Thus $\mathbf{Pr}(I = \perp) \leq 2/3$ as required.

(d) This is clearly true if $V = \emptyset$. If $V \neq \emptyset$ and $v = \max V \in I_0$ then, by induction

$$\mathbf{Pr}(I = I_0) = \frac{N_1}{N_1 + N_2}\phi\frac{N_1 + N_2}{N_1} = \phi$$

and similarly $\mathbf{Pr}(I = I_0) = \phi$ if $v \notin I_0$.

(e) Let $\mathcal{E}$ denote the event that some output of APPROXCOUNT is bad in the iteration that produces output. Then for $A \subseteq \Omega$,

$$\begin{aligned}\hat{\pi}(A) - \pi(A) &\leq \mathbf{Pr}(I \in A \mid \bar{\mathcal{E}}) + \mathbf{Pr}(\mathcal{E}) - \pi(A)\\ &\leq \frac{|A|}{|\Omega|} + \delta - \frac{|A|}{|\Omega|}\\ &\leq \delta.\end{aligned}$$

□

We have therefore shown that by running UGENX for *constant* expected number of times, we will with probability at least $1 - \delta$ output a randomly chosen independent set. The expected running time of UGEN is clearly as given in (1.11) which is small enough to make it a good sampler.

Having dealt with a specific example we see how to put the above ideas into a formal framework. Before doing this we enumerate some basic facts about Markov Chains.

1.3 Markov Chains

Throughout $\mathbb{N} = \{0, 1, 2, \ldots\}$, $\mathbb{N}_+ = \mathbb{N} \setminus \{0\}$, $\mathbb{Q}_+ = \{q \in \mathbb{Q} : q > 0\}$, and $[n] = \{1, 2, \ldots, n\}$ for $n \in \mathbb{N}_+$.

A Markov chain $\mathcal{M}$ on the finite state space Ω, with transition matrix P is a sequence of random variables X_t, $t = 0, 1, 2, \ldots$, which satisfy

$$\mathbf{Pr}(X_t = \sigma \mid X_{t-1} = \omega, X_{t-2}, \ldots, X_0) = P(\omega, \sigma) \qquad (t = 1, 2, \ldots),$$

We sometimes write P_σ^ω. The value of X_t is referred to as the *state* of $\mathcal{M}$ at *time* t.

Consider the digraph $D_{\mathcal{M}} = (\Omega, A)$ where $A = \{(\sigma, \omega) \in \Omega \times \Omega : P(\sigma, \omega) > 0\}$. We will by and large be concerned with chains that satisfy the following assumptions:

M1 The digraph $D_{\mathcal{M}}$ is strongly connected.

M2 $\gcd\{|C| : C \text{ is a directed cycle of } D_{\mathcal{M}}\} = 1$

Under these assumptions, $\mathcal{M}$ is *ergodic* and therefore has a unique stationary distribution π i.e.

$$\lim_{t\to\infty} \mathbf{Pr}(X_t = \omega \mid X_0 = \sigma) = \pi(\omega) \tag{1.12}$$

i.e. the limit does not depend on the starting state X_0. Furthermore, π is the unique left eigen-vector of P with eigenvalue 1 i.e. satisfying

$$P^T\pi = \pi. \tag{1.13}$$

Another useful fact is that if τ_σ denotes the expected number of steps between successive visits to state σ then

$$\tau_\sigma = \frac{1}{\pi(\sigma)}. \tag{1.14}$$

In most cases of interest, $\mathcal{M}$ is *reversible*, i.e.

$$Q(\omega,\sigma) = \pi(\omega)P(\omega,\sigma) = \pi(\sigma)P(\sigma,\omega) \qquad (\forall\omega,\sigma\in\Omega). \tag{1.15}$$

The central role of reversible chains in applications rests on the fact that π can be deduced from (1.15). If $\mu : \Omega \longrightarrow \mathbb{R}$ satisfies (1.15), then it determines π up to normalization. Indeed, if (1.15) holds and $\sum_{\omega\in\Omega}\pi(\omega) = 1$ then

$$\sum_{\omega\in\Omega}\pi(\omega)P(\omega,\sigma) = \sum_{\omega\in\Omega}\pi(\sigma)P(\sigma,\omega) = \pi(\sigma)$$

which proves that π is a left eigenvector with eigenvalue 1.

In fact, we often *design* the chain to satisfy (1.15). Without reversibility, there is no apparent method of determining π, other than to explicitly construct the transition matrix, an exponential time (and space) computation in our setting.

As a canonical example of a reversible chain we have a *random walk* on a graph. A random walk on the undirected graph $G = (V, E)$ is a Markov chain with state space V associated with a particle that moves from vertex to vertex according to the following rule: the probability of a transition from vertex i, of degree d_i, to vertex j is $\frac{1}{d_i}$ if $\{i, j\} \in E$, and 0 otherwise. Its stationary distribution is given by

$$\pi(v) = \frac{d_v}{2|E|} \qquad\qquad v \in V. \tag{1.16}$$

To see this note that $Q(v, w) = Q(w, v)$ if v, w are not adjacent and otherwise

$$Q(v,w) = \frac{1}{2|E|} = Q(w,v),$$

verifying the *detailed balance* equations (1.15).

Note that if G is a regular graph then the steady state is uniform over V.

If G is bipartite then the walk as described is not ergodic. This is because all cycles are of even length. This is usually handled by adding d_v loops to vertex v for each vertex v. (Each loop counts as a single exit from v.) The net effect of this is to make the particle stay put with probability $\frac{1}{2}$ at each step. The steady state is unaffected. The chain is now *lazy*.

A chain is lazy if $P(\omega,\omega) \geq \frac{1}{2}$ for all $\omega \in \Omega$.

If $p_0(\omega) = \mathbf{Pr}(X_0 = \omega)$, then $p_t(\sigma) = \sum_\omega p_0(\omega)P^t(\omega,\sigma)$ is the distribution at time t. As a measure of convergence, the natural choice in this context is variation distance.

The *mixing time* of the chain is then

$$\tau(\varepsilon) = \max_{p_0} \min_t \{D_{\mathrm{tv}}(p_t,\pi) \leq \varepsilon\},$$

and it is easy to show that the maximum occurs when $X_0 = \omega_0$, with probability one, for some state ω_0. This is because $D_{\mathrm{tv}}(p_t,\pi)$ is a convex function of p_0 and so the maximum of $D_{\mathrm{tv}}(p_t,\pi)$ occurs at an extreme point of the set of probabilities p_0.

I think this should be moved to the next chapter

We now provide a simple lemma which indicates that variation distance $D_{\mathrm{tv}}(p_t,\pi)$ goes to zero exponentially. We define several related quantities: $p_t^{(i)}$ denotes the t-fold distribution, conditional on $X_0 = i$.

$$d_i(t) = D_{\mathrm{tv}}(p_t^{(i)},\pi),\ d(t) = \max_i d_i(t),\ \bar{d}(t) = \max_{i,j} D_{\mathrm{tv}}(p_t^{(i)},p_t^{(j)}).$$

Lemma 1.3.1 *For all $s,t \geq 0$,*

(a) $\bar{d}(s+t) \leq \bar{d}(s)\bar{d}(t)$.

(b) $d(s+t) \leq 2d(s)d(t)$.

(c) $d(s) \leq 2\bar{d}(s)$.

(d) $d(s) \leq d(t)$ *for* $s \leq t$.

Proof We will use the characterisation of variation distance as

$$D_{\mathrm{tv}}(\mu_1,\mu_2) = \min \mathbf{Pr}(X_1 \neq X_2) \tag{1.17}$$

where the minimum is taken over pairs of random variables X_1, X_2 such that X_i has distribution $\mu_i, i = 1,2$.

