Mixing for Markov Chains and Spin Systems

DRAFT of August 31, 2005

Yuval Peres

U.C. Berkeley

peres@stat.berkeley.edu
http://www.stat.berkeley.edu/users/peres

Lectures given at the 2005 PIMS Summer School in Probability held at the University of British Columbia from June 6 through June 30. Special thanks is due to Jesse Goodman, Jeffrey Hood, Ben Hough, Sandra Kliem, Lionel Levine, Yun Long, Asaf Nachmias, Alex Skorokhod and Terry Soo for help preparing these notes. **These notes have not been subjected to the usual scrutiny reserved for formal publications.**

–Yuval

Lecture 1: Introduction and total variation distance

1.1 Introduction

Let Ω be a finite set of cardinality N. A *stochastic matrix* on Ω is a function $\mathrm{P} : \Omega \times \Omega \to \mathbf{R}$ such that

$$\mathrm{P}(x, y) \geq 0 \text{ for all } x, y \in \Omega, \quad \text{and} \quad \sum_{y \in \Omega} \mathrm{P}(x, y) = 1 \text{ for all } x \in \Omega. \tag{1.1}$$

A *Markov chain* with transition matrix P, is a sequence of random variables $X_0, X_1, \ldots$ such that

$$\mathbf{P}\left\{ X_{t+1} = y \middle| X_0 = x_0, X_1 = x_1, \ldots, X_{t-1} = x_{t-1}, X_t = x \right\} = \mathrm{P}(x, y) \tag{1.2}$$

for all sequences $(x_0, x_1, \ldots, x_{t-1}, x, y)$ of elements of Ω. We use the notation $\mathbf{P}_\mu$ and $\mathbb{E}_\mu$ to indicate probabilities and expectations for the chain started with distribution μ. We write $\mathbf{P}_x$ and $\mathbb{E}_x$ when the chain is started in the state x.

A Markov chain is called *irreducible* if for any two states $x, y \in \Omega$, there exists an integer n (possibly depending on x and y) so that $\mathrm{P}^n(x, y) > 0$. This means that it is possible to get from any state to any other state. The chain will be called *aperiodic* if for every state x

$$\mathrm{GCD}\{k \geq 0 \mid \mathrm{P}^k(x, x) > 0\} = 1.$$

where GCD stands for the greatest common divisor.

Proposition 1.1 *Show that the chain is both irreducible and aperiodic if and only if there exists $k \geq 0$ such that for every pair of states x, y we have $\mathrm{P}^k(x, y) > 0$.*

The proof leaves as an exercise.

1.2 Total variation distance

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

$$\|\mu - \nu\|_{TV} = \max_{A \subset \Omega} |\mu(A) - \nu(A)| \,. \tag{1.3}$$

We denote $X \sim \mu$ if the random variable X has distribution μ.

Proposition 1.2 *Let μ and ν be two probability distributions on Ω. Then*

$$\begin{aligned} \|\mu - \nu\|_{TV} &= \frac{1}{2} \sum_{x \in \Omega} |\mu(x) - \nu(x)| && (1.4) \\ &= \inf \left\{ \mathbf{P}\{X \neq Y\} \,:\, X \sim \mu,\, Y \sim \nu \right\} && (1.5) \end{aligned}$$

Proof: To prove the first equality, take $B = \{x : \mu(x) > \nu(x)\}$. We use the set B to decompose the set A:

$$\mu(A) - \nu(A) = \sum_{x \in A \cap B} [\mu(x) - \nu(x)] - \sum_{x \in A \cap B^c} [\nu(x) - \mu(x)] \,. \tag{1.6}$$

Notice that $[\mu(x) - \nu(x)] > 0$ for $x \in B$, and $[\nu(x) - \mu(x)] \geq 0$ for $x \in B^c$. It follows that

$$\mu(A) - \nu(A) \leq \sum_{x \in B} [\mu(x) - \nu(x)] = \mu(B) - \nu(B).$$

We obtain similarly that

$$\mu(A) - \nu(A) \geq \mu(B^c) - \nu(B^c) \tag{1.7}$$

and since $\mu(B) - \nu(B) = \nu(B^c) - \mu(B^c)$. Thus,

$$\begin{aligned} \|\mu - \nu\|_{TV} &= \frac{1}{2} [\mu(B) - \nu(B) + \nu(B^c) - \mu(B^c)] \\ &= \frac{1}{2} \left[\sum_{x \in B} (\mu(x) - \nu(x)) + \sum_{x \in B^c} (\nu(x) - \mu(x)) \right] = \frac{1}{2} \left[\sum_{x \in \Omega} |\mu(x) - \nu(x)| \right]. \end{aligned} \tag{1.8}$$

This establishes (1.4).

Define

$$p(x) := \frac{\mu(x) \wedge \nu(x)}{z},$$

where $z = 1 - \|\mu - \nu\|_{TV}$. It's easy to verify that p is a probability distribution. Now define probability distributions $\tilde{\mu}$ and $\tilde{\nu}$ through the equations

$$\mu = zp + (1 - z)\tilde{\mu} \tag{1.9}$$

$$\nu = zp + (1 - z)\tilde{\nu}. \tag{1.10}$$

Since $z \leq 1$ and $zp(x) \leq \mu(x)$, this does implicitly define legal probability distributions $\tilde{\mu}$ and $\tilde{\nu}$.

We can generate a pair (X, Y) as follows: A coin with probability z of "heads" is tossed. If "heads", generate X according to p, and set $Y = X$. If "tails", generate X according to $\tilde{\mu}$ and, independently, generate y according to $\tilde{\nu}$.

The reader should convince herself that the marginal distributions of (X, Y) are μ and ν.

