + r_k + q_k = 1$, where p_k is the probability of moving from k to $k = 1$, q_k is the probability of moving from k to $k - 1$, and r_k is the probability of remaining at k. We also have that $q_0 = p_n = 0$.

Proposition 15. Every birth-and-death chain is reversible.

Proof. We have that a function w on Ω satisfies the detailed balance equations if and only if

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

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

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

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

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

for $0 \leqslant k \leqslant n$. **Q.E.D**

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

- For any $k \in \{0, 1, \dots, l-1\}$, the chain makes the 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 $r_l + p_l$, respectively.

We write $\bar{\mathbb{E}}$ for expectations for this new chain. By the proof of reversibility of this chain, the stationary probability $\bar{\pi}$ of the truncated chain is given by

$$\bar{\pi} = \frac{w_k}{\sum_{j=0}^{l} w_j}$$

for $0 \leqslant k \leqslant l$. Since in the truncated chain the only possible moves from l are to stay put or step down to $l-1$, the expected first return time $\bar{\mathbb{E}}_l(\tau_l^+)$ satisfies

$$\bar{\mathbb{E}}_l(\tau_l^+) = (r_l + p_l) \cdot 1 + q_l \left(\bar{\mathbb{E}}_{l-1}(\tau_l) + 1 \right) = 1 + q_l \bar{\mathbb{E}}_{l-1}(\tau_l).$$

By Proposition 4 part 2,

$$\bar{\mathbb{E}}_l(\tau_l^+) = \frac{1}{\bar{\pi}(l)} = \frac{1}{w_l} \sum_{j=0}^{l} w_j.$$

We have constructed the truncated chain so that $\bar{\mathbb{E}}_{l-1}(\tau_l) = \mathbb{E}_{l-1}(\tau_l)$. Rearranging the above equations gives

$$\mathbb{E}_{l-1}(\tau_l) = \frac{1}{q_l} \left(\sum_{j=0}^{l} \frac{w_j}{w_l} - 1 \right) = \frac{1}{q_l w_l} \sum_{j=0}^{i-1} w_j.$$

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

$$\mathbb{E}_a(\tau_b) = \sum_{l=a+1}^{b} \mathbb{E}_{l-1}(\tau_l).$$

There were some special cases the book considered, but I skipped over it.

2.6 Random Walks on Groups

Given a probability distribution μ on a group $(G, \cdot)$, we define the random walk 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 of a random element of G selected according to μ. Equivalently, the transition matrix P of this chain has entries

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

for all $g, h \in G$.

Remark. We multiply the current state by the increment on the left because it is generally more natural in non-commutative examples, such as the symmetric group. For commutative examples, such as the two described below, it does not matter which side we multiply it by.

Example 3. (The n-cycle) Let μ assign probability 1/2 to each of 1 and $n-1 \equiv -1 \pmod{n}$ in the additive cyclic group $\mathbb{Z}_n = \{0, 1, \ldots, n-1\}$. The simple random walk on the n-cycle, discussed in the set-up section, is the random walk on $\mathbb{Z}_n$, with increment distribution μ. Similarly, let v assign weight 1/4 to both 1 and $n-1$ and weight 1/2 to 0. Then lazy random walk on the n-cycle, discussed prior as well, is the random walk on $\mathbb{Z}_n$ with increment distribution v.

Example 4. (The hypercube) The hypercube random walks defined earlier are random walks on the group $\mathbb{Z}_2^n$, which is the direct product of n copies of the two element group $\mathbb{Z}_2$. For the simple random walk the increment distribution is uniform on the set $\{e_i : 1 \leqslant i \leqslant n\}$, where the vector e_i has a 1 in the i-th place and a 0 in all other entires. For the lazy version, the increment distribution gives the vector 0 (with all zero entries) weight 1/2 and each e_i weight $1/2n$.

Proposition 16. Let P be the transition matrix of a random walk on a finite group G and let U be the uniform probability distribution on G. Then U is a stationary distribution for P.

Proof. Let μ be the increment distribution of the random walk. For any $g \in G$

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

The first equality comes from re-indexing $k = gh^{-1}$. **Q.E.D**

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 $\langle H \rangle H$ 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 17. Let μ be a probability distribution on a finite group G. The random walk on G with increment distribution μ is irreducible if and only if $S = \{g \in G : \mu(g) > 0\}$ generates G.

Proof. Select an arbitrary $a \in G$. If the random walk is irreducible, then there exists an $r > 0$ so that $P^r(e, a) > 0$, where $e \in G$ is the identity element. In order for this to occur, there must be some sequence $s_1, \ldots, s_r \in G$ such that $a = s_r s_{r-1} \cdots s_1$ and $s_i \in S$ for $i = 1, \ldots, r$. Thus, $a \in \langle S \rangle$.

Now assume that S generates G, and consider $a, b \in G$. We know that ba^{-1} can be written as a word in the elements of S and their inverses. Since every element of G has finite order, any inverse appearing in the expression for ba^{-1} can be written as a positive power of the same group element. Let the resulting expression be $ba^{-1} = s_r s_{r-1} \cdots s_1$ where $s_i \in S$, for $i = 1, \ldots, r$. Then

$$P^m(a,b) \geqslant P(a, s_1 a)P(s_1 a, s_2 s_1 a) \cdots P(s_{r-1}s_{r-2} \cdots s_1 a, (ba^{-1})a)$$

$$= \mu(s_1)\mu(s_2)\cdots\mu(s_r) > 0.$$

Q.E.D

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

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

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

Proposition 18. The random walk on a finite group G with increment distribution μ is reversible if μ is symmetric.

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

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

and

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

which are equal if and only if $\mu(hg^{-1} = \mu((hg^{-1})^{-1})$ **Q.E.D**

Remark. The converse of the prior proposition is also true. It is an exercise to do this.

Definition. A Markov chain is called transitive if for each pair $(x, y) \in \Omega \times \Omega$ there is a bijection $\phi = \phi_{(x,y)} : \Omega \to \Omega$ such that

$$\phi(x) = y$$

and

$$P(z, w) = P(\phi(z), \phi(w))$$

for all $z, w \in \Omega$. Roughly, this means that the chain 'looks the same' from any point in the state space Ω. Clearly, any random walk on a group is transitive; set $\phi_{(x,y)}(g) = gx^{-1}y$. However, there are examples of transitive chains that are not random walks on groups.

Many properties of random walks on groups generalize to the transitive case, including Proposition 10.

Proposition 19. Let P be the transition matrix of a transitive Markov chain on a finite state space Ω. Then the uniform probability distribution on Ω is stationary for P.

Proof. Fix $x, y \in \Omega$ and let $\phi(x) = y$. Let U be the uniform probability on Ω. Then

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

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

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

Q.E.D

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

A nearest-neighbor random walk on $\mathbb{Z}$ moves right and left by at most one step on each move, and each move is independent of the past. More precisely, if (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).

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

$$P(k, k+1) = 0, P(k,k) = r, P(k, k-1) = q,$$

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

$$P_0\{X_t = k\} = \left\{ \begin{array}{ll} \binom{t}{\frac{t-k}{2}} 2^{-t}, & \text{if } t-k \text{ is even,} \\ 0, & \text{otherwise} \end{array} \right\}$$

since there are $\binom{t}{\frac{t-k}{2}}$ possible paths of length t from 0 to k. When $p = q = 1/4$ and $r = 1/2$, the chain is the lazy simple random walk on $\mathbb{Z}$.

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

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

is the first time the walk hits zero. Then

$$P_k\{\tau_0 > r\} \leqslant \frac{12k}{\sqrt{r}}$$

for any integers $k, r > 0$.

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

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

$$P_k\{\tau_0 < r, X_r = j\} = P_k\{X - r = -j\}$$

and

$$P_k\{\tau_0 < r, X_r > 0\} = P_k\{X_r < 0\}.$$

Proof. We proceed using the Markov property. The walk starts anew from 0 when it hits 0, meaning that the walk viewed from the first time it hits zero is independent of its past and has the same distribution as a walk which started at 0. Hence, for any $s < r$ and $j > 0$, we have

$$P_k\{\tau_0 = s, X_r = j\} = P_k\{\tau_0 = s\}P_0\{X_{r-s} = j\}.$$

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

$$P_k\{\tau_0 = s\}P_0\{X_{r-s} = j\} = P_k\{\tau_0 = s, X_r = -j\}.$$

Summing over $s < r$, we obtain

$$P_k\{\tau_0 < r, X_r = j\} = P_k\{\tau_0 < r, X_r = -j\} = P_k\{X_r = -j\}.$$

Summing over $j > 0$ yields our result. **Q.E.D**

Remark. A simpler combinatorial interpretation is that 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. To obtain the bijection, reflect a path after the first time it hits 0.

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

$$P_k\{\tau_0 > r\} = P_0\{-k < X_r \leqslant k\}$$

for any $k > 0$.

Proof. We have that

$$P_k\{X_r > 0\} = P_k\{X_r > 0, \tau_0 \leqslant r\} + P_k\{\tau_0 > r\}.$$

By the Reflection Principle,

$$P_k\{X_r > 0\} = P_k\{X_r < 0\} + P_k\{\tau_0 > r\}.$$

By the symmetry of the walk, $P_k\{X_r < 0\} = P_k\{X_r > 2k\}$, and so combining this gives the desired result. **Q.E.D**

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

$$P_0\{X_t = k\} \leqslant \frac{3}{\sqrt{t}}.$$

Proof. If $X_{2r} = 2k$, there are $r + k$ up moves and $r - k$ down moves. The probability of this is

$$\binom{2r}{r+k} 2^{-2r}.$$

We check that $\binom{2r}{r+k}$ is maximized at $k = 0$ for $k = 0, 1, \dots, r$. In other words, after simplifying, we would like to show

$$\frac{1}{(r-k)!(r+k)!} < \frac{1}{r!^2}.$$

We need a few claims to proceed (both are clear from the definition of factorial).

Claim 12.

$$\frac{r!}{(r-k)!} = (r - (k-1)) \cdot (r - (k-2)) \cdots r$$

Claim 13.

$$\frac{r!}{(r+k)!} = \frac{1}{(r+1) \cdot (r+2) \cdots (r+k)}$$

It is clear that, after multiplying this together, we get the product is less than 1. In other words, we have that it is maximized at $k = 0$. So, using this fact, we have

$$P_0\{X_{2r} = 2k\} \leqslant \binom{2r}{r} 2^{-2r} = \frac{(2r)!}{(r!)^2 2^{2r}}.$$

Using Stirling's formula, we obtain

$$P_0\{X_{2r} = 2k\} \leqslant \sqrt{\frac{8}{\pi}} \frac{1}{\sqrt{2r}}.$$

We now condition on the first step of the walk and use the bound found above. Use as well the simple bound

$$\sqrt{\frac{t}{t-1}} \leqslant \sqrt{2}$$

to see

$$P_0\{X_{2r+1} = 2k+1\} \leqslant \frac{4}{\sqrt{\pi}} \frac{1}{\sqrt{2r+1}}.$$

Note that $\frac{4}{\sqrt{\pi}} \leqslant 3$, and we get the bound. **Q.E.D**

Remark. The bijection described earlier has a very nice consequence. Define an up-right path to be a path through the two-dimensional grid in which every segment heads either up or to the right.