Fix states i_1, i_2 and times s,t and let Y^1, Y^2 denote the chains started at i_1, i_2 respectively. By (1.17) we can construct a joint distribution for (Y_s^1, Y_s^2) such that

$$\mathbf{Pr}(Y_s^1 \neq Y_s^2) = D_{\mathrm{tv}}(p_s^{(i_1)}, p_s^{(i_2)}) \leq \bar{d}(s).$$

Now for each pair j_1, j_2 we can use (1.17) to construct a joint distribution for (Y^1_{s+t}, Y^2_{s+t}) such that

$$\mathbf{Pr}(Y^1_{s+t} \neq Y^2_{s+t} \mid Y^1_s = j_1, Y^2_s = j_2) = D_{\mathrm{tv}}(p^{(j_1)}_t, p^{(j_2)}_t).$$

The RHS is 0 if $j_1 = j_2$ and otherwise at most $\bar{d}(t)$. So, unconditionally,

$$\mathbf{Pr}(Y^1_{s+t} \neq Y^2_{s+t}) \leq \bar{d}(s)\bar{d}(t)$$

and (1.17) establishes part (a) of the lemma.

For part (b), the same argument, with Y^2 now being the stationary chain shows

$$d(s+t) \leq d(s)\bar{d}(t) \tag{1.18}$$

and so (b) will follow from (c), which follows from the triangular inequality for variation distance. Finally note that (d) follows from (1.18). □

We will for the most part use carefully defined Markov chains as our good samplers. As an example, we now define a simple chain with state space Ω equal to the collection of independent sets of a graph G. The chain is ergodic and its steady state is uniform over Ω. So, running the chain for sufficiently long will produce a near uniformly chosen independent set, see (1.12). Unfortunately, this chain does not have a small enough mixing time for this to qualify as a good sampler, unless $\Delta(G) \leq 4$.

We define the chain as follows: suppose $X_t = I$. Then we choose a vertex v of G uniformly at random. If $v \in I$ then we put $X_{t+1} = I \setminus \{v\}$. If $v \notin I$ and $I \cup \{v\}$ is an indepedent set then we put $X_{t+1} = I \cup \{v\}$. Otherwise we let $X_{t+1} = X_t = I$. Thus the transition matrix can be described as follows: $n = |V|$ and I, J are independent sets of G.

$$P(I,J) = \begin{cases} \frac{1}{n} & |I \Delta J| = 1 \\ 0 & \text{otherwise} \end{cases}$$

Here $I \Delta J$ denotes the symmetric difference $(I \setminus J) \cup (J \setminus I)$.

The chain satisfies M1 and M2: In $D_{\mathcal{M}}$ every vertex can reach and is reachable from $\emptyset$, implying M1 holds. Also, $D_{\mathcal{M}}$ contains loops unless G has no edges. In both cases M2 holds trivially.

Note finally that $P(I,J) = P(J,I)$ and so (1.15) holds with $\pi(I) = \frac{1}{|\Omega|}$. Thus the chain is reversible and the steady state is uniform.

1.4 A formal computational framework

The sample spaces we have in mind are sets of combinatorial objects. However, in order to discuss the computational complexity of generation, it is necessary to consider a sequence of instances of increasing size. We therefore work within the following formal

CHAPTER 3

Markov Chain Monte Carlo: Metropolis and Glauber Chains

3.1. Introduction

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

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

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

For some graphs, e.g. trees, there are simple recursive methods for generating a random proper coloring (see Example 14.12). However, for other graphs it can be challenging to directly construct a random sample. One approach is to use Markov chains to sample: suppose that (X_t) is a chain with state space $\mathcal{X}$ and with stationary distribution uniform on $\mathcal{X}$ (in Section 3.3, we will construct one such chain). By the Convergence Theorem (Theorem 4.9, whose proof we have not yet given but have often foreshadowed), X_t is approximately uniformly distributed when t is large.

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

3.2. Metropolis Chains

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

3.2.1. Symmetric base chain. Suppose that Ψ is a symmetric transition matrix. In this case, Ψ is reversible with respect to the uniform distribution on $\mathcal{X}$.

We now show how to modify transitions made according to Ψ to obtain a chain with stationary distribution π, given an arbitrary probability distribution π on $\mathcal{X}$.

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

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

By Proposition 1.20, the transition matrix P has stationary distribution π if

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

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

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

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

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

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

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

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

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

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

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

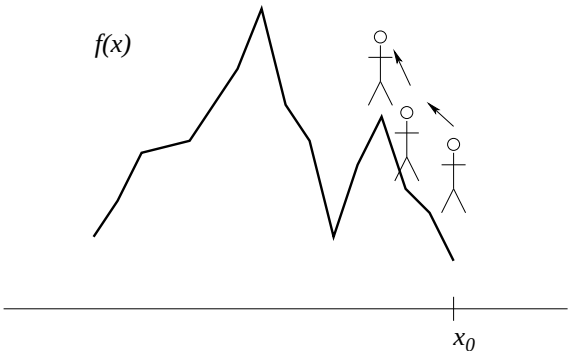

FIGURE 3.1. A hill climb algorithm may become trapped at a local maximum.

A ***hill climb*** is an algorithm which attempts to locate the maximum values of f as follows: when at x, if there is at least one neighbor y of x satisfying $f(y) > f(x)$, move to a neighbor with the largest value of f. The climber may become stranded at local maxima — see Figure 3.1.

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

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

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

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

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

Define

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

Then

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

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

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

y with probability

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

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

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

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

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

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

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

3.3. Glauber Dynamics

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

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

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

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

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

which is a useful identity.

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

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

PROPOSITION 4.5. *Let μ and ν be two probability distributions on $\mathcal{X}$. Then the total variation distance between them satisfies*

$$\|\mu - \nu\|_{\mathrm{TV}} \;=\; \frac{1}{2} \sup \left\{ \sum_{x \in \mathcal{X}} f(x)\mu(x) - \sum_{x \in \mathcal{X}} f(x)\nu(x) \;:\; \max_{x \in \mathcal{X}} |f(x)| \leq 1 \right\}. \tag{4.7}$$

PROOF. If $\max_{x \in \mathcal{X}} |f(x)| \leq 1$, then

$$\frac{1}{2} \left| \sum_{x \in \mathcal{X}} f(x)\mu(x) - \sum_{x \in \mathcal{X}} f(x)\nu(x) \right| \leq \frac{1}{2} \sum_{x \in \mathcal{X}} |\mu(x) - \nu(x)| = \|\mu - \nu\|_{\mathrm{TV}} \,.$$

Thus, the right-hand side of (4.7) is at most $\|\mu - \nu\|_{\mathrm{TV}}$.

For the other direction, define

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

Then

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

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

4.2. Coupling and Total Variation Distance

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

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

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

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

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

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

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

and

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

Conversely, given a probability distribution q on the product space $\mathcal{X} \times \mathcal{X}$ which satisfies

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

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

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

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

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

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

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

PROPOSITION 4.7. *Let μ and ν be two probability distributions on $\mathcal{X}$. Then*

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

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

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

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

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

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

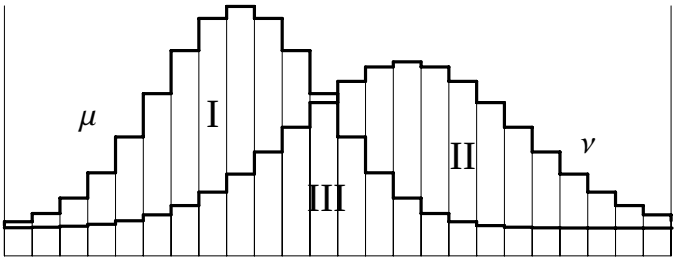

FIGURE 4.2. Since each of regions I and II has area $\|\mu - \nu\|_{\mathrm{TV}}$ and μ and ν are probability measures, region III has area $1 - \|\mu - \nu\|_{\mathrm{TV}}$.