Notice that $\tilde{\mu}(x) > 0$ if and only if $\nu(x) < \mu(x)$, and likewise $\tilde{\nu}(x) > 0$ if and only if $\mu(x) < \nu(x)$. Thus the supports of μ and ν are disjoint. This means that if the coin lands "tails", then $X \neq Y$. We conclude that

$$\mathbf{P}\{X \neq Y\} = 1 - z = \|\mu - \nu\|_{TV}.$$

Consequently,

$$\inf \{\mathbf{P}\{X \neq Y\} : X \sim \mu, Y \sim \nu\} \leq \|\mu - \nu\|_{TV}. \tag{1.11}$$

On the other hand, for any pair (X, Y) with the correct marginals, and any set $A \subset \Omega$,

$$\mu(A) - \nu(A) = \mathbf{P}\{X \in A\} - \mathbf{P}\{Y \in A\} \leq \mathbf{P}\{X \in A, Y \notin A\} \leq \mathbf{P}\{X \neq Y\}.$$

Taking the maximum over A and the minimum over pairs (X, Y), we get that

$$\|\mu - \nu\|_{TV} \leq \inf \{\mathbf{P}\{X \neq Y\} : X \sim \mu, Y \sim \nu\} \,. \tag{1.12}$$

Together (1.11) and (1.12) prove (1.5). ∎

Lecture 2: Convergence Theorem and Coupling

2.1 Convergence Theorem

Consider an irreducible aperiodic Markov chain with transition matrix P on a finite state space Ω. A measure π on Ω is called *stationary* if it satisfies

$$\pi = \pi \mathrm{P}. \tag{2.1}$$

If in addition, π is a probability measure ($\pi(\Omega) = 1$) then it is called a *stationary distribution*. The condition (2.1) can be rewritten as a system of linear equations:

$$\pi(y) = \sum_{x \in \Omega} \pi(x) \mathrm{P}(x, y), \quad y \in \Omega. \tag{2.2}$$

Theorem 2.1 *Let $\{X_j\}_{j \geq 0}$ be a Markov Chain on a finite space Ω. Fix any state $a \in \Omega$, and let $\tau_a^+ := \inf\{n \geq 1 : X_n = a\}$ be the first positive time that the chain hits the state a. We also define another hitting time $\tau_a = \inf\{n \geq 0 : X_n = a\}$. τ_a^+ differs from τ_a only when $X_0 = a$. Define*

$$\tilde{\pi}(x) = \mathbb{E}_a \sum_{n=0}^{\tau_a^+ - 1} \mathbf{1}_{\{X_n = x\}}.$$

In words, $\tilde{\pi}(x)$ is the expected number of visits to x, starting at a and before the chain returns to a. Then, $\tilde{\pi}(x)$ gives a stationary measure.

Proof: We will prove a stronger result as Theorem 2.6. ■

Theorem 2.2 *Define*

$$\pi(x) = \frac{\tilde{\pi}(x)}{\mathbb{E}_a \tau_a^+}. \tag{2.3}$$

Then, π is a stationary distribution.

Proof: Since $\sum_{x \in \Omega} \tilde{\pi}(x) = \mathbb{E}_a \tau_a^+$. We only have to prove that $E_a \tau_a^+ < \infty$.
Let r be an integer such that $\mathrm{P}^r(x, y) > 0$ for all x and y in Ω. Such an r is guaranteed to exist by irreducibility and aperiodicity. Let $\epsilon = \min_{x, y \in \Omega} \mathrm{P}^r(x, y)$.

Suppose that for some k,

$$\mathbf{P}_a\{\tau_a^+ > kr\} \leq (1 - \epsilon)^k \tag{2.4}$$

Then

$$\begin{aligned}
\mathbf{P}_a\{\tau_a^+ > (k+1)r\} &\leq \mathbf{P}_a\{\tau_a^+ > kr,\, X_{(k+1)r} \neq a\} \\
&\leq \mathbf{P}_a\{\tau_a^+ > kr\} \sup_{x \in \Omega} \mathbf{P}_x\{X_r \neq a\} \\
&\leq (1 - \epsilon)^k (1 - \epsilon)
\end{aligned}$$

So, by induction, equation (2.4) is true for all k. Thus:

$$\mathbb{E}_a\{\tau_a^+\} = \sum_{n=0}^{\infty} \mathbf{P}_a\{\tau_a^+ > n\} = \sum_{k=0}^{\infty}\sum_{j=0}^{r-1} \mathbf{P}_a\{\tau_a^+ > kr + j\} \leq \sum_{k=0}^{\infty} r\mathbf{P}_a\{\tau_a^+ > kr\} < \infty.$$

■

Definition 2.3 *A function h is* harmonic *at $x \in \Omega$ if*

$$h(x) = \sum_{y\in\Omega} \mathrm{P}(x,y)h(y). \tag{2.5}$$

A function is harmonic on $D \subset \Omega$ if it is harmonic at every point x of D. If h is regarded as a column vector, then a function which is harmonic on all of Ω satisfies the matrix equation $\mathrm{P}h = h$.

Lemma 2.4 *Functions which are harmonic everywhere on Ω must be constant functions.*

Proof: Let h be a harmonic function on Ω. Let $A = \{x \in \Omega : h(x) = \max_\Omega h\}$ be the set of points where h takes its maximum value. Since Ω is finite, A is nonempty. Let $x \in A$, and let $N_x = \{y : P(x,y) > 0\}$. If there is $y_0 \in N_x$ with $h(y_0) < h(x)$, then

$$h(x) = \mathrm{P}(x,y_0)h(y_0) + \sum_{y\in N_x\setminus\{y_0\}} \mathrm{P}(x,y)h(y) < h(x),$$

a contradiction. It follows that $h(y) = h(x)$ for all y such that $P(x,y) > 0$. Since the chain is irreducible, h must be constant. ■

Theorem 2.5 *The stationary distribution is unique.*

Proof: From equation (2.1), we see that in order to prove π is unique, we only have to prove that the matrix $(\mathrm{P} - I)$ has rank $N-1$. This is equivalent to the fact that $(\mathrm{P}-I)f = 0$ has only constant solutions, which is a consequence of Lemma 2.4. ■

Theorem 2.6 *Suppose $\tau > 0$ is a stopping time with $\mathbf{P}_a[X_\tau = a] = 1$, $\mathbb{E}_a\tau < \infty$. Then the measure*

$$\mu(x) := \mathbb{E}_a \sum_{t=0}^{\tau-1} \mathbf{1}_{\{X_t = x\}}$$

is stationary.