Theorem (The Ballot Theorem)**.** Fix positive integers a and b with $a < b$. An up-right path from $(0, 0)$ to (a, b) chosen uniformly at random has probability $\frac{b-a}{a+b}$ of lying strictly above the line $x = y$ (except for its initial point).

Remark. There is a very nice interpretation of this in terms of votes, which I'll leave to the book.

Proof. The total number of up-right paths from $(0,0)$ to (a,b) is $\binom{a+b}{b}$, since there are exactly $a+b$ steps total, of which exactly b steps go right.

How many paths never touch the line $x=y$ after the first step? Any such path must have its first step up, and there are $\binom{a+b-1}{b-1}$ such paths. How many of those paths touch the line $x=y$?

Given a path whose first step is up and that touches the line $x=y$, reflecting the portion after the first touch of $x=y$ yields a path from $(0,0)$ whose first step is up and which ends at (b,a). Since every up-right path whose first step is up and which ends at (b,a) must cross $x=y$, we obtain every such path via this reflection. Hence, there are $\binom{a+b-1}{b}$ 'bad' paths to subtract, and the desired probability is

$$\frac{\binom{a+b-1}{b-1}-\binom{a+b-1}{b}}{\binom{a+b}{b}}=\frac{b-a}{a+b}.$$

Q.E.D

Exercises

Problem 1. Show that the system of equations for $0<k<n$

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

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

Solution. This was done above.

Problem 2. Consider a hesitant gambler: at each time, they flip 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 tops playing. Assuming that her initial fortune is k, find the expected number of rounds she will play, in terms of n, k, and p.

Solution. This is almost analogous to what we did before, except now we need to modify a few factors. Again, we write f_k for the expected time $\mathbb{E}_k(\tau)$ to be absorbed, starting at position k. Clearly, $f_0=f_n=0$. In order to move up, we need to get 2 heads; so there is a $\frac{p}{2}$ chance of moving up. To move down, we need to get a heads and a tails, so there is a $\frac{p}{2}$ chance of moving down. To remain, we need to get a tails on the first flip, so there is a $(1-p)$ chance of this happening. Our system is now

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

We start with f_1 to try to find some sort of inductive argument. Solving, we find

$$f_1=\frac{1}{2p}(pf_2+2).$$

Likewise,

$$f_2 = \frac{2}{3p}(pf_3 + 3)$$

which leads us to our claim.

Claim 14. We have

$$f_k = \frac{k}{p(k+1)}\big(pf_{k+1} + (k+1)\big).$$

Proof. We've shown the base case above. Assume it holds for k. We must show it holds for $k+1$. Using the system above, we have

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

Simplifying this on Maple gives us the desired equality. **Q.E.D**

Now, we find f_{n-1}; this gives us

$$f_{n-1} = \frac{n-1}{p}.$$

Moving backwards, we see

$$f_{n-2} = \frac{2(n-2)}{p}$$

and

$$f_{n-3} = \frac{3(n-3)}{p}$$

leading us to our next claim.

Claim 15. We have

$$f_k = \frac{k(n-k)}{p}.$$

Proof. Again, we use induction. We have done the base cases above. Assume it holds for k. We must show it holds for $k-1$. We have

$$f_{k-1} = \frac{k-1}{pk}(pf_k + k).$$

Using the inductive hypothesis and Maple, we have

$$\frac{\underline{(k-1)(n-(k-1))}}{p}$$

which is what we desired. **Q.E.D**

Putting things together, we get

$$\mathbb{E}_k(\tau) = \frac{k(n-k)}{p}.$$

Problem 3. Consider a random walk on the path $\{0, 1, \ldots, n\}$ in which the walk moves left or right with equal probability except when at n and 0. At n, it remains at n with probability 1/2 and moves to $n-1$ with probability 1/2, and once the walk hits 0 it remains there forever. Compute the expected time of the walks's absorption at state 0, given that it starts at state n.

Solution. We proceed like prior. Set $f_0 = 0$ and

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

for $0 < k < n$. We then find that

$$f_n = 2 + f_{n-1}.$$

Solving

$$f_{n-1} = \frac{1}{2}(1 + f_n) + \frac{1}{2}(1 + f_{n-2})$$

we find

$$f_{n-1} = 4 + f_{n-2}.$$

This leads us to our claim.

Claim 16.

$$f_k = 2(n - k + 1) + f_{k-1}.$$

Proof. We go by induction. We have the base case above. Assume it holds for k. We must show it holds for $k-1$. Since it holds for k, we have

$$f_{k-1} = 2 + n - k + \frac{1}{2}f_{k-1} + \frac{1}{2}f_{k-2}.$$

Simplifying this, we get

$$f_{k-1} = 2(2 + n - k) + f_{k-2}$$

as desired. **Q.E.D**

We find then that $f_1 = 2n$, $f_2 = 4n - 2$, and $f_3 = 6n - 6$. This leads us to our next claim.

Claim 17. We have $f_k = k(2n + 1 - k)$.

Proof. It's another induction argument. **Q.E.D**

Thus, we just need to substitute n in for k to find

$$\mathbb{E}_n(\tau) = n(n+1).$$

Problem 4. By comparing the integral of $1/x$ with its Riemann sums, show that

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

Solution. It's clear that

$$\int_1^n \frac{1}{x} dx \leqslant \sum_{k=1}^n k^{-1}$$

by examining rectangles of height $1/k$ above the interval $[k, k+1]$, and noticing the union of rectangles is the upper Riemann sum. Thus, we get

$$\log(n) \leqslant \sum_{k=1}^n k^{-1}.$$

For the upper bound, we can fit all the terms but the first to get

$$\sum_{k=1}^n k^{-1} \leqslant 1 + \int_1^n \frac{1}{x} dx = 1 + \log(n).$$

Problem 5. Let P be the transition matrix for the Ehrenfest chain described earlier. Show that the binomial distribution with parameters n and $1/2$ is the stationary distribution for this chain.

Solution. Recall we have

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

So we just need to check

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

for arbitrary $y \in \Omega$. Since there are only two things to consider, just above and just below y, we have

$$\sum_{x \in \Omega} \pi(x) P(x,y) = \binom{n}{y-1} \frac{n-(y-1)}{n} \frac{1}{2^n} + \frac{y+1}{n} \binom{n}{y+1} \frac{1}{2^n} = \frac{1}{2^n} \binom{n}{y} = \pi(y)$$

as desired.

Problem 6. Give an example of a random walk on a finite abelian group which is not reversible.

Solution. Consider the biased random walk on the n-cycle where $p \neq \frac{1}{2}$. Then we have

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

Problem 7. Show that if a random walk on a group is reversible, then the increment distribution is symmetric.

Solution. We have $\mu(k) = P(g, kg)$ and $\mu(k^{-1}) = P(kg, g)$. Therefore, by the detailed balance equations, since it's reversible we have

$$\pi(g)P(g, kg) = \pi(kg)P(kg, g).$$

However, since π is the uniform distribution, we have $\pi(g) = \pi(kg)$. This then gives

$$P(g, kg) = P(kg, g) \to \mu(k) = \mu(k^{-1}).$$

3 Markov Chain Mixing

Definition. The total variation distance between two probability distribution μ and ν on Ω is defined by

$$||\mu - \nu||_{TV} = \max_{A \subseteq \Omega} |\mu(A) - \nu(A)|.$$

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

Example 5. A certain frog lives in a pond with two lily pads, east and west. He has two coins on each lily pad, and each day the frog decides whether two jump by tossing the current lily pad's coin. If the coin lands heads up, the frog jumps to the other lily pad. If the coin lands tails up, he remains where he is. Say he has probability p from jumping 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

$$\pi = \left(\frac{q}{p+q}, \frac{p}{p+q} \right).$$

Claim 18. The stationary distribution is as above.

Proof. Notice that we have

$$\sum_{x \in \Omega} \pi(x)P(x, 1) = \frac{q(1-p)}{p+q} + \frac{pq}{p+q} = \frac{q}{p+q} = \pi(1)$$

and

$$\sum_{x \in \Omega} \pi(x)P(x, 2) = \frac{pq}{p+q} + \frac{p(1-q)}{p+q} = \frac{p}{p+q} = \pi(2).$$

Q.E.D

Assume the frog starts at the east pad ($\mu_0 = (1, 0)$) and define

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

Since there are only two states, there are only four possible events $A \subseteq \Omega$.

Claim 19. We have

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

Proof. This is very easy and intuitive. It just involves checking manually that the calculation comes out fine. **Q.E.D**

Also notice that $\triangle_t = (1 - p - q)^t \triangle_0$. Hence, for this two-state chain, the total variation distance decreases exponentially fast as t increases.

Remark. Note that $(1 - p - q)$ is an eigenvalue of P.

The definition of total variation distance is a maximum over all subsets of Ω, so using this definition is extremely inconvenient. We follow this up with three alternatives.

Proposition 20. Let μ and v be two probability distributions on Ω. Then

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

Proof. Let $B = \{x : \mu(x) \geqslant \nu(x)\}$ and let $A \subseteq \Omega$ be any event. Then

$$\mu(A) - \nu(A) \leqslant \mu(A \cap B) - \nu(A \cap B) \leqslant \mu(B) - \nu(B).$$

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 probability.