It will suffice to construct a coupling for which $\mathbf{P}\{X \neq Y\}$ is exactly equal to $\|\mu - \nu\|_{\mathrm{TV}}$. We will do so by forcing X and Y to be equal as often as they possibly can be. Consider Figure 4.2. Region III, bounded by $\mu(x) \wedge \nu(x) = \min\{\mu(x), \nu(x)\}$, can be seen as the overlap between the two distributions. Informally, our coupling proceeds by choosing a point in the union of regions I and III, and setting X to be the x-coordinate of this point. If the point is in III, we set $Y = X$ and if it is in I, then we choose independently a point at random from region II, and set Y to be the x-coordinate of the newly selected point. In the second scenario, $X \neq Y$, since the two regions are disjoint.

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

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

Write

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

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

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

By equation (4.5) and the immediately preceding equation,

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

Flip a coin with probability of heads equal to p.

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

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

and set $X = Y = Z$.

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

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

and independently choose Y according to the probability distribution

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

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

Clearly,

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

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

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

■

4.3. The Convergence Theorem

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

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

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

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

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

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

for all $x, y \in \mathcal{X}$. Let $\theta = 1 - \delta$. The equation

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

defines a stochastic matrix Q.

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

Next, we use induction to demonstrate that

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

i.e., we change the ith component from x_i to y_i. Note that some of the edges may be loops (if $x_i = y_i$). To compute $\bar{\varrho}$, fix attention on a particular (oriented) edge

$$t = (w, w') = \big((w_0, \ldots, w_i, \ldots w_{n-1}), (w_0, \ldots, w'_i, \ldots w_{n-1})\big),$$

and consider the number of canonical paths γ_{xy} that include t. The number of possible choices for x is 2^i, as the final $n - i$ positions are determined by $x_j = w_j$, for $j \geq i$; and by a similar argument the number of possible choices for y is 2^{n-i-1}. Thus the total number of canonical paths using a particular edge t is 2^{n-1}; furthermore, $Q(w, w') = \pi(w)P(w, w') \geq 2^{-n}(2n)^{-1}$, and the length of every canonical path is exactly n. Plugging all these bounds into the definition of $\bar{\rho}$ yields $\bar{\rho} \leq n^2$. Thus, by Theorem 2.2.4, the mixing time of $\mathcal{W}_n$ is $\tau(\varepsilon) \leq n^2(n \ln q + \ln \varepsilon^{-1})$.

2.2.4 Comparison Theorems

2.2.5 Decomposition Theorem

2.3 Coupling

A *coupling* $\mathcal{C}(\mathcal{M})$ for $\mathcal{M}$ is a stochastic process (X_t, Y_t) on $\Omega \times \Omega$ such that each of X_t, Y_t is marginally a copy of $\mathcal{M}$,

$$\begin{aligned} \mathbf{Pr}(X_t = \sigma_1 \mid X_{t-1} = \omega_1) &= P(\omega_1, \sigma_1), \\ \mathbf{Pr}(Y_t = \sigma_2 \mid Y_{t-1} = \omega_2) &= P(\omega_2, \sigma_2), \end{aligned} \qquad (\forall t > 0). \tag{2.18}$$

The following simple but powerful inequality then follows easily from these definitions.

Lemma 2.3.1 (Coupling Lemma) *Let X_t, Y_t be a coupling for $\mathcal{M}$ such that Y_0 has the stationary distribution π. Then, if X_t has distribution p_t,*

$$D_{\mathrm{tv}}(p_t, \pi) \leq \mathbf{Pr}(X_t \neq Y_t). \tag{2.19}$$

Proof Suppose $A_t \subseteq \Omega$ maximizes in (1.3). Then, since Y_t has distribution π,

$$\begin{aligned} D_{\mathrm{tv}}(p_t, \pi) &= \mathbf{Pr}(X_t \in A_t) - \mathbf{Pr}(Y_t \in A_t) \\ &\leq \mathbf{Pr}(X_t \in A_t, Y_t \notin A_t) \\ &\leq \mathbf{Pr}(X_t \neq Y_t). \end{aligned}$$

□

It is important to remember that the Markov chain Y_t is simply a proof construct, and X_t the chain we actually observe. We also require that $X_t = Y_t$ implies $X_{t+1} = Y_{t+1}$,

since this makes the right side of (2.19) nonincreasing. Then the earliest epoch T at which $X_T = Y_T$ is called *coalescence*, making T a random variable. A *successful coupling* is such that $\lim_{t\longrightarrow\infty} \mathbf{Pr}(X_t \neq Y_t) = 0$. Clearly we are only interested in successful couplings.

As an example consider our random walk on the cube Q_n. We can define a coupling as follows: Given (X_t, Y_t) we

(a) Choose i uniformly at random from $[n]$.

(b) Put $X_{t+1,j} = X_{t,j}$ and $Y_{t+1,j} = Y_{t,j}$ for $j \neq i$.

(c) If $X_{t,i} = Y_{t,i}$ then

$$X_{t+1,i} = Y_{t+1,i} = \begin{cases} X_{t,i} & \text{prob } \frac{1}{2} \\ 1 - X_{t,i} & \text{prob } \frac{1}{2} \end{cases}$$

(d) otherwise

$$(X_{t+1,i}, Y_{t+1,i}) = \begin{cases} (X_{t,i}, 1 - Y_{t,i}) & \text{prob } \frac{1}{2} \\ (1 - X_{t,i}, Y_{t,i}) & \text{prob } \frac{1}{2} \end{cases}$$

It should hopefully be clear that this is a coupling i.e. the marginals are correct and $X_t = Y_t$ implies $X_{t+1} = Y_{t+1}$.

Now let $I_t = \{j : \ i \text{ is chosen in (a) of steps } 1, 2, \ldots, t$. Then $I_t = [n]$ implies that $X_\tau = Y_\tau$ for $\tau \geq t$. So

$$\begin{aligned} \mathbf{Pr}(X_t \neq Y_t) &\leq \mathbf{Pr}(I_t \neq [n]) \\ &= \mathbf{Pr}(\bar{I}_t \neq \emptyset) \\ &\leq \mathbf{E}(|\bar{I}_t|) \\ &= n\left(1 - \tfrac{1}{n}\right)^t. \end{aligned}$$

So if $t = n(\log n + \log \epsilon^{-1})$ we have $d_{TV}(p_t, \pi) \leq \epsilon$.

A coupling is a *Markovian coupling* if the process $\mathcal{C}(\mathcal{M})$ is a Markov chain on $\Omega \times \Omega$. There always exists a *maximal coupling*, which gives equality in (2.19). This maximal coupling is in general non-Markovian, and is seemingly not constructible without knowing p_t $(t = 1, 2, \ldots)$. But coupling has little algorithmic value if we already know p_t. More generally, it seems difficult to prove mixing properties of non-Markovian couplings in our setting. Therefore we restrict attention to Markovian couplings, at the (probable) cost of sacrificing equality in (2.19).

Let $\mathcal{C}(\mathcal{M})$ be a Markovian coupling, with Q its transition matrix, i.e. the probability of a joint transition from (ω_1, ω_2) to (σ_1, σ_2) is $Q^{\omega_1\omega_2}_{\sigma_1\sigma_2}$. The precise conditions required of

Q are then

$$Q^{\omega\,\omega}_{\sigma_1\sigma_2} \neq 0 \qquad \text{implies} \quad \sigma_1 = \sigma_2 \qquad (\forall \omega \in \Omega), \tag{2.20}$$

$$\sum_{\sigma_2 \in \Omega} Q^{\omega_1\omega_2}_{\sigma_1\sigma_2} = P^{\omega_1}_{\sigma_1} \quad (\forall \omega_2 \in \Omega), \quad \sum_{\sigma_1 \in \Omega} Q^{\omega_1\omega_2}_{\sigma_1\sigma_2} = P^{\omega_2}_{\sigma_2} \quad (\forall \omega_1 \in \Omega). \tag{2.21}$$

Here (2.20) implies equality after coalescence, and (2.21) implies the marginals are copies of $\mathcal{M}$. Our goal is to design Q so that $\mathbf{Pr}(X_t \neq Y_t)$ quickly becomes small. We need only specify Q to satisfy (2.21) for $\omega_1 \neq \omega_2$. The other entries are completely determined by (2.20) and (2.21).