Proof: Fix a state y. From all those paths that hit y at time $j+1$, considering the position of the path at time j, we have

$$\sum_{j=0}^{\tau-1} \mathbf{1}_{\{X_j=y\}} = \sum_{j=0}^{\tau-1} \mathbf{1}_{\{X_{j+1}=y\}} = \sum_x \sum_{j=0}^{\tau-1} \mathbf{1}_{\{X_j=x\}}\mathbf{1}_{\{X_{j+1}=y\}}.$$

Taking the expectation $\mathbb{E}_a$, we obtain

$$\begin{aligned}
\mu(y) &= \sum_x \sum_{j=0}^{\infty} \mathbf{P}(\tau > j, X_j = x, X_{j+1} = y) \\
&= \sum_x \sum_{j=0}^{\infty} \mathbf{P}(\tau > j, X_j = x)\mathbf{P}(X_{j+1} = y|\tau > j, X_j = x) \qquad (2.6)
\end{aligned}$$

Since $\mathbf{1}_{\{\tau>j\}}$ is a function of $X_1, \dots, X_j$, by the Markov property, $\mathbf{P}(X_{j+1} = y | \tau > j, X_j = x) = \mathrm{P}(x, y)$. We also have $\sum_{j=0}^{\infty} \mathbf{P}(\tau > j, X_j = x) = \mu(x)$. So the right side of (2.6) reduces to $\sum_x \mu(x)\mathrm{P}(x, y)$, which shows that μ is stationary.

To prevent cheating, we also have to prove that each $\mu(x)$ is finite. This is easy since $\sum_x \mu(x) = \mathbb{E}_a \tau < \infty$ ∎

2.2 Coupling

Theorem 2.7 *Suppose that* P *is both irreducible and aperiodic. Let* π *be the stationary distribution. Then for any* $x \in \Omega$*,*

$$\lim_{n\to\infty} \|\mathrm{P}^n(x, \cdot) - \pi\|_{TV} = 0.$$

Moreover, the convergence is geometrically fast.

Proof: Let r be such that P^r has all strictly positive entries. Then for sufficiently small ϵ there exists a stochastic matrix Q satisfying

$$\mathrm{P}^r = \epsilon\pi + (1 - \epsilon)\mathrm{Q}, \tag{2.7}$$

where π is the matrix with $N = |\Omega|$ rows, each identical to the row vector π.

According to equation (2.7), we can generate an r-step move of the Markov chain P as follows: If an ϵ-coin lands "heads", generate an observation from π and take this as the new state, while if the coin lands "tails", make a move according to the transition matrix Q. The first time the coin lands "heads", the distribution of the new position is exactly stationarity. Then using the coupling characterization of total variation distance (Proposition 1.2) , the total variation distance $\|(\mathrm{P}^r)^k(x, \cdot) - \pi\|_{TV}$ can be bounded by $(1 - \epsilon)^k$, the probability of no heads in k tosses.
Again by Proposition 1.2, we know that if we run two chains one step forward respectively, then their total variation distance will become smaller. So, we can easily extend the last result to the following one, which leads to our conclusion:

$$\|\mathrm{P}^n(x, \cdot) - \pi\|_{TV} \le (1 - \epsilon)^{\lfloor n/r \rfloor}.$$

∎

For the rest of this subsection, we assume that the two Markov chains $\{(X_t, Y_t)\}$ are coupled so that

$$\text{if } X_s = Y_s, \quad \text{then } X_t = Y_t \text{ for } t \ge s, \tag{2.8}$$

and τ_{couple} is the first time the two chains meet.

Theorem 2.8 *Suppose that for each pair of states* x, y *there is a coupling* $\{(X_t, Y_t)\}$ *with* $X_0 = x$ *and* $Y_0 = y$*. Let*

$$\tau_{couple} := \min\{t \,:\, X_t = Y_t\}. \tag{2.9}$$

Then

$$\max_{\mu} \|\mu\mathrm{P}^t - \pi\|_{TV} \le \max_{x,y\in\Omega} \mathbf{P}\{\tau_{couple} > t\}.$$

First, we show that the distribution of the chain started from μ and the distribution of the chain started from ν are bounded by the meeting time of coupled chains.

Proposition 2.9 *If μ is the distribution of X_0 and ν is the distribution of Y_0, then*

$$\|\mu \mathrm{P}^t - \nu \mathrm{P}^t\|_{TV} \leq \mathbf{P}\{\tau_{couple} > t\}.$$

Proof:

$$\begin{aligned}\mu \mathrm{P}^t(z) - \nu \mathrm{P}^t(z) = \mathbf{P}\{X_t = z, \tau_{couple} \leq t\} + \mathbf{P}\{X_t = z, \tau_{couple} > t\} \\ - \mathbf{P}\{Y_t = z, \tau_{couple} \leq t\} - \mathbf{P}\{Y_t = z, \tau_{couple} > t\}\end{aligned}$$

Now since $X_t = Y_t$ when $\tau_{couple} \leq t$, the first and the third terms cancel, so

$$\mu \mathrm{P}^t(z) - \nu \mathrm{P}^t(z) = \mathbf{P}\{X_t = z, \tau_{couple} > t\} - \mathbf{P}\{Y_t = z, \tau_{couple} > t\}.$$

Thus

$$\left\|\mu \mathrm{P}^t - \nu \mathrm{P}^t\right\|_{TV} \leq \frac{1}{2} \sum_z \left[\mathbf{P}\{X_t = z, \tau_{couple} > t\} + \mathbf{P}\{Y_t = z, \tau_{couple} > t\}\right] = \mathbf{P}\{\tau_{couple} > t\}.$$

■

The following lemma combined with Proposition 2.9 establishes Theorem 2.8.

Lemma 2.10

$$\max_\mu \|\mu \mathrm{P}^t - \pi\|_{TV} = \max_x \|\mathrm{P}^t(x, \cdot) - \pi\|_{TV} \leq \max_{x,y \in \Omega} \|\mathrm{P}^t(x, \cdot) - \mathrm{P}^t(y, \cdot)\|_{TV}. \tag{2.10}$$

Proof: As π is stationary, $\pi(A) = \sum_y \pi(y) \mathrm{P}^t(y, A)$ for any set A. Using this shows that

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

Maximize over A, we get for any state x:

$$\|\mathrm{P}^t(x, \cdot) - \pi\|_{TV} \leq \max_y \|\mathrm{P}^t(x, \cdot) - \mathrm{P}^t(y, \cdot)\|_{TV}. \tag{2.11}$$

Which proves the second half. The first equality is similar, and left to the reader as an exercise. ■

Lemma 2.11 *If $\overline{d}(t) := \max_{x,y \in \Omega} \|\mathrm{P}^t(x, \cdot) - \mathrm{P}^t(y, \cdot)\|_{TV}$, then $\overline{d}(s + t) \leq \overline{d}(s)\overline{d}(t)$.*