By exactly parallel reasoning,

$$\nu(v) - \mu(A) \leqslant \nu(B^c) - \mu(B^c).$$

The upper bounds on the right-hand sides of the above equations are actually the same. Furthermore, when we take $A = B$ (or B^c), then $|\mu(A) - \nu(A)|$ is equal to the upper bound. Thus

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

Q.E.D

Remark. The proof of the prior proposition also shows that

$$||\mu - \nu||_{TV} = \sum_{x \in \Omega, \mu(x) \geqslant \nu(V)} |\mu(x) - \nu(x)|.$$

Remark. From the prior proposition and the triangle inequality for real numbers, it is easy to see that total variation distance satisfies the triangle inequality: for probability distributions μ, ν, and ζ,

$$||\mu - \nu||_{TV} \leqslant ||\mu - \zeta||_{TV} + ||\zeta - \nu||_{TV}.$$

Proposition 21. Let μ and ν be two probability distributions on Ω. Then the total variation distance between them satisfies

$$||\mu{-}\nu||_{TV} = \frac{1}{2}\sup\left\{\sum_{x\in\Omega} f(x)\mu(x){-}\sum_{x\in\Omega} f(x)\nu(x) : f \text{ satisfying } \max_{x\in\Omega}|f(x)| \leqslant 1\right\}.$$

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

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

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

$$f^*(x) = \begin{cases} 1, & \text{if x satisfies } \mu(x) \geqslant \nu(x), \\ -1, & \text{otherwise.}\end{cases}$$

Then

$$\frac{1}{2}\left[\sum_{x\in\Omega} f^*(x)\mu(x) - \sum_{x\in\Omega} f^*(x)\nu(x)\right] = \frac{1}{2}\sum_{x\in\Omega} f^*(x)[\mu(x) - \nu(x)]$$

$$= \frac{1}{2}\left[\sum_{x\in\Omega,\mu(x)\geqslant\nu(x)}[\mu(x) - \nu(x)] + \sum_{x\in\Omega,\nu(x)>\mu(x)}[\nu(x) - \mu(x)]\right].$$

Using the prior proposition shows that the right-hand side above equals $||\mu - \nu||_{TV}$. Hence, the right-hand side of the proposition is at least $||\mu - \nu||_{TV}$. **Q.E.D**

Definition. 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 $P\{X = x\} = \mu(x)$ and $P\{Y = y\} = \nu(y)$.

Example 6. Let μ and ν both be the 'fair coin' measure giving 1/2 to the elements of $\{0, 1\}$.

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

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

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

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

and

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

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

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

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

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

Proposition 22. Let μ and ν be two probability distributions on Ω. Then

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

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

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

$$\begin{aligned} \mu(A) - \nu(A) &= P\{X \in A\} - P\{Y \in A\} \\ &\leqslant P\{X \in A, Y \notin A\} \leqslant P\{X \neq Y\}. \end{aligned}$$

It immediately follows that

$$||\mu - \nu||_{TV} \leqslant \inf\{P\{X \neq Y\} : (X, y) \text{ is a coupling of } \mu \text{ and } \nu\}.$$

It will suffice to construct a coupling for which $P\{X \neq Y\}$ is exactly equal to $||\mu - \nu||_{TV}$. We will do so by forcing X and Y to be equal as often as they possible can. There is a good outline in the book that I will not replicate here. **Q.E.D**

The survey paper (which can be found here) talks a little about total variation distance. To keep things together, I'll outline some of it here as well.

Definition. The paper defines total variation distance (which is really equivalent) between two probability measures μ and ν on Ω as

$$d_{TV}(x,y) = ||\mu - \nu||_{TV} = \sup_{A \subseteq \Omega} \{\mu(A) - \nu(A)\}.$$

Definition. The maximal distance is defined to be

$$d(t) := \max_{x \in \Omega} ||P^t(x,\cdot) - \pi||.$$

If we have two places that we're looking at, we'll use an alternative function. To help, we make the definition

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

Remark. Be careful on the notation here; though it uses d, it is not the same as the prior definition.

Definition. We set the maximal distance (in terms of the paper) as

$$d_{\pi,p}(\mu,\nu) = ||f-g||_p = \left(\sum_{x \in \Omega} |f(x) - g(x)|^p \pi(x) \right)^{1/p}$$

and we customarily set

$$d_{\pi,\infty}(\mu,\nu) = \max\{|f-g|\}.$$

Remark. Setting $\mu(f) = \sum f\mu$ and $p = 1$, we have

$$d_{\pi,p}(\mu,\nu) = ||f-g|| = \sum_{x\in\Omega} |f(x) - g(x)|\pi(x) \leqslant \sum_{x \in \Omega} |f(x) - g(x)|$$

$$= 2\left(\frac{1}{2} \sum_{x \in \Omega} |f(x) - g(x)| \right) = 2d_{TV}(\mu,\nu) = 2||\mu - \nu||_{TV} = \max_{||f||_\infty = 1} \{|\mu(f) - \nu(f)|\}.$$

For $p = 2$, notice we have

$$d_{\pi,2}(\mu,\nu) = \left(\sum_{x \in \Omega} \left| \frac{\mu(x)}{\pi(x)} - \frac{\nu(x)}{\pi(x)} \right|^2 \pi(x) \right)^{1/2}$$

Remark. Notice that we can use Jensen's inequality to establish the mapping $d \mapsto d_{\pi,p}$ is a non-decreasing function, which means

$$d_{\pi,1}(\mu,\nu) = 2d_{TV}(\mu,\nu) \leqslant d_{\pi,2}(\mu,\nu) \leqslant d_{\pi,\infty}(\mu,\nu).$$

We now want to talk about Markov chains converging to their stationary distributions.

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

$$\max_{x\in\Omega} ||P^t(x,\cdot) - \pi||_{TV} \leqslant C\alpha^t.$$

Remark. This can be improved/expanded upon, as we see in the paper. In the language of the paper, let K be a Markov kernel with invariant probability distribution π. Then for any fixed $1 \leqslant p \leqslant \infty$, $n \mapsto \sup_{x\in\Omega} d_{\pi,p}(K_n(x,\cdot),\pi)$ is non-decreasing subadditive function. Moreover, if we have

$$\sup_{x\in\Omega} d_{\pi,p}(K_m(x,\cdot) \leqslant \beta$$

for some fixed integer n, then we have $\forall m \in \mathbb{N}$,

$$\sup_{x\in\Omega} d_{\pi,p}(K_m(x,\cdot),\pi) \leqslant \beta^{m/n}.$$

Remark. Sub-remark – I forgot if this relies on the Markov chain being over a group or not.

Proof. To save some time, I'll skip over the proof for now (we did it in Peterson's class). It is rather intuitive (as noticed by Graham's talk). **Q.E.D**

Remark. Because of Theorem 4.9, the distribution π is also called the equilibrium distribution.

We can make some relationships between d and $\bar{d}$.

Lemma. If $d(t)$ and $\bar{d}(t)$ are as defined above, then

$$d(t) \leqslant \bar{d}(t) \leqslant 2d(t).$$

Proof. It is very easy to show the upper bound. We have

$$\bar{d}(t) = \max_{x,y\in\Omega} ||P^t(x,\cdot) - P^t(y,\cdot)||_{TV}$$

$$= ||(P^t(x,\cdot) - \pi) - (P^t(y,\cdot) - \pi)||_{TV} \leqslant 2||P^t(x,\cdot) - \pi|| = 2d(t)$$

by the triangle inequality.

The lower bound is more interesting. To show that $d(t) \leqslant \bar{d}(t)$, note that first, since π is stationary, we have $\pi(A) = \sum_{y\in\Omega} \pi(y)P^t(y,A)$ for any set A. Using this shows that

$$||P^t(x,\cdot) - \pi||_{TV} = \max_{A\subset\Omega} |P^t(x,A) - \pi(A)|$$

$$= \max_{A\subset\Omega} \left| \sum_{y\in\Omega} \pi(y)[P^t(x,A) - P^t(y,A)] \right|.$$

Now, use the triangle inequality and the fact that the maximum of the sum is not larger than the sum over a maximum to move things around and get

$$\max_{A\subset\Omega}\sum_{y\in\Omega}|P^t(x,A)-P^t(y,A)|\leqslant\sum_{y\in\Omega}\pi(y)\max_{A\subset\Omega}|P^t(x,A)-P^t(y,A)|$$

$$=\sum_{y\in\Omega}\pi(y)||P^t(x,\cdot)-P^t(y,\cdot)||_{TV}\leqslant\max_{y\in\Omega}||P^t(x,\cdot)-P^t(y,\cdot)||_{TV}$$

which is the desired result. **Q.E.D**

Lemma. We have equivalently, for $\mathfrak{P}$ a collection of all probability distributions on Ω,

$$d(t):=\sup_{\mu\in\mathfrak{P}}||\mu P^t-\pi||_{TV},$$

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

Proof. We will show the first part (which consequently will give us the second pretty easily). For the first, notice that one direction is easy; We have that

$$\max_{x\in\Omega}||P^t(x,\cdot)-\pi||\leqslant\sup_{\mu\in\mathfrak{P}}||P^t(x,\cdot)-\pi||$$

since we can choose μ to be the vector which choose our maximal point x. For the other direction, notice that we have that the weighted average will be less than the maximum, and that μ will simply weight our average. We can also get an alternative proof from one of the remarks above. **Q.E.D**

Lemma. The function $\bar{d}$ is submultiplicative. That is to say, $\bar{d}(s+t)\leqslant\bar{d}(s)\bar{d}(t)$.

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 the optimal coupling proposition. Hence,

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

As P^{s+t} is the matrix product of P^t and P^s and the distribution of X_s is P^s, we have

$$P^{s+t}(x,w)=\sum_z P^s(x,z)P^t(z,w)=\sum_z P\{X_s=z\}P^t(z,w)=\mathbb{E}\big(P^t(X_s,w)\big).$$

Combining this with the symmetric identity, $P^{s+t}(y,w)=\mathbb{E}\big(P^t(Y_s,w)\big)$ allows us to write

$$P^{s+t}(x,w)-P^{s+t}(y,w)=\mathbb{E}\big(P^t(X_s,w)-P^t(Y_s,w)\big).$$

(Here, we implicitly used the linearity of expectation.) Combining expectations is possible since X_s and Y_s are defined together on the same probability space. Summing this over $w\in\Omega$ and applying an earlier proposition gives us

$$||P^{s+t}(x,\cdot)-P^{s+t}(y,\cdot)||_{TV}=\frac{1}{2}\sum_w|\mathbb{E}\big(P^t(X_s,w)-P^t(Y_s,w)\big)|.$$

The right hand side is less than or equal to

$$\mathbb{E}\bigg(\frac{1}{2}\sum_{w}|P^t(X_s,w)-P^t(Y_s,w)|\bigg).$$

Applying the same proposition to before, we have that the quantity inside the expectation is exactly

$$||P^t(X_s,\cdot)-P^t(Y_s,\cdot)||_{TV},$$

which is zero whenever $X_s = Y_s$. Moreover, this distance is always bounded by $\bar{d}(t)$. So, this shows that

$$||P^{s+t}(x,\cdot)-P^{s+t}(y,\cdot)||_{TV} \leqslant \bar{d}(t)\mathbb{E}(\nVdash_{\{X_s\neq Y_s\}} = \bar{d}(t)\mathbb{P}\{X_s \neq Y_s\}.$$

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 this over x and y completes the proof. **Q.E.D**

Using the exercise below, we have that $\bar{d}(t)$ is non-increasing in t. From this and an above lemma, it follows that when c is any non-negative integer and t is any non-negative integer, we have

$$d(ct) \leqslant \bar{d}(ct) \leqslant \bar{d}(t)^c.$$

Definition. A useful parameter to study is the mixing time, which we define by

$$t_{\text{mix}}(\epsilon) := \min\{t : d(t) \leqslant \epsilon\}$$

and

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

Corollary. The prior lemma gives us that, for $l \in \mathbb{Z}_{\geqslant 0}$,

$$d(lt_{\text{mix}}(\epsilon)) \leqslant \bar{d}(lt_{\text{mix}}(\epsilon)) \leqslant \bar{d}(t_{\text{mix}}(\epsilon))^l \leqslant (2\epsilon)^l.$$

In particular, taking $\epsilon = 1/4$, we have

$$d(lt_{\text{mix}}(1/4)) \leqslant 2^{-l}.$$

This then gives us

$$t_{\text{mix}}(\epsilon) \leqslant \lceil \log_2 \epsilon^{-1} \rceil t_{\text{mix}}.$$

Definition. For a distribution μ on a group G, the inverse distribution $\hat{\mu}$ is defined by $\hat{\mu} := \mu(g^{-1})$ for all $g \in G$.