In general, to prove rapid mixing using coupling, it is usual to map $\mathcal{C}(\mathcal{M})$ to a process on $\mathbb{N}$ by defining a function $\psi : \Omega \times \Omega \longrightarrow \mathbb{N}$ such that $\psi(\omega_1, \omega_2) = 0$ implies $\omega_1 = \omega_2$. We call this a *proximity function*. Then $\mathbf{Pr}(X_t \neq Y_t) \leq \mathbf{E}(\psi(X_t, Y_t))$, by Markov's inequality, and we need only show that $\mathbf{E}(\psi(X_t, Y_t))$ converges quickly to zero.

2.4 Path coupling

A major difficulty with coupling is that we are obliged to specify it, and show improvement in the proximity function, for every pair of states. The idea of *path coupling*, where applicable, can be a major saving in this respect. We describe the approach below.

As a simple example of this approach consider a Markov chain where $\Omega \subseteq S^m$ for some set S and positive integer m. Suppose also that if $\omega, \sigma \in \Omega$ and $h(\omega, \sigma) = d$ (Hamming distance) then there exists a sequence $\omega = x_0, x_1, \ldots, x_d = \sigma$ of members of Ω such that (i) $\{x_0, x_1, \ldots, x_d\} \subseteq \Omega$, (ii) $h(x_i, x_{i+1}) = 1$, $i = 0, 1, \ldots, d-1$ and (iii) $P(x_i, x_{i+1}) > 0$.

Now suppose we define a coupling of the chains (X_t, Y_t) *only* for the case where $h(X_t, Y_t) = 1$. Suppose then that

$$\mathbf{E}(h(X_{t+1}, Y_{t+1}) \mid h(X_t, Y_t) = 1) \leq \beta \tag{2.22}$$

for some $\beta < 1$. Then

$$\mathbf{E}(h(X_{t+1}, Y_{t+1})) \leq \beta h(X_t, Y_t), \tag{2.23}$$

in all cases. It then follows that

$$d_{TV}(p_t, \pi) \leq \mathbf{Pr}(X_t \neq Y_t) \leq n\beta^t.$$

Equation (2.23) is shown by choosing a sequence $X_t = Z_0, Z_1, \ldots, Z_d = Y_t, d = h(X_t, Y_t)$ $Z_0, Z_1, \ldots, Z_d$ satisfy (i),(ii),(iii) above. Then we can couple Z_i and Z_{i+1}, $1 \leq i < d$ so that $X_{t+1} = Z'_0, Z'_1, \ldots, Z'_d = Y_{t+1}$ and (i) $\mathbf{Pr}(Z'_i = \sigma \mid Z_i = \omega) = P(\omega, \sigma)$ and (ii)

$\mathbf{E}(h(Z_i', Z_{i+1}')) \le \beta$. Therefore

$$\mathbf{E}(h(X_{t+1}, Y_{t+1})) \le \sum_{i=1}^{d} \mathbf{E}(h(Z_i', Z_{i+1}')) \le \beta d$$

and (2.23) follows.

As an example, let $G = (V, E)$ be a graph with maximum degree Δ and let $k \ge 2\Delta + 1$ be an integer. Let Ω_k be the set of proper k- vertex colourings of G i.e. $\{c : V \to [k]\}$ such that $(v, w) \in E$ implies $c(v) \ne c(w)$. We describe a chain which provides a good sampler for the uniform distribution over Ω_k. We let $\Omega = V^k$ be all k-colourings, including improper ones and describe a chain on Ω for which only proper colourings have a positive steady state probability.

To describe a general step of the chain asume $X_t \in \Omega$. Then

Step 1 Choose w uniformly from V and x uniformly from $[k]$.

Step 2 Let $X_{t+1}(v) = X_t(v)$ for $v \in V \setminus \{w\}$.

Step 3 If no neighbour of w in G has colour x then put $X_{t+1}(w) = x$, otherwise put $X_{t+1}(w) = x$.

Note that $P(\omega, \sigma) = P(\sigma, \omega) = \frac{1}{nk}$ for two proper colourings which can be obtained from each other by a single move of the chain. It follows from (1.15) that the steady state is uniform over Ω_k.

We first describe a coupling which is extremely simple but needs $k > 3\Delta$ in order for (2.22) to be satisfied. Let $h(X_t, Y_t) = 1$ and let v_0 be the unique vertex of V such that $X_t(v) \ne Y_t(v)$. In our coupling we choose w, x as in Step 1 and try to colour w with x in both chains.

We claim that

$$\mathbf{E}(h(X_{t+1}, Y_{t+1}) \le 1 - \frac{1}{n}\left(1 - \frac{\Delta}{k}\right) + \frac{\Delta}{n}\frac{2}{k} = 1 - \frac{k - 3\Delta}{kn}. \tag{2.24}$$

and so we can take $\beta \le 1 - \frac{1}{kn}$ in (2.23) if $k > 3\Delta$.

The term $\frac{1}{n}\left(1 - \frac{\Delta}{k}\right)$ in (2.24) lower bounds the probability that $w = v_0$ and that x is not used in the neighbourhood of v_0. In which case we will have $X_{t+1} = Y_{t+1}$. Next let $c_X \ne c_Y$ be the colours of v_0 in X_t, Y_t respectively. The term $\frac{\Delta}{n}\frac{2}{k}$ in (2.24) is an upper bound for the probability that w is in the neighbourhood of v_0 and $x \in \{c_X, c_Y\}$ and in which case we might have $h(X_{t+1}, Y_{t+1}) = 2$. In all other cases we find that $h(X_{t+1}, Y_{t+1}) = h(X_t, Y_t) = 1$.

A better coupling gives the desired result. We proceed as above except for the case where w is a neighbour of v_0 and $x \in \{c_X, c_Y\}$. In this case with probability $\frac{1}{2}$ we try to colour w with c_X in X_t and colour w with c_Y in Y_t, and fail in both cases. With probability $\frac{1}{2}$ we try to colour w with c_Y in X_t and colour w with c_X in Y_t, in which case the hamming distance may increase by one. Thus for this coupling we have

$$\mathbf{E}(h(X_{t+1}, Y_{t+1}) \leq 1 - \frac{1}{n}\left(1 - \frac{\Delta}{k}\right) + \frac{1}{2}\frac{\Delta}{n}\frac{2}{k} = 1 - \frac{k - 2\Delta}{kn}$$

and we can take $\beta \leq 1 - \frac{1}{kn}$ in (2.23) if $k > 2\Delta$.

We now give a more general framework for the definition of path coupling. Recall that a *quasi-metric* satisfies the conditions for a metric except possibly the symmetry condition. Any metric is a quasi-metric, but a simple example of a quasi-metric which is not a metric is directed edge distance in a digraph.