Proof: Let (X_s, Y_s) be the optimal coupling of $\mathrm{P}^s(x, \cdot)$ and $\mathrm{P}^s(y, \cdot)$ which attains the total variation distance in equation 1.5. Then

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

We have the same equality for Y_s. Subtract the two equalities gives:

$$\mathrm{P}^{s+t}(x, w) - \mathrm{P}^{s+t}(y, w) = \mathbb{E}\{\mathrm{P}^t(X_s, w) - \mathrm{P}^t(Y_s, w)\}. \tag{2.12}$$

Summing over w establishes:

$$\begin{aligned}\|\mathrm{P}^{s+t}(x,\cdot)-\mathrm{P}^{s+t}(y,\cdot)\|_{\mathrm{TV}} &= \frac{1}{2}\sum_{w}\left|\mathbb{E}\{\mathrm{P}^{t}(X_s,w)-\mathrm{P}^{t}(Y_s,w)\}\right| \\ &\leq \mathbb{E}\|\mathrm{P}^{t}(X_s,\cdot)-\mathrm{P}^{t}(Y_s,\cdot)\|_{\mathrm{TV}}\end{aligned}$$

The total variation distance inside the expectation is zero whenever $X_s = Y_s$. Moreover, this distance is always bounded by $\overline{d}(t)$. Since (X_s, Y_s) is an optimal coupling, we obtain

$$\begin{aligned}\|\mathrm{P}^{s+t}(x,\cdot)-\mathrm{P}^{s+t}(y,\cdot)\|_{\mathrm{TV}} &\leq \overline{d}(t)\mathbf{P}\{X_s \neq Y_s\} \\ &= \overline{d}(t)\|\mathrm{P}^{s}(x,\cdot)-\mathrm{P}^{s}(y,\cdot)\|_{\mathrm{TV}} \\ &\leq \overline{d}(t)\overline{d}(s)\end{aligned}$$

Maximizing over x, y completes the proof. ■

2.3 Examples of coupling time

Example 2.12 Consider a Markov chain on the hypercube $\Omega = \{0,1\}^d$. Each state x in Ω can be represented by the ordered d-tuple $(x_1, \ldots, x_d)$, in which each x_j may assume two different values, 0 and 1. The chain is run as follows. At each time, we pick a coordinate x_j uniformly, and change the value of x_j with probability 1/2. Let τ be the first time that all the coordinates have been selected at least once and $\tau^{(l)}$ be the first time that l distinct coordinates have been chosen. Let see what what can be said about the distribution of the stopping time $\tau^{(l)}$.

The event $\{\tau^{(l+1)} - \tau^{(l)} \geq k\}$ means that in the next $(k-1)$ steps following $\tau^{(l)}$, we never pick a new coordinate. Thus

$$\mathbf{P}((\tau^{(l+1)} - \tau^{(l)}) \geq k) = \left(\frac{l}{d}\right)^{k-1}.$$

From this, we get

$$\mathbb{E}(\tau^{(l+1)} - \tau^{(l)}) = \sum_{k\geq 1}\left(\frac{l}{d}\right)^{k-1} = \frac{d}{d-l}.$$

Summing over l gives

$$\mathbb{E}\tau = \mathbb{E}\tau^{(d)} \sim d\log d.$$

Example 2.13 Lazy random walk on cycle. Consider n points $\{0, 1, \ldots, n-1\}$ on a circle. Two points a and b are neighbors if and only if $a \equiv b \pm 1 \mod n$. The chain (X_t, Y_t) starts at two points x, y on this circle. At each time, if X_t and Y_t have not yet met, first toss a coin to decide if X_t or Y_t moves, then make a move of that chosen point to the left neighbor or to the right neighbor, with probability 1/2 each. If X_t and Y_t have already met, then move them together.

We look at the process X_t, it is a lazy random walk. Each time the chain has probability 1/2 to be still, and probability 1/4 to move left and right respectively. Let τ be the first meeting time of X_t and Y_t. We want to compute $\mathbb{E}\tau$.

Let $\{Z_t\}$ be a simple random walk on $\mathbb{Z}$ starting at $k = x - y \mod n$. The two processes $\{X_t - Y_t \mod n\}$ and $\{Z_t \mod n\}$ has the same distributions. Let $\tau_0 := \inf\{t \geq 0 : Z_t = 0 \text{ or } n\}$. Then, $\mathbb{E}\tau = \mathbb{E}\tau_0$.

We have two methods to find $\mathbb{E}\tau_0$. One is to write f_k for the expected time $\mathbb{E}_k(\tau_0)$ started at state k. Clearly, $f_0 = f_n = 0$. For other values of k, considering the first step gives

$$f_k = \frac{1}{2}\mathbb{E}(\tau_0 | \text{ walk moves to } k+1) + \frac{1}{2}\mathbb{E}(\tau_0 | \text{ walk moves to } k-1)$$

This gives the recurrence formula:

$$f_k = 1 + \frac{1}{2}\left(f_{k+1} + f_{k-1}\right). \tag{2.13}$$

Exercise 2.14 *Check that the recurrence 2.13 has a unique solution $f_k = k(n-k)$.*

The other way to get $\mathbb{E}\tau_0$ is indicated by the following exercise.

Exercise 2.15 *Prove that $Z_t^2 - t$ is a martingale. Use the Optional Sampling Theorem to prove $\mathbb{E}\tau_0 = k(n-k)$.*

In general, no matter what k is,

$$\mathbb{E}_k\tau \le (n/2)^2. \tag{2.14}$$

Example 2.16 D-dimensional discrete torus The d-dimensional torus has vertices in $\mathbb{Z}_n^d$. Two vertices $x = (x^1, \dots, x^d)$ and $y = (y^1, \dots, y^d)$ are neighbors if for some coordinate j, $x^j \equiv y^j \pm 1 \mod n$, and $x^i = y^i$ for all $i \neq j$

We couple two lazy random walks started at $x, y \in \mathbb{Z}_n^d$ as follows: First, we pick one of the d coordinates at random. If the two chains agree in this coordinate, we move each of the chains identically, adding ± 1 to the chosen coordinate with probability $\frac{1}{4}$ each, and doing nothing with probability $\frac{1}{2}$. If the two chains do not agree in the chosen coordinate, we pick one of the chains at random to move, leaving the other fixed. For the chain selected to move, we add ± 1 to the chosen coordinate with probability $\frac{1}{2}$ each.