Definition. Let P be the transition matrix of the random walk with increment distribution μ. Then the random walk with increment distribution $\hat{\mu}$ is exactly the time reversal $\hat{P}$ of P.

Lemma. Let P be the transition matrix of a random walk on a group G with increment distribution μ, and let $\hat{P}$ be that of the walk on G with increment distribution $\hat{\mu}$. Let π be the uniform distribution on G. Then for any $t \geqslant 0$,

$$||P^t(\mathrm{id}, \cdot) - \pi||_{TV} = ||\hat{P}^t(\mathrm{id}, \cdot) - \pi||_{TV}.$$

Proof. Let $(X_t) = (\mathrm{id}, X_1, X_2, \ldots)$ be a Markov chain with transition matrix P and initial state id. We can write $X_k = g_1 g_2 \cdots g_k$, where $g_i \in G$ are independent choices from the distribution μ. Similarly, let (Y_t) be a chain with transition matrix $\hat{P}$, with increments $h_1, h_2, \ldots \in G$ chosen independently from $\hat{\mu}$. For any fixed elements $a_1, \ldots, a_t \in G$,

$$P\{g_1 = a_1, \ldots, g_t = a_t\} = P\{h_1 = a_t^{-1}, \ldots, h_t = a_1^{-1}\},$$

by definition of $\hat{P}$. Summing over all strings such that $a_1 a_2 \cdots a_t = a$ yields

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

Hence,

$$\sum_{a \in G} \left|P^t(\mathrm{id}, a) - |G|^{-1}\right| = \sum_{a \in G} \left|\hat{P}^t(\mathrm{id}, a^{-1}) - |G|^{-1}\right| = \sum_{a \in G} \left|\hat{P}^t(\mathrm{id}, a) - |G|^{-1}\right|$$

and using the fact that

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

we have the result. **Q.E.D**

Corollary. If t_{mix} is the mixing time of a random walk on a group and $\hat{t}_{\text{mix}}$ is the mixing time of the inverse walk, then we have $t_{\text{mix}} = \hat{t}_{\text{mix}}$.

We finish by talking about the ergodic theorem. The philosophy of the ergodic theorem is ”time averages equals space averages”.

Definition. Let f be a real valued function on Ω and μ be any probability distribution on Ω. We define

$$\mathbb{E}_\mu(f) = \sum_{x \in \Omega} f(x)\mu(x).$$

Theorem (Ergodic Theorem)**.** Let f be a real value function defined on Ω. If (X_t) is an irreducible Markov chain (notice we don’t need aperiodic), then for any starting distribution μ we have

$$P_\mu\left\{ \lim_{t \to \infty} \frac{1}{t} \sum_{s=0}^{t-1} f(X_s) = \mathbb{E}_\pi(f) \right\} = 1.$$

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

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

Since the chain starts anew 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. If we set

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

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

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

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

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

Thus,

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

We can then show that $\mathbb{E}_x(Y_1) = \mathbb{E}_\pi(f)\mathbb{E}_x(\tau^+_x)$. Thus, we get

$$P_x\left\{\lim_{n\to\infty} \frac{S_{\tau^+_{x,n}}}{\tau^+_{x,n}} = \mathbb{E}_\pi(f)\right\} = 1.$$

By the second problem, we have that the theorem holds when $\mu = \delta_x$, the probability distribution with unit mass at x. Averaging over the starting state completes the proof. **Q.E.D**

Exercises

Problem 8. Let P be the transition matrix of a Markov chain with state space Ω and let μ and ν be any two distributions on Ω. Prove that

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

Solution. We can simplify this to

$$||\mu P - \nu P||_{TV} = ||(\mu - \nu)P||_{TV}.$$

Rewriting this, we have

$$= \max_{A\subset\Omega} |(\mu - \nu)P(A)| \leqslant \max_{A\subset\Omega} |(\mu - \nu)(A)| = ||\mu - \nu||_{TV}$$

Problem 9. Let (a_n) be a bounded sequence. If, for a sequence of integers (n_k) satisfying $\lim_{k\to\infty} n_k/n_{k+1} = 1$, we have

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

then

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

Solution. I don't recall this from Analysis, but it seems to be a reasonable assumption.

Let $y_n = \frac{1}{n}\sum_{i=1}^{n} a_i$. Then we have that $y_{n_k} \to 1$ as $k \to \infty$, with $\lim_{k\to\infty} n_k/n_{k+1} = 1$. Since these are asymptotically equivalent, we can deduce monotonicity on a domain extending to infinity. Notice that $\lim_{k\to\infty} y_{n_k}/y_{n_k+1} = 1$, thus giving us monotonicity in the sums of the a_i on some infinite domain. We have monotone and bounded, so we get convergence, and they must converge to the same limit.

4 Coupling

We'll start by recalling the definition

Definition. 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 Y is ν.

Remark. We have $P_{x,y}$ will be the probability on the sapce where X_t and Y_t are defined.

Theorem. Let $\{(X_t, Y_t)\}$ be a coupling of two Markov Chain's satisfying $X_0 = x$ and $Y_0 = y$. Let τ_{couple} be defined by

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

Then

$$||P^t(x, \cdot) - P^t(y, \cdot)||_{TV} \leqslant P_{x,y}\{\tau_{\text{couple}} > t\}.$$

Proof. Notice $P^t(x, z) = P_{x,y}\{X_t = z\}$, $P^t(y, z) = P_{x,y}\{Y_t = z\}$. Then by a prior proposition, we can write this as

$$||P^t(x, \cdot) - P^t(y, \cdot)||_{TV} \leqslant P_{x,y}\{X_t \neq Y_t\}$$

and we can notice that this last part is obviously $P_{x,y}\{\tau_{\text{couple}} < t\}$. **Q.E.D**

Corollary. We have

$$d(t) \leqslant \max_{x,y \in \Omega} P_{x,y}\{\tau_{\text{couple}} > t\}.$$

Definition. A Markovian coupling of P is a Markov Chain with state space $\Omega \times \Omega$ whose transition matrix Q satisfies

(a) $\forall x, y, x'$ we have $\sum_{y'} Q((x, y), (x', y')) = P(x, x')$.

(b) $\forall x, y, y'$ we have $\sum_{x'} Q((x, y), (x', y')) = P(y, y')$.

Remark. In order to proceed moving forward, we'll need Markov's inequality. Markov's inequality is as follows; if X is a nonnegative random variable $a > 0$, then we have

$$P(X \geqslant a) \leqslant \frac{\mathbb{E}(X)}{a}.$$

Example 7 (Random Walk on $\mathbb{Z}_n$)**.** We have that (X_t, Y_t) moves as follows – flip a coin. If it's heads, X_t moves, and if it's tails then Y_t moves. In each case, we have that they move by flipping another coin. Once they connect, they move together. Let D_t be the distance between them – in other words, $D_t \in \{0, 1, \ldots, n\}$, and it gets absorbed at either 0 or n. Recall $\mathbb{E}_{x,y}(\tau) = k(n-k)$ from a prior exercise. Then using Corollary 5.3, the Markov Inequality, and noticing that $\mathbb{E}_{x,y}(\tau)$ is maximized at $k = n/2$, we have

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

For the next example, we'll need Wald's identity.

Remark. Let (X_n) be a collection of random variables which are i.i.d. and let N be a nonnegative integer-valued random variable that is independent of the X_i. If the X_i and N have finite expectation, then we have

$$\mathbb{E}\left(\sum_{i=1}^{N} X_i\right) = \mathbb{E}(N)\mathbb{E}(X_1).$$

Theorem. For the lazy random walk on the d-dimensional taurus $\mathbb{Z}_n^d$, we have

$$\tau_{\text{mix}}(\epsilon) \leqslant c(d)n^2 \log_2(\epsilon^{-1}),$$

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

Proof. Couple a random walk $(\vec{X}_t)$ startng at $\vec{x}$ and $(\vec{Y}_t)$ starting at $\vec{y}$. Randomly choose a coordinate d uniformly. If $(\vec{X}_t)$, $(\vec{Y}_t)$ agree on the chosen coordinate, then move them bot $+1$, -1, or 0 with probability $1/4$, $1/4$, and $1/2$ respectively. If they disagree, choose on of the chains at random and fix the other. Move $+1$ or -1 in the coordinate with probability $1/2$. Let

$$\vec{X}_t = (X_t^t, \ldots, X_t^d)$$

and

$$\vec{Y}_t = (T_t^1, \ldots, Y_t^d)$$

and let

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

Using Wald's identity and the fact that there is a geometric waiting time between each coordinate with mean d, we have $\mathbb{E}_{x,y}(\tau_i) = \frac{dn^2}{4}$. Now $\tau_{\text{couple}} = \max_{1\leqslant i\leqslant d} \tau_i$ and bounding it above by a sum gets us

$$\mathbb{E}_{x,y}(\tau_{\text{couple}}) \leqslant \frac{d^2n^2}{4}.$$

So $P_{x,y}\{\tau_{\text{couple}} > t\} \leqslant \frac{d^2n^2}{4t}$. Using Proposition 4.36, we get $\tau_{\text{mix}}(\epsilon) \leqslant d^2n^2\lceil\log_2(\epsilon^{-1})\rceil$.
Q.E.D

For the next example, we'll need some basic graph theory definitions, which we'll review.

Definition. A tree is a connected graph with no cycles.

Definition. A rooted tree has a distinguished vertex, called the root.

Definition. The depth of a vertex v is its graph distance to the root.

Definition. A level of the tree consists of all the vertices at the same depth.

Definition. The children of v are the neighbors of v with depth larger than v.

Definition. A leaf is a vertex of degree one.

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

(a) v_0 has degree b.

(b) Every vertex with distance j from the root, where $1 \leqslant j \leqslant k-1$, has degree $b+1$.