Suppose we have a relation $S \subseteq \Omega \times \Omega$ such that S has transitive closure $\Omega \times \Omega$, and suppose that we have a proximity function defined for all pairs in S, i.e. $\psi : S \longrightarrow \mathbb{N}$. Then we may lift ψ to a quasi-metric $\phi(\omega, \omega')$ on Ω as follows. For each pair $(\omega, \omega') \in \Omega \times \Omega$, consider the set $\mathcal{P}(\omega, \omega')$ of all sequences

$$\omega = \omega_1, \omega_2, \ldots, \omega_{r-1}, \omega_r = \omega' \quad \text{with} \quad (\omega_i, \omega_{i+1}) \in S \quad (i = 1, \ldots, r-1). \tag{2.25}$$

Then we set

$$\phi(\omega, \omega') \;=\; \min_{\mathcal{P}(\omega,\omega')} \sum_{i=1}^{r-1} \psi(\omega_i, \omega_{i+1}). \tag{2.26}$$

It is easy to prove that ϕ is a quasi-metric. We call a sequence minimizing (2.26) *geodesic*. We now show that, without any real loss, we may define the (Markovian) coupling only on pairs in S. Such a coupling is a called a path coupling. We give a detailed development below. Clearly $S = \Omega \times \Omega$ is always a relation whose transitive closure is $\Omega \times \Omega$, but path coupling is only useful when we can define a suitable S which is "much smaller" than $\Omega \times \Omega$. A relation of particular interest is $\mathcal{R}_\sigma$ from Section 1.4, but this is not always the best choice.

As in Section 2.3, we use σ (or σ_i) to denote a state obtained by performing a single transition of the chain from the state ω (or ω_i). Let P^ω_σ denote the probability of a transition from state ω to state σ in the Markov chain, and let $Q^{\omega\omega'}_{\sigma\sigma'}$ denote the probability of a joint transition from (ω, ω') to (σ, σ'), where $(\omega, \omega') \in S$, as specified by the path coupling. Since this coupling has the correct marginals, we have

$$\sum_{\sigma' \in \Omega} Q^{\omega\omega'}_{\sigma\sigma'} = P^\omega_\sigma, \qquad \sum_{\sigma \in \Omega} Q^{\omega\omega'}_{\sigma\sigma'} = P^{\omega'}_{\sigma'} \qquad (\forall (\omega, \omega') \in S). \tag{2.27}$$

We extend this to all pairs $(\omega, \omega') \in \Omega \times \Omega$, as follows. For each pair, fix a sequence $(\omega_1, \omega_2, \ldots, \omega_r) \in \mathcal{P}(\omega, \omega')$. We do not assume the sequence is geodesic here, or indeed

the existence of any proximity function, but this is our eventual purpose. The implied global coupling $\bar{Q}^{\omega_1\omega_r}_{\sigma_1\sigma_r}$ is then defined along this sequence by successively conditioning on the previous choice. Using (2.27), this can be written explicitly as

$$\bar{Q}^{\omega_1\omega_r}_{\sigma_1\sigma_r} = \sum_{\sigma_2\in\Omega}\sum_{\sigma_3\in\Omega}\cdots\sum_{\sigma_{r-1}\in\Omega} Q^{\omega_1\omega_2}_{\sigma_1\sigma_2}\frac{Q^{\omega_2\omega_3}_{\sigma_2\sigma_3}}{P^{\omega_2}_{\sigma_2}}\cdots\frac{Q^{\omega_{r-1}\omega_r}_{\sigma_{r-1}\sigma_r}}{P^{\omega_{r-1}}_{\sigma_{r-1}}}. \tag{2.28}$$

Summing (2.28) over σ_r or σ_1, and again applying (2.27), causes the right side to successively simplify, giving

$$\sum_{\sigma_r\in\Omega}\bar{Q}^{\omega_1\omega_r}_{\sigma_1\sigma_r} = P^{\omega_1}_{\sigma_1}\quad(\forall\omega_r\in\Omega),\qquad \sum_{\sigma_1\in\Omega}\bar{Q}^{\omega_1\omega_r}_{\sigma_1\sigma_r} = P^{\omega_r}_{\sigma_r}\quad(\forall\omega_1\in\Omega). \tag{2.29}$$

Hence the global coupling satisfies (2.21), as we would anticipate from the properties of conditional probabilities.

Now suppose the global coupling is determined by geodesic sequences. We bound the expected value of $\phi(\sigma_1,\sigma_r)$. This is

$$\begin{aligned}
\mathbf{E}(\phi(\sigma_1,\sigma_r)) &= \sum_{\sigma_1}\cdots\sum_{\sigma_r}\phi(\sigma_1,\sigma_r)\frac{Q^{\omega_1\omega_2}_{\sigma_1\sigma_2}Q^{\omega_2\omega_3}_{\sigma_2\sigma_3}\cdots Q^{\omega_{r-1}\omega_r}_{\sigma_{r-1}\sigma_r}}{P^{\omega_2}_{\sigma_2}\cdots P^{\omega_{r-1}}_{\sigma_{r-1}}}\\
&\le \sum_{\sigma_1}\cdots\sum_{\sigma_r}\sum_{i=1}^{r-1}\phi(\sigma_i,\sigma_{i+1})\frac{Q^{\omega_1\omega_2}_{\sigma_1\sigma_2}Q^{\omega_2\omega_3}_{\sigma_2\sigma_3}\cdots Q^{\omega_{r-1}\omega_r}_{\sigma_{r-1}\sigma_r}}{P^{\omega_2}_{\sigma_2}\cdots P^{\omega_{r-1}}_{\sigma_{r-1}}}\\
&= \sum_{i=1}^{r-1}\sum_{\sigma_1}\cdots\sum_{\sigma_r}\phi(\sigma_i,\sigma_{i+1})\frac{Q^{\omega_1\omega_2}_{\sigma_1\sigma_2}Q^{\omega_2\omega_3}_{\sigma_2\sigma_3}\cdots Q^{\omega_{r-1}\omega_r}_{\sigma_{r-1}\sigma_r}}{P^{\omega_2}_{\sigma_2}\cdots P^{\omega_{r-1}}_{\sigma_{r-1}}}\\
&= \sum_{i=1}^{r-1}\sum_{\sigma_i}\sum_{\sigma_{i+1}}\phi(\sigma_i,\sigma_{i+1})Q^{\omega_i\omega_{i+1}}_{\sigma_i\sigma_{i+1}},
\end{aligned} \tag{2.30}$$

where we have used the triangle inequality for a quasi-metric and the same observation as that leading from (2.28) to (2.29).

Suppose we can find $\beta\le 1$, such that, for all $(\omega,\omega')\in S$,

$$\mathbf{E}(\phi(\sigma,\sigma')) \;=\; \sum_{\sigma}\sum_{\sigma'}\phi(\sigma,\sigma')\,Q^{\omega\omega'}_{\sigma\sigma'} \;\le\; \beta\,\phi(\omega,\omega'). \tag{2.31}$$

Then, from (2.30), (2.31) and (2.26) we have

$$\mathbf{E}(\phi(\sigma_1,\sigma_r)) \;\le\; \sum_{i=1}^{r-1}\beta\,\phi(\omega_i,\omega_{i+1}) \;=\; \beta\sum_{i=1}^{r-1}\phi(\omega_i,\omega_{i+1}) \;=\; \beta\,\phi(\omega_1,\omega_r). \tag{2.32}$$

Thus we can show (2.31) for every pair, merely by showing that this holds for all pairs in S. To apply path coupling to a particular problem, we must find a relation S and

proximity function ψ so that this is possible. In particular we need $\phi(\omega, \omega')$ for $(\omega, \omega') \in S$ to be easily deducible from ψ.