Let τ_i be the time it takes for the i-th coordinate to couple. Each time this coordinate is selected, the process is just like the case of the cycle $\mathbb{Z}_n$ (Example 2.13). Since the i-th coordinate is selected with probability $1/d$ at each move, there is a geometric waiting time between moves with expectation d. It follows from (2.14) that

$$\mathbb{E}(\tau_i) \le \frac{dn^2}{4}. \tag{2.15}$$

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

$$\mathbb{E}(\tau_{couple}) \le \frac{d^2n^2}{4}. \tag{2.16}$$

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

$$\mathbf{P}\{\tau_{couple} > t\} \le \frac{1}{t}\mathbb{E}(\tau_{couple}) \le \frac{1}{t}\frac{d^2n^2}{4}$$

Taking $t_0 = d^2n^2$ shows that $\bar{d}(t_0) \le \frac{1}{4}$. Using Lemma 2.11 shows that if $t = \lceil \log_4(\epsilon^{-1}) \rceil d^2n^2$ then $\bar{d}(t) \le \epsilon$. In other words, $\tau(\epsilon) = O\left(c(d)n^2\log(\epsilon^{-1})\right)$.

Exercise 2.17 *Starting from equation (2.15), prove that there exists a constant A, such that*

$$\mathbb{E}(\tau_{couple}) \le A \cdot d\log d \cdot n^2 \tag{2.17}$$

which is a stronger version of (2.16).

Lecture 3: Path Coupling and the Kantorovich Metric

3.1 Markov Chain Review

Consider a Markov chain on a finite state space Ω. The chain is *irreducible* if for any two states $x, y \in \Omega$ there exists $k \geq 0$ such that $p^k(x, y) > 0$. If the chain is irreducible and aperiodic, then, there exists $k \geq 0$ such that for every pair of states x, y, $\mathrm{P}^k(x, y) > 0$.

Convergence theorem: If the chain is irreducible and aperiodic, then

$$||\mathrm{P}^t(x, \cdot) - \pi||_{\mathrm{TV}} \to 0 \qquad \text{as } t \to \infty,$$

where π is the stationary distribution. Note that this implies the uniqueness of the stationary distribution; in fact, this uniqueness holds even without the assumption of aperiodicity. To deduce this from the convergence theorem, use the lazy chain, with transition matrix $\frac{\mathrm{P}+\mathrm{Id}}{2}$. The lazy chain is aperiodic, so the convergence theorem implies uniqueness of stationary distribution. Since $\pi\mathrm{P} = \pi$ if and only if $\pi\frac{\mathrm{P}+\mathrm{Id}}{2} = \pi$, the uniqueness applies to the original chain as well. This is the second proof of theorem 2.5.

3.2 Glauber Dynamics for Graph Coloring

Let $G = (V, E)$ be a finite undirected graph with all vertex degrees $\leq \Delta$. A *q-coloring* of G is a map $f : V \to \{1, \ldots, q\}$; the coloring is *proper* if adjacent vertices receive distinct colors: $x \sim y \Rightarrow f(x) \neq f(y)$. The minimal q for which there exists a proper q-coloring is called the *chromatic number* of G, denoted $\chi(G)$.

We would like to understand the geometry of the space of proper colorings; in particular, we are interested in sampling a uniform or close-to-uniform proper q-coloring. Define a graph structure on the space of all q-colorings of G by putting f and g adjacent if they differ at a single vertex. Denote by $d(f, g)$ the Hamming distance between colorings f and g; this is also the length of the shortest path from f to g. Note, however, that if f and g are proper, the shortest path joining f and g in the space of *proper* colorings may be strictly longer than $d(f, g)$.

The *Glauber dynamics* on proper colorings are defined as follows: at each time step, choose a vertex uniformly at random, and change its color to one chosen uniformly at random from among those different from the colors of the neighboring vertices. This rule ensures that if we start from a proper coloring, the dynamics will continue to produce proper colorings. Note that

- If $q \geq \Delta + 1$, there exists a proper q-coloring (use a greedy algorithm).
- If $q \geq \Delta + 2$, the graph of proper q-colorings is connected, and hence the Glauber dynamics are irreducible. To see this, suppose f and g are distinct colorings, and let x be a vertex with $f(x) \neq g(x)$. Let $c = g(x)$. Since $q \geq \Delta + 2$, for each neighbor y of x satisfying $f(y) = c$, we can find a different color c' such that changing $f(y)$ to c' will result in a proper coloring. After making these changes, no neighbor of x has f-color c, so changing $f(x)$ to c will again result in a proper coloring. We have produced a proper coloring $\tilde{f}$ along with a path in the space of proper colorings from f to $\tilde{f}$, such that $d(\tilde{f}, g) < d(f, g)$. By induction on the distance, we can produce a path from f to g.

- If $q \geq \Delta + 3$, the Glauber dynamics are aperiodic.

Exercise 3.1 *Show that on a finite binary tree, the space of* 3*-colorings is connected. This shows that the above bounds can be far from sharp. Hint: Induct on the depth of the tree.*

Open Question: Is $q \geq \Delta + C$ or $\Delta(1+\epsilon) + C$ enough to ensure polynomial time mixing for the Glauber dynamics? We want polynomial time in $n = |V|$ for constant q, Δ, i.e.

$$\tau_1\left(\frac{1}{4}\right) \leq C_1 n^\ell$$

for some constants C_1 and ℓ which may depend on q and Δ.

What's known:

- If $q > 2\Delta$, then $\tau_1 = O(n \log n)$. (Jerome '95 / Kotecky)
- Later improvements: $q > (2-\epsilon)\Delta \Rightarrow \tau_1 = O(n \log n)$
- $q > \frac{11\Delta}{6} \Rightarrow \tau_1 = O(n^2 \log n)$. (Vigoda '99)

3.3 Path Coupling

The following lemma shows the basic connection between coupling and mixing.