(c) The vertex at distance k are leafs.

Remark. There are $n = \frac{b^{k+1}-1}{b-1}$ vertices in $T_{b,k}$.

Example 8. In this example, we consider the random walk on the finite binary tree, $T_{2,k}$. The walk remains at its current position with probability 1/2. Consider the following coupling (X_t, Y_t) of two lazy random walks, started from states x_0 and y_0 on the tree. Assume without loss of generality that x_0 is at least as close to the root as y_0 (can do this arbitrarily). At each move, toss a fair coin to decide which of the two chains moves; if heads, $Y_{t+1} = Y_t$ while X_{t+1} is chosen from the neighbors of X_t uniformly at random. If the coin toss is tails, then $X_{t+1} = X_t$ and Y_{t+1} is chosen from the neighbors of Y_t uniformly at random. Run the two chains according to this rule until the first time they are at the same level of the tree. Once the two chains are at the same level, change the coupling to the following update rule: let X_t evolve as a lazy random walk, and couple Y_t to X_t so that Y_t moves closer to (further from) the root if and only if X_t moves closer to (further from) the root, respectively. Let B be the set of leaves. Observe that if (X_t) has first visited B and then visited the root, it must have coupled at this time. The expected value of this time is less than the commute time τ from the root to B, the time it takes starting from the root to first visit the set B and then return to the root. It will be shown later that $\mathbb{E}(\tau) \leqslant 4n$. Thus, if τ_{couple} is the time when the two particles meet, we have $P_{x,y}\{\tau_{\text{couple}} > t\} \leqslant \frac{4n}{t}$. We conclude that $t_{\text{mix}} \leqslant 16n$.

Proposition 23. Let Q be an irreducible transition matrix and consider the lazy chain with transition matrix $P = (Q+I)/2$. The distribution at time t and $t+1$ satisfy

$$||P^t(x,\cdot) - P^{t+1}(x,\cdot)||_{TV} \leqslant \frac{12}{\sqrt{t}}.$$

Proof. The proof involves clever coupling and Proposition 2.17. **Q.E.D**

Definition (Grand Coupling)**.** Construct a collection of random variables $\{X_t^x : x \in \Omega, t = 0, 1, 2, \ldots\}$ such that for each $x \in \Omega$, the sequence $(X_t^x)_{t=0}^{\infty}$ is a Markov chain with transition matrix P which started from x. We can use the random mapping construction to make grand couplings. Let $f : \Omega \times \Lambda \to \mathbb{R}$ be a function and Z and Λ-valued random variable such that $P(x, y) = P\{f(x, Z) = y\}$. Proposition 1.5 guarantees such a (f, Z) pair exists. Let $\{Z_i\}_{i \geqslant 0}$ be an i.i.d. sequence with the same distribution as Z, and define inductively $X_0^x = x$, $X_t^x = f(X_{t-1}^x, Z)$ for $T \geqslant 1$. This yields a grand coupling.

Exercises

Problem 10. (a) Show that when (X_t, Y_t) is a coupling satisfying the normal properties for which $X_0 \sim \mu$ and $Y_0 \sim \nu$, then

$$||\mu P^t - \nu P^t||_{TV} \leqslant P\{\tau_{\text{couple}} > t\}.$$

(b) If (X_t) and (Y_t) are independent, then they surely coalesce. That is, $P\{\tau_{\text{couple}} < \infty\} = 1$.

Solution. (a) This is just an application of two different results. Proposition 4.7 says $||\mu - \nu||_{TV} \leqslant P\{\tau_{\text{couple}} > t\}$. From Exercise 4.3, we know $||\mu P - \nu P||_{TV} \leqslant ||\mu - \nu||_{TV}$, and so as a consequence we have $||\mu P^t - \nu P^t||_{TV} \leqslant ||\mu - \nu||_{TV}$. This then gives us the desired result.

(b) This is a clever trick. We have $P(X_t \neq Y_t | X_0, Y_0) \leqslant 1 - \epsilon$ by ergodicity. Now, we have $P(X_{2t} \neq Y_{2t} | X_t \neq Y_t) \leqslant 1 - \epsilon$ by the Markov property. We then have $P(X_{2t} \neq Y_{2t} \cap X_t \neq Y_t | X_0, Y_0) = P(X_{2t} \neq Y_{2t} | X_0, Y_0) \leqslant (1-\epsilon)^2$. Continue to see that eventually they must coalesce.

5 Strong Stationary Times

Consider the top-to-random shuffle. Let τ_{top} be the time one move after the first occasion when the original bottom card has moved to the top of the deck.

Proposition 24. Let (X_t) be the random walk on S_n corresponding to the top-to-random shuffle on n cards. Given at time t that there are k cards under the original bottom card, each of the $k!$ possible orderings of these cards are equally likely. Therefore, if τ_{top} is one shuffle after the first time that the original bottom and moves to the top of the deck, then the distribution of $X_{\tau_{\text{top}}}$ is uniform over S_n, and the time τ_{top} is independent of $X_{\tau_{\text{top}}}$.

Proof. It's a simple induction proof. For $k = 0, 1$ this is clear, and so we have base cases established. Assume it holds for k. Then when we take the top card and place it randomly in the deck, it is either below the original bottom card or above it. If it is above it, nothing changes and we continue. If it is below it, then we have that it is placed with uniform probability in any of the $k + 1$ remaining places, and thus we have that the $k + 1$ cards beneath the original bottom card are all uniformly random. Once the bottom card is now on the top of the deck, we have that all the cards below it are uniformly random, and we place it uniformly at random in any of the deck, completing the procedure. **Q.E.D**

Definition. Given a sequence $(X_t)_{t=0}^{\infty}$ of Ω-valued random variables, a $\{0, 1, 2, \ldots, \infty\}$-value random variable τ is a stopping time for (X_t) if, for each $t \in \{0, 1, \ldots\}$, there is a set $B_t \subseteq \Omega^{t+1}$ such that $\{\tau = t\} = \{(X_0, X_1, \ldots, X_t) \in B_t\}$.

Remark. The random mapping representation of the random walk on the hypercube is given by $\{1, 2, \ldots, n\} \times \{0, 1\}$ where you are selecting an element

(j, B), where the coordinate J of the current state is updated with B. Define $\tau_{\text{refresh}} := \min\{t \geqslant 0 : \{j_1, \dots, j_t\} = \{1, 2, \dots, n\}\}$. We have then $X_{t_{\text{refresh}}}$ is exactly the sample from the stationary distribution π. Notice τ_{refresh} is a stopping time for (Z_t). Recall we defined $(X_t)_{t=0}^{\infty}$ inductively as follows: $X_0 = x, X_t = f(X_{t-1}, Z_t)$.

Definition. A randomized stopping time for the Markov chain (X_t) is a stopping time τ for the sequence (Z_t).

Definition. Let (X_t) be an irreducible Markov chain with stationary distribution π. A stationary time τ for (X_t) is a randomized stopping time such that the distribution of X_τ is π: $P_x\{X_\tau = y\} = \pi(y)$.

Definition. A strong stationary time for a Markov chain (X_t) with stationary distribution π is a randomized stopping time τ, possibly depending on a starting position x, such that $P_x\{\tau = t, X_t = y\} = P_x\{\tau = t\}\pi(y)$.

Example 9. The top-to-random shuffle forms a strong stationary time, s we outlined in the proposition.

Lemma. Let (X_t) be an irreducible Markov chain with stationary distribution π. It τ is a strong stationary time for (X_t), then for all $t \geqslant 0$,

$$P_x\{\tau \leqslant t, X_t = y\} = P\{\tau \leqslant t\}\pi(y).$$

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

$$P_x\{\tau = S, X_t = y\} = \sum_{z \in \Omega} P_x\{X_t = y | \tau = s, X_s = z\} P_x\{\tau = s, X_s = z\}.$$

Since τ is a stopping time for (Z_t), the event $\{\tau = s\}$ equals $\{(Z_1, Z_2, \dots, Z_s) \in B\}$ for some set $B \subset \Omega^s$. Also, for integers $r, s \geqslant 0$, there exists a function $\bar{f}_r : \Omega^{r+1} \to \Omega$ such that $X_{s+r} = \bar{f}_r(X_s, Z_{s+1}, \dots, Z_{s+r})$. Since $(Z_1, \dots, Z_s$ and $(Z_{s+1}, \dots, Z_t)$ are independent,

$$\begin{aligned} P_x\{X_1 = y | \tau = s, X_s = z\} &= P_x\{\bar{f}_{t-s}(z, Z_{s+1}, \dots, Z_t) = y | (X_1, \dots, X_s) \in B, X_s = Z\} \\ &= P^{t-s}(z, y). \end{aligned}$$

Summing over the $s \leqslant t$ gives $P\{\tau \leqslant t\}\pi(y)$. **Q.E.D**

We want to eventually show that $d(t) \leqslant \max_{x \in \Omega} P_x\{\tau > t\}$. In order to do so, we will need some definitions and lemmas.

Definition. Define the separation distance by

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

and

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

Lemma. If τ is a strong stationary time, then $s_x(t) \leqslant P_x\{\tau > t\}$.

Proof. Fix $x \in \Omega$. For all $y \in \Omega$, notice that we have

$$1 - \frac{P^t(x,y)}{\pi(y)} = 1 - \frac{P_x\{X_t = y\}}{\pi(y)} \leqslant 1 - \frac{P_x\{X_t = y, \tau \leqslant t\}}{\pi(y)}.$$

By the prior lemma, we have

$$1 - \frac{P_x\{X_t = y, \tau \leqslant t\}}{\pi(y)} \leqslant 1 - \frac{P_x\{\tau \leqslant t\}\pi(y)}{\pi(y)} = P_x\{\tau > t\}$$

as we desired. **Q.E.D**

Definition. Given a starting state x a state y is a halting time for a stopping time τ if $X_t = y$ implies $\tau \leqslant t$.

Remark. The inequality in the prior lemma is an equality if and only if y is a halting state for the starting state x, for some y.

Lemma. The seperation distance $s_x(t)$ satisfies

$$||P^t(x,\cdot) - \pi||_{TV} \leqslant s_x(t)$$

thus giving us $d(t) \leqslant s(t)$.

Proof. We have

$$\begin{aligned} d(t) := \max_{x\in\Omega} ||P^t(x,\cdot) - \pi|| &= \sum_{\substack{y\in\Omega \\ P^t(x,y)<\pi(y)}} [\pi(y) - P^t(x,y)] \\ &= \sum_{\substack{y\in\Omega \\ P^t(x,y)<\pi(y)}} \pi(y)\left[1 - \frac{P^t(x,y)}{\pi(y)}\right] \\ &\leqslant \max_{y\in\Omega}\left[1 - \frac{P^t(x,y)}{\pi(y)}\right] = s_x(t) \leqslant s(t). \end{aligned}$$

Q.E.D

Combining the two above lemmas gives us the following corollary.