Suppose that Ω has *diameter* D, i.e. $\phi(\omega, \omega') \leq D$ for all $\omega, \omega' \in \Omega$. Then, $\mathbf{Pr}(X_t \neq Y_t) \leq \beta^t D$ and so if $\beta < 1$ we have, since $\log \beta^{-1} \geq 1 - \beta$,

$$D_{\mathrm{tv}}(p_t, \pi) \leq \varepsilon \quad \text{for} \quad t \geq \log(D\varepsilon^{-1})/(1-\beta). \tag{2.33}$$

This bound is polynomial even when D is exponential in the problem size. It is also possible to prove a bound when $\beta = 1$, provided we know the quasi-metric cannot "get stuck". Specifically, we need an $\alpha > 0$ (inversely polynomial in the problem size) such that, in the above notation,

$$\mathbf{Pr}(\phi(\sigma, \sigma') \neq \phi(\omega, \omega')) \geq \alpha \qquad (\forall \omega, \omega' \in \Omega). \tag{2.34}$$

Observe that it is not sufficient simply to establish (2.34) for pairs in S. However, the structure of the path coupling can usually help in proving it. In this case, we can show that

$$D_{\mathrm{tv}}(p_t, \pi) \leq \varepsilon \quad \text{for} \quad t \geq \lceil eD^2/\alpha \rceil \lceil \ln(\varepsilon^{-1}) \rceil. \tag{2.35}$$

This is most easily shown using a martingale argument. Here we need D to be polynomial in the problem size.

Consider a sequence $(\omega_0, \omega_0'), (\omega_1, \omega_1') \ldots, (\omega_t, \omega_t')$ and define the random time $T^{\omega,\omega'} = \min\{t : \phi(\omega_t, \omega_t') = 0\}$, assuming that $\omega_0 = \omega, \omega_0' = \omega'$. We prove that

$$\mathbf{E}(T^{\omega,\omega'}) \leq D^2/\alpha. \tag{2.36}$$

Let

$$Z(t) = \phi(\omega_t, \omega_t')^2 - 2D\phi(\omega_t, \omega_t') - \alpha t$$

and let

$$\delta(t) = \phi(\omega_{t+1}, \omega_{t+1}') - \phi(\omega_t, \omega_t').$$

Then

$$\mathbf{E}(Z(t+1) \mid Z(0), Z(1), \ldots, Z(t)) - Z(t) = \\ 2(\phi(\omega_t, \omega_t') - D)\mathbf{E}(\delta(t) \mid \omega_t, \omega_t') + (\mathbf{E}(\delta(t)^2 \mid \omega_t, \omega_t') - \alpha) \geq 0.$$

Hence $Z(t)$ is a submartingale. The stopping time $T^{\omega,\omega'}$ has finite expectation and $|Z(t+1) - Z(t)| \leq D^2$. We can therefore apply the Optional Stopping Theorem for submartingales to obtain

$$\mathbf{E}(Z(T^{\omega,\omega'})) \geq Z(0).$$

This implies

$$-\alpha \mathbf{E}(T^{\omega,\omega'}) \geq \delta(0)^2 - 2D\delta(0)$$

and (2.36) follows.

So for any ω, ω'

$$\mathbf{Pr}(T^{\omega,\omega'} \geq eD^2/\alpha) \leq e^{-1}$$

and by considering k consecutive time intervals of length k we obtain

$$\mathbf{Pr}(T^{\omega,\omega'} \geq keD^2/\alpha) \leq e^{-k}$$

and (2.35) follows.

2.5 Hitting Time Lemmas

For a finite Markov chain $\mathcal{M}$ let $\mathbf{Pr}_i, \mathbf{E}_i$ denote probability and expectation, given that $X_0 = i$.

For a set $A \subseteq \Omega$ let

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

Then for $i \neq j$ the *hitting time*

$$H_{i,j} = \mathbf{E}_i(T_j)$$

is the expected number of steps needed to get from state i to state j.

The *commute time*

$$C_{i,j} = H_{i,j} + H_{j,i}.$$

Lemma 2.5.1 *Assume $X_0 = i$ and S is a stopping time with $X_S = i$. Let j be an arbitrary state. Then*

$$\mathbf{E}_i(\textit{number of visits to state } j \textit{ before time } S) = \pi_j \mathbf{E}_i(S).$$

Proof Consider the renewal process whose inter-renewal time is distributed as S. The reward-renewal theorem states that the asymptotic proportion of time spent in state j is given by

$$\mathbf{E}_i(\text{number of visits to } j \text{ before time } S)/\mathbf{E}_i(S).$$

This also equal to π_j, by the ergodic theorem. □

Lemma 2.5.2

$$\mathbf{E}_j(\textit{number of visits to } j \textit{ before } T_i) = \pi_j C_{i,j}.$$

Proof Let S be the time of the first return to i after the first visit to j. Apply Lemma 2.5.1. □

The *cover time* $C(\mathcal{M})$ of $\mathcal{M}$ is $\max_i C_i(\mathcal{M})$ where $C_i(\mathcal{M}) = \mathbf{E}_i(\max_j T_j)$ is the expected time to visit all states starting at i.

Chapter 2

Bounding the Mixing Time

2.1 Spectral Gap

Let P be the transition matrix of an ergodic, reversible Markov chain on state space Ω, Let π be its stationary distribution. Let $N = |\Omega|$ and assume w.l.o.g. that $\Omega = \{0, 1, \ldots, N-1\}$. Let the eigenvalues of P be $1 = \lambda_0 > \lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_{N-1}$. They are all real valued. Let $\lambda_{\max} = \max\{|\lambda_i| : i > 0\}$. The fact that $\lambda_{\max} < 1$ is a classical result of the theory of non-negative matrices. The *spectral gap* $1 - \lambda_{\max}$ determines the mixing rate of the chain in an essential way. The larger it is, the more rapidly does the chain mix. For $U \subseteq \Omega$ let

$$\Delta_U(t) = \max_{i,j \in U} \left\{ \frac{|P^t(i,j) - \pi(j)|}{\pi(j)} \right\}.$$

Theorem 2.1.1 *For all $U \subseteq \Omega$ and $t \geq 0$,*

$$\Delta_U \leq \frac{\lambda_{\max}^t}{\min_{i \in U} \pi(i)}.$$

Proof Let $D^{1/2}$ be the diagonal $\Omega \times \Omega$ matrix with diagonal entries $\sqrt{\pi(\omega)}$, $\omega \in \Omega$ and let $D^{-1/2}$ be its inverse. Then the reversibility of of the chain (1.15) implies that the matrix $S = D^{1/2} P D^{-1/2}$ is symmetric. It has the same eigenvalues as P and its symmetry means that these are all real. We can select an orthonormal basis of column vectors $\mathbf{e}^{(i)}$, $i \in \Omega$ for $\mathbb{R}^\Omega$ consisting of left eigenvectors of S where $\mathbf{e}^{(i)}$ has associated eigenvalue λ_i and $\mathbf{e}^{(0)} = \pi^T D^{-1/2}$. S has the *spectral decomposition*

$$S = \sum_{i=0}^{N-1} \lambda_i \mathbf{e}^{(i)} \mathbf{e}^{(i)T} = \sum_{i=0}^{N-1} \lambda_i E^{(i)},$$

where $E^{(i)} = \mathbf{e}^{(i)}\mathbf{e}^{(i)^T}$. Note that $E^{(i)}E^{(j)} = 0$ for $i \neq j$ and $E^{(i)^2} = E^{(i)}$. It follows that for any $t = 0, 1, 2, \ldots$, $S^t = \sum_{i=0}^{N-1} \lambda_i^t E^{(i)}$. Hence

$$\begin{aligned} P^t = D^{-1/2}S^tD^{1/2} &= \sum_{i=0}^{N-1} \lambda_i^t (D^{-1/2}\mathbf{e}^{(i)}))(\mathbf{e}^{(i)^T}D^{1/2}) \\ &= \mathbf{1}_N\pi^T + \sum_{i=1}^{N-1} \lambda_i^t (D^{-1/2}\mathbf{e}^{(i)}))(\mathbf{e}^{(i)^T}D^{1/2}), \end{aligned}$$

where $\mathbf{1}_N$ is the N-vector all of whose components are 1. In component form, we get with the help of the Cauchy-Schwartz inequality:

$$\begin{aligned} |P^t(j,k) - \pi_k| &= \left| \sqrt{\frac{\pi_k}{\pi_j}} \sum_{i=1}^{N-1} \lambda_i^t \mathbf{e}_j^{(i)} \mathbf{e}_k^{(i)} \right| \\ &\leq \sqrt{\frac{\pi_k}{\pi_j}} \lambda_{\max}^t \left(\sum_{i=0}^{N-1} \mathbf{e}_j^{(i)^2} \right)^{1/2} \left(\sum_{i=0}^{N-1} \mathbf{e}_k^{(i)^2} \right)^{1/2} \\ &= \sqrt{\frac{\pi_k}{\pi_j}} \lambda_{\max}^t. \end{aligned} \tag{2.1}$$