Lemma 3.2 *For any coupling of two copies of the chain X_t, Y_t started from $X_0 = x, Y_0 = y$, we have*

$$||\mathrm{P}^t(x,\cdot) - \mathrm{P}^t(y,\cdot)||_{\mathrm{TV}} \leq \mathbf{P}(X_t \neq Y_t). \tag{3.1}$$

Proof: This is directly follows the proposition of the total variation distance:

$$||\mathrm{P}^t(x,\cdot) - \mathrm{P}^t(y,\cdot)||_{TV} = \inf\{\mathbf{P}\{X \neq Y\} \,:\, X \sim \mathrm{P}^t(x,\cdot),\, Y \sim \mathrm{P}^t(y,\cdot)\}$$

■

Let d be any metric on Ω satisfying $d(x,y) \geq 1$ whenever $x \neq y$. Then the right side of (3.1) is bounded above by $\mathbb{E}d(X_t, Y_t)$. This suggests a contraction approach: find a coupling such that

$$\mathbb{E}d(X_t, Y_t) \leq e^{-\gamma}\mathbb{E}d(X_{t-1}, Y_{t-1}) \tag{3.2}$$

for some $\gamma > 0$, so that

$$||\mathrm{P}^t(x,\cdot) - \mathrm{P}^t(y,\cdot)||_{\mathrm{TV}} \leq e^{-\gamma t} d(x,y) \leq e^{-\gamma t}\mathrm{Diam}(\Omega).$$

Solving for the time t which makes this distance $\leq \epsilon$, we obtain a bound on the mixing time:

$$\tau_1(\epsilon) \leq \frac{1}{\gamma} \log \frac{\mathrm{Diam}(\Omega)}{\epsilon}. \tag{3.3}$$

Now the question becomes how to verify the contraction condition (3.2) for a reasonable value of γ. In order to get polynomial time mixing we need γ to be polynomial in $\frac{1}{n}$.

3.4 Kantorovich Metric

To reduce the amount of work involved in checking (3.2), Bubbley and Dyer noticed that under certain conditions it suffices to check (3.2) on neighboring vertices $x \sim y$ and for a single step of the chain. In part they were redoing work of Kantorovich (1942). Given a finite metric space (Ω, d), the *Kantorovich metric* d_K is a distance on probability measures on Ω, defined by

$$d_K(\mu, \nu) = \inf_{X\sim\mu, Y\sim\nu} \mathbb{E}d(X, Y). \tag{3.4}$$

The infimum is over all couplings of random variables X and Y distributed as μ and ν, or joint distributions having marginals μ and ν.

The joint distribution of X and Y is specified by an $\Omega \times \Omega$ matrix $M(x, y) = \mathbf{P}(X = x, Y = y)$, with given row and column sums

$$\sum_x M(x, y) = \nu(y);$$
$$\sum_y M(x, y) = \mu(x).$$

Given a coupling M, we have

$$\mathbb{E}d(X, Y) = \sum_{x,y} M(x, y)d(x, y). \tag{3.5}$$

The Kantorovich distance is obtained by minimizing the linear functional in (3.5), and hence the infimum in (3.4) is attained (so it can be replaced by a minimum).

Two simple properties of the Kantorovich metric are worth mentioning. For $x \in \Omega$ let δ_x be the probability distribution concentrated on x. Then

$$d_K(\delta_x, \delta_y) = d(x, y).$$

Secondly, if d is the discrete metric ($d(x, y) = 1$ for all $x \neq y$), then the Kantorovich metric coincides with the total variation distance:

$$d_K(\mu, \nu) = \inf_{X\sim\mu, Y\sim\nu} \mathbf{P}(X \neq Y) = ||\mu - \nu||_{\mathrm{TV}}.$$

Lemma 3.3 *d_K is a metric.*

Proof: Only the triangle inequality is nontrivial. Given random variables X, Y, Z distributed as μ, ν, λ, let $p_1(x, y)$ be the coupling of X and Y that realizes $d_K(\mu, \nu)$, and let $p_2(y, z)$ be the coupling of Y and Z that realizes $d_K(\mu, \lambda)$. Define a coupling of all three random variables by the joint distribution

$$p(x, y, z) = \frac{p_1(x, y)p_2(y, z)}{\nu(y)}.$$

Then

$$\sum_x p(x, y, z) = \frac{\nu(y)p_2(y, z)}{\nu(y)} = p_2(y, z).$$
$$\sum_z p(x, y, z) = \frac{p_1(x, y)\nu(y)}{\nu(y)} = p_1(x, y).$$

Thus in our coupling we have $\mathbb{E}d(X,Y) = d_K(\mu,\nu)$ and $\mathbb{E}d(Y,Z) = d_K(\nu,\lambda)$. Our underlying metric d obeys the triangle inequality; taking expectations in $d(X,Z) \le d(X,Y) + d(Y,Z)$ we obtain

$$d_K(\mu,\lambda) \le \mathbb{E}d(X,Z) \le d_K(\mu,\nu) + d_K(\nu,\lambda).$$

∎

The Kantorovich metric has a simple interpretation in terms of transportation of goods which makes the triangle inequality intuitively obvious. Suppose that the supply of some good is distributed in different cities (elements of Ω) according to μ, and the demand for the good is distributed according to ν. Suppose further that the cost of transporting a given quantity of goods between cities x and y is proportional to $d(x,y)$. We wish to find the most cost-effective way of transporting the goods from distribution μ to distribution ν. If we choose to transport $M(x,y)$ units from x to y, then the sum on the right side of (3.5) is the total cost of transporting the goods. The Kantorovich distance $d_K(\mu,\nu)$ minimizes this sum, so it is the lowest possible cost of transporting the goods.

Using this transportation analogy, $d_K(\mu,\lambda)$ is the minimum possible cost of transporting goods from distribution μ to distribution λ. One way to do this is to transport them via an intermediate distribution ν, which explains the triangle inequality.

Given a finite edge-weighted graph Γ, the *path metric* for Γ is the distance on vertices of Γ

$$d(v,w) = \inf_{v=x_0\sim x_1\sim\ldots\sim x_k=w} \sum_{i=0}^{k-1} \ell(x_i, x_{i+1}),$$

where $\ell(x,y)$ is the length (weight) of the edge (x,y). Informally, $d(v,w)$ is the length of the shortest path from v to w.

Theorem 3.4 *(Bubbley-Dyer) Suppose that the underlying metric d is the path metric for some graph Γ. Moreover suppose that $d(x,y) \ge 1$ whenever $x \ne y$. If the contraction condition*

$$d_K(\delta_x P, \delta_y P) \le e^{-\gamma} d(x,y) \tag{3.6}$$

holds for neighboring vertices $x,y \in \Gamma$, then it holds more generally for any pair of vertices v,w and at any time t:

$$d_K(\delta_v P^t, \delta_w P^t) \le e^{-\gamma t} d(v,w) \tag{3.7}$$

Remark. The graph Γ need not be related to the dynamics of the Markov chain, even though both live on the same space Ω. In the Glauber dynamics for graph colorings, for example, we'll take Γ to be the Hamming graph on the space of all colorings (proper or otherwise), in which two colorings are adjacent if they differ on just one vertex. The Hamming distance on colorings is the corresponding path metric. The transition probabilites for the Glauber dynamics, however, are not the same as those for nearest-neighbor random walk on the Hamming graph.