Corollary. If τ is a strong stationary time, then

$$d(t) = \max_{x\in\Omega} ||P^t(x,\cdot) - \pi||_{TV} \leqslant \max_{x\in\Omega} P_x\{\tau > t\}.$$

Example 10. Take two complete graphs on n-vertices and "glue" them together at one vertex. Add n loops to all other vertices, and one lop to the glued vertex. This makes the graph regular of degree $2n-1$ (here, the loops contribute one degree). Let τ be the time one step after $v*$ (the glued vertex) has been visited for the first time. Then τ is a strong stationary time. We have that the probability of going to v^* is $\frac{1}{2n-1}$, and this is geometric. Hence, we get $\mathbb{E}(\tau) = 2n$. By Markov's inequality, we find

$$P_x\{\tau \geqslant t\} \leqslant \frac{\mathbb{E}(\tau)}{t} = \frac{2n}{t}.$$

Taking $t = 8n$ gives us

$$P_x\{\tau \geqslant t\} \leqslant \frac{2n}{8n} = \frac{1}{4}.$$

So, we have $d(t) \leqslant \frac{1}{4}$ if $t = 8n$, and so $t_{\text{mix}} \leqslant 8n$ by definition.

Example 11. Consider the top-to-random shuffle. The probability that a card moves below the original bottom card is $\frac{k}{n}$ if there are k-cards beneath it. We see this is the coupon collector again. Proposition 2.4 gives us

$$P_x\{\tau > [n\log(n) + cn]\} \leqslant e^{-c},$$

and Proposition 6.10 gives

$$d(n\log(n) + cn) \leqslant e^{-c} \to t_{\text{mix}}(\epsilon) \leqslant n\log(n) + \log(\epsilon^{-1})n.$$

Example 12. Imagine a line of books, and after randomly selecting a book you move it to the front. This is the time reversal Markov chain of the top-to-random shuffle, and so using Lemma 4.13 we can bound

$$t_{\text{mix}} \leqslant n\log(n) + n\log(\epsilon^{-1}).$$

Consider a finite chian (X_t) with transition matrix P and stationary distribution π on Ω. Given $t \geqslant 1$, suppose that we chose uniformly a time $\sigma \in \{0, 1, \ldots, t-1\}$ and run the given Markov chain for σ-steps. Then the state X_σ has distribution

$$v_x^t := \frac{1}{t}\sum_{s=0}^{t-1} P^s(x, \cdot).$$

Definition. The Cesaro mixing time $t^*_{\text{mix}}(\epsilon)$ is defined as teh first t such that $\forall x \in \Omega$, $||v_x^t - \pi||_{TV} \leqslant \epsilon$.

Theorem. Consider a finite chain with transition matrix P and stationary distribution π on Ω. If τ is a stationary distribution for the chain, then $t^*_{\text{mix}}(1/4) \leqslant 4\max_{x\in\Omega}\mathbb{E}_x(\tau) + 1$.

Proof. Proof omitted for now. **Q.E.D**

Remark. The converse was proven by Lovasz and Winkler.

Exercises

Problem 11. Show that if τ and τ' are stopping times for the sequence (X_t), then $\tau + \tau'$ is a stopping time for (X_t).

Solution. Simply note that for $n \in \{0, 1, \ldots\}$ we have

$$\{\tau + \tau' = n\} = \bigcup_{i=0}^{n} \big(\{\tau = i\} \cap \{\tau = n - i\}\big).$$

Problem 12. Consider the top-to-random shuffle. Show that the time until the card initially one card from the bottom rises to the top, plus one more move, is a strong stationary time, and find it's expectation.

Solution. The argument is essentially the same as the bottom card argument. The mean is still the coupon collector mean, except we skip the last (first?) one. So it will be $\mathbb{E}(\tau) = n/2 + n/3 + \cdots + 1$.

Problem 13. Let $s(t)$ be the seperation distance. Show that there is a stochastic matrix Q so that $P^t(x, \cdot) = [1 - s(t)]\pi + s(t)Q^t(x, \cdot)$ and $\pi = \pi Q$.

Solution. Showing that it is stochastic is simple. We have

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

Rewrite this as

$$1 = [1 - s(t)] + s(t)Q^t(x, y).$$

Solving for $Q^t(x, y)$, we find 1. More importantly, taking $t = 1$, we get that the matrix Q is stochastic. Next, instead of x, use the stationary distribution π. We have then

$$\pi P = \pi = [1 - s(t)]\pi + s(t)\pi Q.$$

Solving this gives

$$\pi = \pi Q.$$

Problem 14. Show that if

$$\max_{x \in \Omega} P_x\{\tau > t_0\} \leqslant \epsilon$$

then

$$d(t) \leqslant \epsilon^{\lfloor t/t_0 \rfloor}.$$

Solution. We use the submultiplicativity of $s(t)$; that is, $s(t + u) \leqslant s(t)s(u)$. We also use the fact that $d(t) \leqslant s(t)$. We'll just show it for the case of $t = 2t_0$ (all other cases are essentially the same argument). For $t = 2t_0$, we have

$$d(2t_0) \leqslant s(t_0 + t_0) \leqslant s(t_0)^2 \leqslant \max_{x \in \Omega} P_x^2\{\tau > t_0\} \leqslant \epsilon^2 = \epsilon^{2t_0/t_0}.$$

6 Lower Bounds on Mixing Times

The idea of this is simple; if the possible locations of a chain after t steps do not form a significant function of the state space, then the distribution of the chain at time t cannot be close to uniform

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

Denote by Ω_t^x the set of states accessible from x in t steps, and observe that $|\Omega_t^x| \leqslant \Delta^t$. If $\Delta^t \leqslant (1-\epsilon)|\Omega|$, then we get

$$||P^t(x,\cdot) - \pi||_{TV} \geqslant P_t(x, \Omega_t^x) - \pi(\Omega_t^x) \geqslant 1 - \frac{\Delta^t}{|\Omega|} > \epsilon.$$

We just need an upper bound on this t. Solving for t, we get

$$\frac{\Delta^t}{|\Omega|} \leqslant 1 - \epsilon \to t \leqslant \frac{\log(|\Omega|(1-\epsilon))}{\log(\Delta)}$$

giving us

$$t_{\text{mix}}(\epsilon) \geqslant \frac{\log(|\Omega|(1-\epsilon))}{\log(\Delta)}.$$

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

We can find something called the diameter bound. Let P be an irreducible and aperiodict transition matrix on Ω with diameter L, and suppose that x_0 and y_0 are states at maximal graph distance L. Then $P^{\lfloor (L-1)/2 \rfloor}(x_0, \cdot)$ and $P^{\lfloor (L-1)/2 \rfloor}(y_0, \cdot)$ are positive on disjoint vertex sets. Hence, it's clear that $\bar{d}(\lfloor (L-1)/2 \rfloor) = 1$, and for any $\epsilon < 1/2$, $t_{\text{mix}}(\epsilon) \geqslant \frac{1}{2}$.

Definition. The edge measure Q is defined by

$$Q(x,y) := \pi(x)P(x,y),$$

and

$$Q(A,B) := \sum_{x \in A, y \in B} Q(x,y).$$

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

Definition. The bottleneck ratio of the whole chain is defined to be

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

and the bottleneck ratio of the whole chain is

$$\Phi_\star := \min_{\pi(S)\leqslant\frac{1}{2}} \Phi(S).$$

Example 13. For a simple random walk with vertices Ω and edge set E,

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

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. In this case, we get

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

Theorem. If $\Phi_\star$ is the bottleneck ratio, then $t_{\text{mix}} \geqslant \frac{1}{4\Phi_\star}$.

Proof. Proof omitted for now. **Q.E.D**

Example 14. Consider the lazy random walk on the rooted binary tree of depth k. We have $n = 2^{k+1}-1$ is the number of vertices. The number of edges is $n-1$. Let v_0 be the root, and denote v_l, v_r as its descendants. Let S consist of the right hand side of the tree; that is, the descendants of v_r. By Example 1.12, we have

$$\pi(v) = \begin{cases} \frac{2}{2n-2} & \text{for } v = v_0 \\ \frac{3}{2n-2} & \text{for } 0 < |v| < k \\ \frac{1}{2n-2} & \text{for } |v| = k. \end{cases}$$

Notice that in S, we have $2^{k-1}-1$ vertices with $\pi(v) = \frac{3}{2n-2}$ and 2^{k-1} vertices with $\pi(v) = \frac{1}{2n-2}$. Multiplying and adding, we get $\frac{2^{k+1}-3}{2n-2} = \frac{n-2}{2n-2} = \pi(S)$. Since there is only one edge connecting S and S^c, we get $Q(S,S^c) = \pi(v_r)P(v_r,v_0) = \frac{1}{2(n-1)}$. Therefore, $\Phi(S) = \frac{1}{n-2}$. Using the prior theorem, we get

$$t_{\text{mix}} \geqslant \frac{n-2}{4}.$$

Definition. Let f be a statistic, or a real-valued function on Ω. Let μ be a probability distribution on Ω. Then

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

Likewise, $\text{Var}_\mu(f)$ indicates variance computed with respect to the probability distribution μ.

Proposition 25. For $f:\Omega\to R$, define $\sigma_\star^2 := \max\{\text{Var}_\mu(f), \text{Var}_\nu(f)\}$. If

$$|E_\nu(f) - E_\mu(f)| \geqslant r\sigma_\star$$

then

$$||\mu - \nu||_{TV} \geqslant 1 - \frac{8}{r^2}.$$

In particular, if for a Markov chain (X_t) with transition matrix P the function f satisfies

$$|\mathbb{E}_x[f(X_t)] - E_\pi(f)| \geqslant r\sigma_\star,$$

then

$$||P^t(x,\cdot) - \pi||_{TV} \geqslant 1 - \frac{8}{r^2}.$$

We'll need a lemma to prove this. When μ is a probability distribution on Ω and $f : \Omega \to \Lambda$, write μf^{-1} for the probability distribution defined by

$$(\mu f^{-1})(A) := \mu(f^{-1}(A))$$

for $A \subseteq \Lambda$. When X is an Ω-valued random variable with distributin μ, then $f(X)$ has distribution μf^{-1} on Λ.

Lemma. Let μ and ν be probability distributions on Ω, and let $f : \Omega \to \Lambda$ be a function on Ω, where Λ is a finite set. Then

$$||\mu - \nu||_{TV} \geqslant ||\mu f^{-1} - \nu f^{-1}||_{TV}.$$

Proof. Since

$$|\mu f^{-1}(B) - \nu f^{-1}(B)| = |\mu(f^{-1}(B)) - \nu(f^{-1}(B))|,$$

then

$$\max_{B \subset \Lambda} |\mu f^{-1}(B) - \nu f^{-1}(B)| \leqslant \max_{A \subset \Omega} |\mu(A) - \nu(A)|.$$

Q.E.D

We now can prove the proposition.