The theorem follows by substitution of the above inequality in the definition of Δ_U. □

In terms of mixing time we have

Corollary 2.1.1

$$\tau(\varepsilon) \leq \left\lceil \frac{\log \varepsilon\pi_{\min}}{\log \lambda_{\max}} \right\rceil.$$

Proof For $A \subseteq \Omega$ we have

$$p_t(A) - \pi(A) \leq \frac{\lambda_{\max}^t}{\pi_{\min}} \pi(A) \leq \frac{\lambda_{\max}^t}{\pi_{\min}}.$$

□

As an example we consider random walk $\mathcal{W}_n$ on the unit hypercube. Here the graph is the n-cube $Q_n = (X_n = \{0,1\}^n, E_n)$ where $x, y \in X_n$ are adjacent in Q_n if their Hamming distance is one i.e. if $|\{i \in [n] : \ x_i \neq y_i\}| = 1$. We add n self loops to each vertex to make the chain lazy.

If G is a d-regular graph without loops and A_G is its adjacency matrix then the probability transition matrix P_G of a random walk on G satisfies $P_G = d^{-1}A_G$.

For graphs $G_i = (V_i, E_i), i = 1, 2$ we can define their product $G = G_1 \times G_2 = (V, E)$ where $V = V_1 \times V_2$ and $E = \{((v_1, v_2), (w_1, w_2)) : \ v_1 = w_1 \text{ and } (v_2, w_2) \in E_2 \text{ or } v_2 = w_2 \text{ and } (v_1, w_1) \in E_1\}$. Then

$$Q_n = K_2 \times K_2 \times \cdots \times K_2 \qquad (n \text{ fold product}). \tag{2.2}$$

Theorem 2.1.2 *If $\mu_i, i = 1, 2, \ldots, m$ and $\nu_i, i = 1, 2, \ldots, n$ are the eigenvalues of matrices A_{G_1}, A_{G_2} respectively, then the eigenvalues of A_G are $\{\mu_i + \nu_j : 1 \le i \le m, 1 \le j \le n\}$.*

Proof A_G can be obtained from A_{G_1} by replacing each 1 by the $|V_2|$ identity matrix I_2, the off-diagonal 0's by the $|V_2| \times |V_2|$ matrix of 0's and replacing each diagonal entry by A_{G_2}. So if $p_G(\lambda) = \det(\lambda I - A_G)$ then

$$p_G(\lambda) = \det p_{G_1}(\lambda I_2 - A_{G_2}).$$

This follows from the following: Suppose the $mn \times mn$ matrix A is decomposed into an $m \times m$ matrix of $n \times n$ blocks $A_{i,j}$. Suppose also that the $A_{i,j}$ commute among themselves. Then

$$\det A = \det \left(\sum_{\sigma} (-1)^{\mathrm{sign}(\sigma)} \prod_{i=1}^{m} A_{i,\sigma(i)} \right),$$

i.e. one can produce an $m \times m$ matrix by a "determinant" calculation and then take its determinant. Needs a proof

So

$$p_G(\lambda) = \det \prod_{i=1}^{n} (\lambda I_2 - A_{G_2} - \mu_i I_2) = \prod_{i=1}^{n} p_{G_2}(\lambda - \mu_i) = \prod_{i=1}^{n} \prod_{j=1}^{n} (\lambda - \mu_i - \nu_j).$$

□

The eigenvalues of K_2 are $\{1, -1\}$ and applying (2.2) we see that the eigenvalues of Q_n are $\{0, \pm 1, \pm 2, \ldots, \pm n\}$ (ignoring multiplicities). To get the eigenvalues for our random walk we (i) divide by n and then (ii) replace each eigenvalue λ by $\frac{1+\lambda}{2}$ to account for adding loops. Thus the second eigenvalue of the walk is $1 - \frac{1}{2n}$.

Applying Corollary 2.1.1 we obtain $\tau(\varepsilon) \le \log(\varepsilon^{-1}) + O(n^2)$. This is a poor estimate, due to our use of the Cauchy-Schwartz inequality in the proof of Theorem 2.1.1. We get an easier and better estimate by using *coupling*.

2.1.1 Decomposition Theorem

2.2 Conductance

The conductance Φ of $\mathcal{M}$ is defined by

$$\Phi = \min\{\Phi_S : S \subseteq \Omega,\ 0 < \pi(S) \le 1/2\}$$

where if $Q(\omega, \sigma) = \pi(\omega) P(\omega, \sigma)$ and $\bar{S} = \Omega \setminus S$,

$$\Phi_S = \pi(S)^{-1} Q(S, \bar{S}).$$

Thus Φ_S is the steady state probability of moving from S to $\bar{S}$ in one step of the chain, conditional on being in S.

Clearly $\Phi \leq \frac{1}{2}$ if $\mathcal{M}$ is lazy.

Note that

$$\Phi_S \pi(S) = Q(S, \bar{S}) = Q(\bar{S}, S) = \Phi_{\bar{S}} \pi(\bar{S}). \tag{2.3}$$

Indeed,

$$Q(S, \bar{S}) = Q(\Omega, \bar{S}) - Q(\bar{S}, \bar{S}) = \pi(\bar{S}) - Q(\bar{S}, \bar{S}) = Q(\bar{S}, S).$$

Let $\pi_{\min} = \min\{\pi(\omega) : \omega \in \Omega\} > 0$ and $\pi_{\max} = \max\{\pi(\omega) : \omega \in \Omega\}$.

2.2.1 Reversible Chains

In this section we show how conductance gives us an estimate of the spectral gap of a reversible chain.

Lemma 2.2.1 *If $\mathcal{M}$ is lazy and ergodic then all eigenvalues are positive.*

Proof $Q = 2P - I \geq 0$ is stochastic and has eigenvalues $\mu_i = 2\lambda_i - 1$, $i = 0, 1, \ldots N-1$. The result follows from $\mu_i > -1$, $i = 0, 1, \ldots N-1$. □

For $y \in \mathbb{R}^N$ let

$$\mathcal{E}(y, y) = \sum_{i<j} \pi_i P_{i,j} (y_i - y_j)^2.$$

Lemma 2.2.2 *If $\mathcal{M}$ is reversible then*

$$1 - \lambda_1 = \min_{\pi^T y = 0} \frac{\mathcal{E}(y, y)}{\sum_i \pi_i y_i^2}.$$

Proof Let $D, S, \mathbf{e}^{(0)}$ be as in Section 2.1. Then by the Rayleigh principle,

$$\lambda_1 = \max_{\pi^T D^{-1/2} x = 0} \frac{x^T D^{1/2} P D^{-1/2} x}{x^T x}.$$

Thus

$$\begin{aligned} 1 - \lambda_1 &= \min_{\pi^T D^{-1/2} x = 0} \frac{x^T D^{1/2} (I - P) D^{-1/2} x}{x^T x} \\ &= \min_{\pi^T y = 0} \frac{y^T D (I - P) y}{y^T D y}. \end{aligned} \tag{2.4}$$