Proof: We will treat the case $t = 1$. The general case is proved in the next lecture. Let $v = x_0, x_1, \ldots, x_{k-1}, x_k = w$ be a path of minimal length in Γ from v to w. Then

$$\begin{aligned} d_K(\delta_v P, \delta_w P) &\le \sum_{i=0}^{k-1} d_K(\delta_{x_i} P, \delta_{x_{i+1}} P) \\ &\le e^{-\gamma} \sum_{i=0}^{k-1} d(x_i, x_{i+1}) \\ &= e^{-\gamma} d(v,w), \end{aligned}$$

where in the last step we have used the fact that d is the path metric for Γ. ∎

Lecture 4: Reversible Chains, Proper Colorings, Ising Model

4.1 Reversible Chains

Definition 4.1 *A Markov chain* P *on* Ω *is* reversible *if*

$$\pi(x)\mathrm{P}(x,y) = \pi(y)\mathrm{P}(y,x). \tag{4.1}$$

for all states x and y.

The conditions (4.1) are also called the *detailed balance equations*. Note that if a vector π satisfies condition (4.1), then it is stationary for P. This can be easily seen by summing both side of (4.1) over x and recalling that P is a stochastic matrix.

One should think about a reversible chain as a chain which looks the same when run backwards, provided it is started according to π. Here is a nice property of reversible chains, due to Coppersmith, Tetali and Winkler. Recall that we denote the first time that state b is reached by a chain by $\tau_b = min(n \geq 0 : X_n = b)$.

Exercise 4.2 *Prove that for any three states a, b, c in a reversible chain*

$$\mathbb{E}_a(\tau_b) + \mathbb{E}_b(\tau_c) + \mathbb{E}_c(\tau_a) = \mathbb{E}_a(\tau_c) + \mathbb{E}_c(\tau_b) + \mathbb{E}_b(\tau_a) \tag{4.2}$$

The "obvious" solution by reversing every path is wrong and here is why. The chain could make a walk *acababca* as a part of an *abca* cycle (which means we start at a, walk until we hit b for the first time, walk until we hit c for the first time and then until we hit a for the first time again). However if we look at this path in reverse, we see that the *acba* cycle is completed in three "steps" instead of seven. However when expectations are considered, for reversable chains things average out.

Hint: think of a chain starting at stationary distribution and then going to a (add this quantity to both sides of 4.2).

Note: The identity in the exercise can be generalized to cycles of any length.

In general, define the *reversal* of a Markov chain P as the Markov chain $\hat{\mathrm{P}}$ which for all $x, y \in \Omega$ satisfies

$$\pi(x)\mathrm{P}(x,y) = \pi(y)\widehat{\mathrm{P}}(y,x) \tag{4.3}$$

It is easy to check that $\hat{\mathrm{P}}$ has π as its stationary distribution (just like P does).

Exercise 4.3 *Show that for a* general *Markov chain*

$$\mathbb{E}_a(\tau_b) + \mathbb{E}_b(\tau_c) + \mathbb{E}_c(\tau_a) = \widehat{\mathbb{E}}_a(\tau_c) + \widehat{\mathbb{E}}_c(\tau_b) + \widehat{\mathbb{E}}_b(\tau_a) \tag{4.4}$$

Let us note that $\mathbb{E}_a(\tau_b)$ and $\mathbb{E}_b(\tau_a)$ can be very different for general Markov chains, including reversible ones. However, for certain types of graphs they are equal. A finite graph G is *transitive* if for any pair of vertices $x, y \in V(G)$ there exists a graph automorphism ψ of G with $\psi(x) = y$.

Exercise 4.4 *Prove that for a simple random walk on a transitive (connected) graph G, for any vertices $a, b \in V(G)$*

$$\mathbb{E}_a(\tau_b) = \mathbb{E}_b(\tau_a). \tag{4.5}$$

Many familiar graphs are transitive, e.g. $\mathbb{Z}_n^d$. The equality (4.5) is trivial if for any vertices $x, y \in V$ we can find an automorphism ψ which flips them: $\psi(x) = y$ and $\psi(y) = x$, which is the case for $\mathbb{Z}_n^d$. Hence an exercise:

Exercise 4.5 *Find the smallest transitive graph G such that there is a pair of vertices $x, y \in V(G)$ for which there is no automorphism ψ of G such that $\psi(x) = y$ and $\psi(y) = x$.*

Now we turn back to estimating mixing times for various Markov chains.

4.2 Review of Kantorovich Metric and Coupling

The *mixing time* $\tau_1(\epsilon)$ of a Markov chain is defined as:

$$\tau_1(\epsilon) = \inf_{x\in\Omega}(\inf\{t : \|\mathbf{P}^t(x,\cdot) - \pi\|_{TV} \leq \epsilon\}) \tag{4.6}$$

Proposition 4.6

$$\tau_1(\epsilon) = \inf_{\mu}(\inf\{t : \|\mathbf{P}^t(\mu,\cdot) - \pi\|_{TV} \leq \epsilon\}) \tag{4.7}$$

where μ is any distributions on Ω.

Proof: The proof is easy and was left as an exercise. ■

Notice that Lemma 2.11 implies

$$\|\mathrm{P}^{k\cdot\tau(1/4)}(x,\cdot) - \pi\|_{TV} \leq \overline{d}(k\cdot\tau_1(1/4)) \leq \overline{d}(\tau_1(1/4))^k \leq (2\cdot\frac{1}{4})^k,$$

and hence $\tau_1(2^{-k}) \leq k\tau_1(1/4)$. So in order to find the order of magnitude of the general mixing time $\tau_1(\epsilon)$, it is enough to consider $\tau_1(1/4)$. Changing ϵ only changes the mixing time by a constant factor. So, we also denote $\tau_1(1/4)$ by τ_1.

In Theorem 3.4 we showed that if d is a path metric on Ω, and the contraction condition

$$d_K(\delta_x\mathrm{P}, \delta_y\mathrm{P}) \leq e^{-\gamma}d(x,y) \tag{4.8}$$

holds for neighboring $x, y \in \Omega$, then (4.8) holds for any $x, y \in \Omega$. Now we will show that we can consider general distributions over Ω.