Proof. Suppose arbitrarily that $E_\mu(f) \leqslant E_\nu(f)$. If $A = (E_\mu(f) + r\sigma_\star/2, \infty)$, then Chebyshev's inequality yields that

$$\mu f^{-1}(A) \leqslant \frac{4}{r^2} \quad \text{and} \quad v f^{-1}(A) \geqslant 1 - \frac{4}{r^2},$$

whence

$$||\mu f^{-1} - \nu f^{-1}||_{TV} \geqslant 1 - \frac{8}{r^2}.$$

The prior lemma now finishes the proof. **Q.E.D**

We can get a better constant for the lower bound with the following proposition.

Proposition 26. Let μ and ν be two probability distributions on Ω, and let f be a real-valued function on Ω. If

$$|E_\mu(f) - E_\nu(f)| \geqslant r\sigma,$$

where $\sigma^2 = [\mathrm{Var}_\mu(f) + \mathrm{Var}_\nu(f)]/2$, then

$$||\mu - \nu||_{TV} \geqslant 1 - \frac{4}{4 + r^2}.$$

Proof. Proof omitted for now. **Q.E.D**

Example 15. We'll use the proposition to bound below the mixing time for the random walk on the hypercube. We'll first prove a lemma.

Lemma. Consider the coupon collector problem with n distinct coupon types, and let $I_j(t)$ be the indicator of the event that the j-th coupon has not been collected by time t. Let $R_t = \sum_{j=1}^n I_j(t)$ be the number of coupon types not collected by time t. The random variables $I_j(t)$ are negatively correlated, and letting $p = (1 - 1/n)^t$, we have for $t \geqslant 0$

$$\mathbb{E}(R_t) = np,$$

$$\mathrm{Var}(R_t) \leqslant np(1-p) \leqslant \frac{n}{4}.$$

Proof. Since we have that $I_j(t)$ is a Bernoulli random variable, we have that $\mathbb{E}(I_j(t)) = p$. Likewise, we get $\mathrm{Var}(I_j(t)) = p(1-p)$. For $j \neq k$, we get

$$\mathbb{E}(I_j(t)I_k(t)) = \left(1 - \frac{2}{n}\right)^t,$$

whence

$$\mathrm{Cov}(I_j(t), I_k(t)) = \left(1 - \frac{2}{n}\right)^t - \left(1 - \frac{1}{n}\right)^{2t} \leqslant 0.$$

The result follows. **Q.E.D**

Exercises

Problem 15. Let $\vec{X}_t = (X_t^1, \ldots, X_t^n)$ be the position of the lazy random walker on the hypercube $\{0,1\}^n$, started at $\vec{X}_0 = \vec{1} = (1, \ldots, 1)$. Show that the covariance between X_t^i and X_t^j is negative. Conclude that if $W(\vec{X}_t) = \sum_{i=1}^n X_t^i$, then $\mathrm{Var}(W(\vec{X}_t)) \leqslant n/4$.

Solution. Let $Y_t^i = 2X_t^i - 1$. Then we have that $Y_t^i = \{-1, 1\}$. We want to then condition on the probabilities. We have that, if the component i is chosen, then it switches between -1 and 1. We construct four events then; event A denotes if i and j are both chosen by a point, B_i denotes only i is chosen, B_j denotes only j is chosen, and C denotes neither were chosen. We get that $\mathbb{E}(X_t^i|A) = 0$

since it is uniform between $\{-1,1\}$, $\mathbb{E}(X_t^i|B_i) = 0$, and the rest are 1. So we find $\mathbb{E}(X_t^i) = P(B_j) + P(C)$. It is a similar situation for X_t^j. For $X_t^iX_t^j$, we have that $\mathbb{E}(X_t^iX_t^j|A) = 0$, $\mathbb{E}(X_t^iX_t^j|B_i) = 0$, $\mathbb{E}(X_t^iX_t^j|B_j) = 0$, and finally $\mathbb{E}(X_t^iX_t^j|C) = 1$. So we get that $\mathbb{E}(X_t^iX_t^j) = P(C)$. So our covariance is $P(C) - (P(B_j) + P(C))^2$. Now, we calculate explicitly $P(B_j)$. Notice that this event is simply $P(B_j) = P(B_i) - P(C)$. Since B_j, B_i are identical, we get that it comes out to $P(C) - P(B_j)^2$. So we have that it is negatively correlated. The result then follows, since $\text{Var}(W(\vec{X}_t)) = \sum_{i=1}^n \text{Var}(X_t^i) + \sum_{i\neq j} \text{Covar}(X_t^i, X_t^j)$. Since it is negatively correlated, this is the same thing as $\text{Var}(W(\vec{X}_t)) \leqslant \sum_{i=1}^n \text{Var}(X_t^i)$. Notice that $\text{Var}(X_t^i) = (1/4)$ and the result follows.

Problem 16. Let $\Omega = GL_n(\mathbb{F}_2)$, the set of invertible $n \times n$ matrices over $\mathbb{F}_2$. Consider the chain which selects uniformly an ordered pair (i,j) of rows $(i \neq j)$ and adds row i to row j, the addition being mod 2.

(a) Show that there is a constant $\gamma > 0$ so that $|\Omega|/2^{n^2} \to \gamma$ as $n \to \infty$.

(b) Show that $t_{\text{mix}} > cn^2/\log(n)$ for a positive constant c.

Solution. I did this on the white board. For the first part, use the exercise from Dummit and Foote to get the limit (show that it's bounded between 0 and 1, which is relatively easy, then show that it's monotonically increasing). For the second part, I noticed that this chain is combinatorially isomorphic to the random walk on $\mathbb{Z}_{|\text{Gl}_n(\mathbb{F}_2)|}$. I then did some analysis by making the walk lazy and getting a lower bound which is similar to the one given in the exercise.

7 The Symmetric Group and Shuffling Cards.

Definition. The set of all bijections from $\{1,\ldots,n\}$ to itself forms the group S_n, also known as the symmetric group on n letters.

Definition. We often use cycle notation. If $a_1,\ldots,a_m$ are elements in our set, then $(a_1a_2\cdots a_m)$ denotes the permutation σ which sends $\sigma(i) \mapsto a_{i+1 \pmod m}$. A transposition is a 2-cycle.

Remark. There is an algorithm for generating an exactly uniform random permutation. Let σ_0 be the identity permutation. For $k = 1,\ldots,n-1$, inductively construct σ_k from σ_{k-1} by swapping the cards (or elements) at locations k and J_k, where J_k is an integer picked uniformly in $\{k,\ldots,n\}$, independently of $\{J_1,\ldots,J_{k-1}\}$. It is rather simple to see that this uniformly creates a random permutation. Let $\eta \in S_n$. Then using this algorithm, we'd like to show that $P\{\sigma_{n-1} = \eta\} = \frac{1}{n!}$. Notice that $P\{\sigma_{n-1} = \eta\} = P\{J_1 = \eta(1) \cap \cdots \cap J_n = \eta(n)\}$. Since the J_i were all independently chosen, this is equivalent to asking $P\{J_1 = \eta(1)\}\cdots P\{J_{n-1} = \eta(n-1)\}$. Since the J_i are chosen uniformly, we have that the probability that these are equal to the $\eta(i)$ is exactly $\frac{1}{n-i}$. Therefore, we get that this product turns out to be $\frac{1}{n!}$.

It also turns out that this algorithm is optimal. Consider the identity permutation and $\sigma = (1 \cdots n)$ on the Cayley graph generated by S_n by permutations. Then it's clear that these are the elements furthest away from eachother, and we also see that it takes $n-1$ edges to go from one another. Hence, the diameter of the graph is $n-1$, and so in order to reach any permutation we must take $n-1$ steps.

Definition. We define the parity of a permutation $\sigma \in S_n$ to be

$$M(\sigma) := \prod_{1 \leqslant i < j \leqslant n} \big(\sigma(j) - \sigma(i)\big).$$

Remark. It is an easy exercise to see that

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

Definition. We call a permutation $\sigma \in S_n$ even if $M(\sigma) > 0$ and odd if $M(\sigma) \leqslant 0$. This is because if we can write σ as a product of even permutations, then we get that all the negatives cancel and so $M(\sigma) > 0$.

Definition. In order to avoid periodicity, the random shuffle transposition is defined as follows: at time t, choose two cards, labeled L_t and R_t, independently and uniformly at random. If L_t and R_t are different, transpose them. Otherwise, do nothing. The resulting distribution μ is then

$$\mu(\sigma) = \begin{cases} \frac{1}{n} & \text{if } \sigma = \text{id} \\ \frac{2}{n^2} & \text{if } \sigma = (ij) \\ 0 & \text{otherwise.} \end{cases}$$

Notice that this walk is irreducible; we have that, for all $h \in S_n$, $P^t(g,h) > 0$ since transpositions generate the group. Aperiodicity follows since $\mu(\text{id}) > 0$, so $\gcd\{t : P^t(g,g) > 0\} = 1$.

Proposition 27. Let $0 < \epsilon < 1$. For the random transposition chain on an n-card deck,

$$t_{\text{mix}}(\epsilon) \geqslant \frac{n-1}{2} \log\left(\frac{1-\epsilon}{6} n\right)$$

Proof. First, we notice that the expected number of fixed points of a permutation $\sigma \in S_n$ is 1. To realize this, let X_i be the indicator random variable for the i-th element in $\{1, \ldots, n\}$. We have $X_i = 1$ if $\sigma(i) = i$ and 0 otherwise. Then $X = \sum_{i=1}^n$ is the random variable which measures the number of fixed points. We have $\mathbb{E}(X_i) = \frac{1}{n}$, and so $\mathbb{E}(X) = \sum_{i=1}^n \frac{1}{n} = 1$.

Let $F(\sigma)$ denote the number of fixed points of the permutation σ. If σ is obtained from the identity by applying t random transpositions, then $F(\sigma)$ is at least as large as the number of cards that were touched by none of the transpositions (there could be more, as you could have a transposition and it's inverse).