Now

$$\begin{aligned} y^T D(I-P)y &= -\sum_{i\neq j} y_i y_j \pi_i P_{i,j} + \sum_i \pi_i (1-P_{i,i}) y_i^2 \\ &= -\sum_{i\neq j} y_i y_j \pi_i P_{i,j} + \sum_{i\neq j} \pi_i P_{i,j} \frac{y_i^2+y_j^2}{2} \\ &= \sum_{i<j} \pi_i P_{i,j} (y_i - y_j)^2 \\ &= \mathcal{E}(y,y), \end{aligned}$$

and the lemma follows from (2.4). □

Theorem 2.2.1 *If $\mathcal{M}$ is a reversible chain then*

$$1-\lambda_1 \geq \frac{\Phi^2}{2}.$$

Proof Assume now that $\pi^T y = 0$, $y_1 \geq y_2 \geq \cdots \geq y_N$ and that

$$\pi_1 + \pi_2 + \cdots + \pi_{r-1} \leq \frac{1}{2} < \pi_1 + \pi_2 + \cdots + \pi_r.$$

Let $z_i = y_i - y_r$ for $i = 1, 2, \ldots, n$. Then

$$z_1 \geq z_2 \geq \cdots \geq z_r = 0 \geq z_{r+1} \geq \cdots \geq z_N,$$

and

$$\begin{aligned} \frac{\mathcal{E}(y,y)}{\sum_i \pi_i y_i^2} &= \frac{\mathcal{E}(z,z)}{-y_r^2 + \sum_i \pi_i z_i^2} \\ &\geq \frac{\mathcal{E}(z,z)}{\sum_i \pi_i z_i^2}. \qquad (2.5) \\ &= \frac{\left(\sum_{i<j} \pi_i P_{i,j}(z_i - z_j)^2\right)\left(\sum_{i<j} \pi_i P_{i,j}(|z_i|+|z_j|)^2\right)}{\left(\sum_i \pi_i z_i^2\right)\left(\sum_{i<j} \pi_i P_{i,j}(|z_i|+|z_j|)^2\right)} \\ &= \frac{A}{B}, \qquad \text{say.} \end{aligned}$$

Now,

$$\begin{aligned} A &\geq \left(\sum_{i<j} \pi_i P_{i,j} |z_i - z_j| (|z_i| + |z_j|)\right)^2 \qquad \text{by Cauchy-Schwartz} \\ &\geq \left(\sum_{i<j} \pi_i P_{i,j} \sum_{k=i}^{j-1} |z_{k+1}^2 - z_k^2|\right)^2. \qquad (2.6) \end{aligned}$$

We verify (2.6) later. Also,

$$\sum_{i<j} \pi_i P_{i,j}(|z_i| + |z_j|)^2 \le 2\sum_{i<j} \pi_i P_{i,j}(z_i^2 + z_j^2) \le 2\sum_i \pi_i z_i^2.$$

So,

$$\frac{\mathcal{E}(y,y)}{\sum_i \pi_i y_i^2} \ge \frac{A}{B} \ge \frac{\left(\sum_{i<j} \pi_i P_{i,j} \sum_{k=i}^{j-1} |z_{k+1}^2 - z_k^2|\right)^2}{2\left(\sum_i \pi_i z_i^2\right)^2}.$$

Now let $S_k = \{1, 2, \ldots, k\}$ and $C_k = \{(i,j) : i \le k < j\}$. Then

$$\begin{aligned}
\sum_{i<j} \pi_i P_{i,j} \sum_{k=i}^{j-1} |z_{k+1}^2 - z_k^2| &= \sum_{k=1}^{N-1} |z_{k+1}^2 - z_k^2| \sum_{(i,j)\in C_k} \pi_i P_{i,j} \\
&\ge \Phi\left(\sum_{k=1}^{r-1}(z_k^2 - z_{k+1}^2)\pi(S_k) + \sum_{k=r}^{N-1}(z_{k+1}^2 - z_k^2)(1-\pi(S_k))\right) \\
&= \Phi\left(\sum_{k=1}^{N-1}(z_k^2 - z_{k+1}^2)\pi(S_k) + (z_N^2 - z_r^2)\right) \\
&= \Phi\left(\sum_{k=1}^{N} \pi_k z_k^2\right)
\end{aligned}$$

since $z_r = 0$.

Thus if $\pi^T y = 0$ then

$$\frac{\mathcal{E}(y,y)}{\sum_i \pi_i y_i^2} \ge \frac{\Phi^2}{2}$$

and Theorem 2.2.1 follows.

Proof of (2.6)

We show that if $i < j$ then

$$|z_i - z_j|(|z_i| + z_j|) \ge \sum_{k=i}^{j-1} |z_{k+1}^2 - z_k^2|. \tag{2.7}$$

If $r \notin \{i, i+1, \ldots, j\}$ i.e. if z_i, z_j have the same sign then LHS(2.7)=RHS(2.7)=$|z_i^2 - z_j^2|$. Otherwise LHS(2.7)=$(|z_i| + |z_j|)^2$ and RHS(2.7)=$z_i^2 + z_j^2$. □

In terms of mixing time we obtain from Corollary 2.1.1,

Corollary 2.2.1 *If $\mathcal{M}$ is a lazy ergodic chain then*

$$\tau(\varepsilon) \le \left\lceil \frac{2|\log \varepsilon \pi_{\min}|}{\Phi^2} \right\rceil.$$

Proof Lemma 2.2.1 implies that $\lambda_1 = \lambda_{\max}$ and then

$$\frac{1}{\log \lambda_{\max}^{-1}} \leq \frac{1}{\log(1-\Phi^2/2)^{-1}} \leq \frac{2}{\Phi^2}.$$

□

Now consider the conductance of a random walk on a graph $G = (V, E)$. For $S, T \subseteq V$ let $E(S,T) = \{(v,w) \in E : \ v \in S, w \in T\}$ and $e(S,T) = |E(S,T)$. Then, by definition,

$$\Phi_S = \frac{\displaystyle\sum_{(v,w)\in E(S,\bar{S})} \frac{d_v}{2|E|}\frac{1}{d_v}}{\displaystyle\sum_{v\in S}\frac{d_v}{2|E|}} = \frac{e(S,\bar{S})}{\displaystyle\sum_{v\in S} d_v}.$$

In particular when G is an r-regular graph

$$\Phi = r^{-1} \min_{|S|\leq \frac{1}{2}|V|} \frac{e(S,\bar{S})}{|S|}. \tag{2.8}$$

The minimand above is referred to as the *expansion* of G. This graphs with good expansion (*expander graphs*) have large conductance and random walks on them mix rapidly.

As an example consider the n-cube Q_n. For $S \subseteq X_n$ let $in(S)$ denote the number of edges of Q_n which are wholly contained in S.

Lemma 2.2.3 *If $\emptyset \neq S \subseteq X_n$ then $in(S) \leq \frac{1}{2}|S|\log_2|S|$.*

Proof We prove this by induction on n. It is trivial for $n = 1$. For $n > 1$ let $S_i = \{x \in S : \ x_n = i\}$ for $i = 1, 2$. Then

$$in(S) \leq in(S_0) + in(S_1) + \min\{|S_0|, |S_1|\}$$

since the term $\min\{|S_0|, |S_1|\}$ bounds the number of edges which are contained in S and join S_0, S_1. The lemma follows from the inequality

$$x\log_2 x + y\log_2 y + 2y \leq (x+y)\log_2(x+y)$$

for all $x \geq y \geq 0$. The proof is left as a simple exercise in calculus. □

By summing the degrees at each vertex of S we see that

$$e(S,\bar{S}) + 2in(S) = n|S|.$$

By the above lemma we have

$$e(S,\bar{S}) \geq n|S| - \tfrac{1}{2}|S|\log_2|S| \geq |S|$$