Our shuffle chain determines transpositions by choosing pairs of cards independently and uniformly at random. Hence, after t shuffles, the number of untouched cards has the same distribution as the number R_{2t} of uncollected coupon types after $2t$ steps of the collector chain. By Lemma 7.13 from the book,

$$\mu := \mathbb{E}(R_{2t}) = n\left(1 - \frac{1}{n}\right)^{2t},$$

and $\text{Var}(R_{2t}) \leqslant \mu$. Let $A = \{\sigma : F(\sigma) > \mu/2\}$; that is, the number of permutations with fixed points greater than $\mu/2$. We will compare the probabilities of A under the uniform distribution π and $P^t(\text{id}, \cdot)$. First,

$$\pi(A) \leqslant \frac{2}{\mu},$$

by Markov's inequality. By Chebyshev's inequality,

$$P^t(\text{id}, A^c) \leqslant P\{R_{2t} \leqslant \mu/2\} \leqslant \frac{\mu}{(\mu/2)^2} = \frac{4}{\mu}.$$

By total variation distance, we get

$$||P^t(\text{id}, \cdot) - \pi||_{TV} \geqslant 1 - \frac{6}{\mu}.$$

We then want to find how small t must be so that $1 - 6/\mu > \epsilon$, or, equivalently,

$$n\left(1 - \frac{1}{n}\right)^{2t} = \mu < \frac{6}{1-\epsilon}.$$

Solving this and using $\log(1 + x) < x$ gets us

$$t \leqslant \frac{n-1}{2} \log\left(\frac{n(1-\epsilon)}{6}\right)$$

so that

$$t_{\text{mix}}(\epsilon) \geqslant \frac{n-1}{2} \log\left(\frac{n(1-\epsilon)}{6}\right).$$

Q.E.D

We now go through the coupling of the random transposition shuffle. At each time t, the shuffler chooses a card with label $X_t \in [x]$, and, independently, a position $Y_t \in [n]$; they then transposes the card labeled X_t with the card in position Y_t. If the card in position Y_t already has the label X_t, the deck is left unchanged. 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. We will see what happens in one step. Let a_t be the number of cards that occupy the same position in both σ_t and σ'_t. If the card labeled X_t is in the same position 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 the specified transposition breaks one alignment bu 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. This leads us to our proposition.

Proposition 28. Let τ be the time required for the two decks to coincide. Then, no matter the initial configuration of the deck, $\mathbb{E}(\tau) < \frac{\pi^2}{6}n^2$.

Proof. Let τ_i denote the time between the first time that $a_t \geqslant i-1$ and $a_t \geqslant i$. When t satisfies $a_t = i$, there are $n-i$ unaligned cards, and the probability of increasing the number of alignments is $(n-i)^2/n^2$ (they're independent, so multiply). In this situation, τ_{i+1} is a random variable with success probability given by $(n-i)^2/n^2$. We may conclude that under these circumstances,

$$\mathbb{E}(\tau_{i+1}|a_t = i) = n^2/(n-i)^2.$$

Now, we see that if $a_t \neq i$ for any t, then $\tau_{i+1} = 0$. Hence,

$$\mathbb{E}(\tau) < \sum \mathbb{E}(\tau_i) < n^2 \sum_{i=1}^{\infty} i^{-2}$$

thus giving the result. **Q.E.D**

We can then combine this with Corollary 5.5 from the book to find $t_{\text{mix}} \leqslant O(n^2)$.

We can also go through this using strong stationary times.

Proposition 29. 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 either R_t is unmarked or 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 stationary time for this chain.

Proof. Proof omitted for now. **Q.E.D**

Lemma. The stopping time τ defined in the prior proposition satisfies

$$\mathbb{E}(\tau) = 2n(\log(n) + O(1))$$

and

$$\text{Var}(\tau) = O(n^2).$$

Proof. We can decompose this into

$$\tau = \sum_{i=0}^{n-1} \tau_i$$

where τ_i is the number of steps after the k-th card is marked, up to and including when the $(k+1)$-st card is marked. Based on the rules in the prior proposition, we can see that this is a geometric random variable, with probability of success being $((k+1)(n-k))/n^2$. Hence, we get

$$\mathbb{E}(\tau) = \sum_{k=0}^{n-1} \frac{n^2}{(k+1)(n-k)}.$$

Using a partial fraction decomposition gives

$$\frac{1}{(k+1)(n-k)} = \frac{1}{n+1}\left(\frac{1}{k+1} + \frac{1}{n-k}\right).$$

Substituting this in gives

$$\frac{n^2}{n+1}\sum_{k=0}^{n-1}\left(\frac{1}{k+1} + \frac{1}{n-k}\right) \sim 2n(\log(n) + O(1)).$$

For the variance, we just use properties of the geometric random variable and bound above. **Q.E.D**

Corollary. For the random transposition chain on an n-card deck,

$$t_{\text{mix}} \leqslant (2 + o(1))n\log(n).$$

The final proposition is on the riffle shuffle.

Proposition 30. Fix $0 < \epsilon, \delta < 1$. Consider the riffle shuffling on an n-card deck. For sufficiently large n,

$$t_{\text{mix}}(\epsilon) \geqslant (1-\delta)\log_2(n).$$

Proof. There are at most 2^n possible states accessible in one step of the time-reversed chain (which, as we saw earlier, gives us equivalent bounds to the normal shuffle). Thus, $\log_2(\Delta)$, where Δ is the maximum out-degree defined in (7.1). The state space has size $n!$, and Stirling's formula shows that $\log_2 n! = (1+o(1))n\log_2(n)$. Using these estimates in (7.2) shows that for all $\delta > 0$, if n is sufficiently large, then the above holds. **Q.E.D**

Exercises

I talked to Graham about most of these exercises, or I did them by hand on the whiteboard.

8 Hitting Times

A preliminary before moving forward.

Definition. A function $h : \Omega \to \mathbb{R}$ is harmonic for P at a vertex x if

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

Definition. Given a Markov chain (X_t) with state space Ω, it is natural to define the hitting time τ_A of a subset $A \subset \Omega$ by

$$\tau_a := \min\{t \geqslant 0 : X_t \in A\}.$$

Remark. We will write τ_w for $\tau_{\{w\}}$.

Definition. We define the first return time as

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

Definition. For a Markov chain with stationary distribution π, let

$$t_\odot^a = \sum_{x \in \Omega} \mathbb{E}_a(\tau_x)\pi(x).$$

Lemma (Random Target Lemma)**.** For an irreducible Markov chain on the state space Ω with stationary distribution π, the target time $t_\odot^a$ does not depend on $a \in \Omega$.

Remark. Due to the prior lemma, we will use $t_\odot$ for all $t_\odot^a$, $a \in \Omega$.

Proof. Set $h_x(a) := \mathbb{E}_a(\tau_x)$. Observe that for all $x \neq a$,

$$h_x(a) = \sum_{y \in \Omega} \mathbb{E}_a(\tau_x | X_1 = y)P(a, y) = \sum_{y \in \Omega}(1 + h_a(y))P(a, y)$$

$$= \sum_{y \in \Omega} P(x, y) + \sum_{y \in \Omega} h_a(y)P(a, y) = 1 + \sum_{y \in \Omega} h_a(y)P(a, y)$$

$$= 1 + (Ph_a)(a).$$

Now, since $\mathbb{E}_a(\tau_a^+) = \pi(a)^{-1}$ (by (1.28))

$$(Ph_a)(a) = \frac{1}{\pi(a)} - 1.$$

Now, letting $h(a) := \sum_{x \in \Omega} h_x(a)\pi(x)$, combining the results above we have

$$(Ph)(a) = \sum_{x \in \Omega}(Ph_x)(a)\pi(x) = \sum_{x \neq a}(h_x(a) - 1)\pi(x) + \pi(a)\left(\frac{1}{\pi(a)} - 1\right).$$

Simplifying the right-hand side and using that $h_a(a) = 0$ yields

$$(Ph)(a) = h(a).$$

That is, h is harmonic. Applying Lemma 1.16 shows that h is a constant function. **Q.E.D**

Since $t_{\odot}$ does not depend on a, we get

$$t_{\odot} = \mathbb{E}_{\pi}(\tau_{\pi}).$$

Lemma. For an irreducible Markov chain with state space Ω and stationary distribution π,

$$t_{\text{hit}} \leqslant 2 \max_{w} \mathbb{E}_{\pi}(\tau_w).$$

Proof. For any $a, y \in \Omega$ we have

$$\mathbb{E}_a(\tau_y) \leqslant \mathbb{E}_a(\tau_{\pi}) + \mathbb{E}_{\pi}(\tau_y).$$

This is a sort of triangle inequality argument. By Lemma 10.1,

$$\mathbb{E}_a(\tau_{\pi}) = \mathbb{E}_{\pi}(\tau_{\pi}) \leqslant \max_{w} \mathbb{E}_{\pi}(\tau_w).$$

It is now clear that the first inequality gives us the desired result. **Q.E.D**

Corollary. For an irreducible transitive Markov chain,

$$t_{\text{hit}} \leqslant 2t_{\odot}.$$

Definition. The commute time between nodes a and b in a network is the expected time to move from a to b and then back to a. We denote by $\tau_{a,b}$ the random amount of time to transit from a to b and then back to a. That is,

$$\tau_{a,b} = \min\{t \geqslant \tau_b : X_t = a\},$$

where $X - 0 = a$. The commute time is then

$$t_{\leftrightarrow} := \mathbb{E}_a(\tau_{a,b}).$$

Note that the maximal commute time is

$$t_{\text{comm}} = \max_{a,b \in \Omega} t_{a \leftrightarrow b}.$$

Lemma. Let (X_t) be a Markov chain with transition matrix P. Suppose that for two probability distributions μ and ν on Ω, there is a stopping time on τ with $P_{\mu}\{\tau < \infty\} = 1$ and such that $P_{\mu}\{X_{\tau} = \cdot\} = \nu$. If ρ is the row vector

$$\rho(x) := \mathbb{E}_{\mu}\left(\sum_{t=0}^{\tau-1} \nVdash_{\{X_t = x\}}\right),$$

then $\rho P = \rho - \mu + \nu$. In particular, if $\mu = \nu$, then $\rho P = \rho$. Thus, if $\mu = \nu$ and $\mathbb{E}_{\mu}(\tau) < \infty$, then $\frac{\rho}{\mathbb{E}_{\mu}(\tau)}$ is a stationary distribution π for P.

Proof. Proof is left as an exercise. **Q.E.D**

In order to continue, we will need to go back a bit and talk about networks.

Definition. A network is a finite undirected connected graph G with vertex set V and edge set E, endowed additionally with non-negative numbers $\{c(e)\}$, called conductances, that are associated to the edges of G. We often write $c(x,y)$ for $c(\{x,y\})$. Notice that this is symmetric as well – $c(x,y) = c(y,x)$. The reciprocal $r(e) = 1/c(e)$ is called the resistance of the edge e.

Definition. A function W which is harmonic on $V \backslash \{a,z\}$ will be called a voltage. It can be shown that a voltage is completely determined by its boundary values $W(a)$ and $W(z)$.

Definition. Define the effective resistance between vertices a and z by

$$R(a \leftrightarrow z) := \frac{W(a) - W(z)}{||I||}.$$