Lemma 4.7 *If (4.8) holds then for any measures μ, ν on (Ω, d)*

$$d_K(\mu\mathrm{P}, \nu\mathrm{P}) \leq e^{-\gamma}d_K(\mu,\nu) \tag{4.9}$$

Proof: Let $M = M(\cdot,\cdot)$ be a coupling that realizes Kantorovich metric $d_K(\mu,\nu)$. Also for each pair $x, y \in \Omega$ we have a coupling $A^{x,y}(\cdot,\cdot)$ which realizes $d_K(\delta_x\mathrm{P}, \delta_y\mathrm{P})$. Then combine all $A^{x,y}$ with weights $M(x,y)$ to get coupling $B(\cdot,\cdot) = \sum_{x,y} M(x,y)A^{x,y}(\cdot,\cdot)$. It is easy to check that B is a coupling of μP and νP and hence

$$d_K(\mu \mathrm{P}, \nu \mathrm{P}) \leq \sum_{x',y'} B(x',y')d(x',y') \quad = \quad \sum_{x,y} M(x,y) \sum_{x',y'} A^{x,y}(x',y')d(x',y') \tag{4.10}$$

$$= \sum_{x,y} M(x,y) d_K(\delta_x \mathrm{P}, \delta_y \mathrm{P}) \quad \leq \quad e^{-\gamma} d_K(\mu,\nu) \tag{4.11}$$

since $d_K(\delta_x\mathrm{P}, \delta_y\mathrm{P}) \leq e^{-\gamma} d(x,y)$ and M is coupling of μ and ν. ■

Iterating Lemma 4.7 completes the proof of Theorem 3.4. Combining this with (3.3), we have proved

Theorem 4.8 *Let d be a metric satisfying for every $x \neq y$, $d(x,y) \geq 1$. If $d_K(\delta_x\mathrm{P}, \delta_y\mathrm{P}) \leq e^{-\gamma}d(x,y)$ for all neighboring pairs $x \sim y$, then*

$$\tau_1(\epsilon) = \frac{1}{\gamma} \log \frac{Diam(\Omega)}{\epsilon}.$$

Example 4.9 Mixing time for random walk on the hypercube Coupling for random walk on the hypercube $\{0,1\}^d$ is the following. Choose a coordinate uniformly at random and update both bits to the same random value, hence possibly reducing the distance between paths.

Two neighbors x, y on the hypercube differ only in one coordinate. If one of the other $d-1$ coordinates is picked, the distance stays the same. If the coordinate in which x and y differ is picked, the distance decreases by 1. Hence

$$d_K(\delta_x\mathrm{P}, \delta_y\mathrm{P}) \leq 1 - \frac{1}{d} = e^{-\gamma}d(x,y) \tag{4.12}$$

This gives us $\gamma = O(1/d)$. Since the diameter of a hypercube is d, Theorem 4.8 gives

$$\tau_1(\epsilon) = O\left(\frac{1}{\gamma}\log \mathrm{Diam}(\Omega)\right) = O(d \log d), \tag{4.13}$$

a result we obtained directly by coupling in Example 2.12.

4.3 Applications: Proper Colorings, Ising Model

4.3.1 Graph Colorings

Recall that a coloring of a graph $G = (V,E)$ with q colors is a function $f : V \longrightarrow S = \{1, 2, \ldots, q\}$. A proper coloring is such that for no neighbors $u, v \in V$, equality $f(u) = f(v)$ holds. We are interested in sampling uniformly from proper colorings when they exist. It is slow to do so directly using rejection sampling by picking a random coloring and testing if it is proper since the size of the state space of proper colorings, $\Omega = S^V$, is exponential in $|V|$ (in case of trees, however, we can sample in linear time; we leave this as an exercise). Rather, we use the Markov Chain Monte Carlo algorithm introduced in section 3.2. Below, we bound the mixing time of this chain.

4.3.2 Mixing time for Glauber Dynamics on Graph Colorings

We briefly recall the Glauber dynamics on graph colorings. At each step of the chain, a vertex is chosen uniformly at random and the color of this vertex is updated. To update, a color is chosen uniformly at random from the allowable colors, which are those colors not seen among the neighbors of the chosen vertex. It can be easily checked that if $q > \Delta + 2$ then this Markov Chain is reversible and its stationary distribution is uniform over all proper colorings. Assume through the rest of the section that $q > \Delta + 2$.

We will use path coupling to bound the mixing time of this chain. Since Glauber dynamics dictates that the color of a vertex is updated by a color *not* among the neighboring colors, it is convenient to write $A(f, w)$ for the set of colors available for a vertex w and a given coloring f:

$$A(f, w) = \{j \in S : \text{for no } u \sim w, f(u) = j\}. \tag{4.14}$$

Write $n = |V|$. We use the usual Hamming distance for two colorings $f, g \in \Omega$:

$$d(f, g) = |\{v : f(v) \neq g(v)\}| \tag{4.15}$$

Note that $\text{Diam}(\Omega) = n$ and $d(f, g) \geq 1$ for $h \neq g$.

Let f and g be two colorings which agree everywhere except for a vertex v; this implies $d(f, g) = 1$. We describe how to simultaneously evolve the two chains so that separately they each have the correct dynamics.

First, we pick a vertex $w \in V$ uniformly at random. If $v \nsim w$, we update the two chains with the same color. This works because in both chains we pick among the available colors uniformly at random, and the available colors are the same for both chains: $A(f, w) = A(g, w)$. This case includes the $w = v$ case for which f and g become the same coloring and distance between them decreases by 1. Otherwise the distance stays the same. Note that $\mathbf{P}(v = w) = \frac{1}{n}$.

The other case is $v \sim w$. This happens with $\mathbf{P}(v \sim w) = \frac{\deg(v)}{n}$. Without loss of generality assume that $|A(g, w)| \leq |A(f, w)|$.

Choose a color c uniformly at random from $A(f, w)$, and use it to update f at w to get a new coloring $\tilde{f}$ with $\tilde{f}(w) = c$. If $c \neq g(v)$, then update the configuration g at w to the same color c: $\tilde{g}(w) = c = \tilde{f}(w)$. We subdivide the case $c = g(v)$ into subcases depending on whether or not $|A(f, w)| = |A(g, w)|$:

case	how to update g at w
$\lvert A(g, w)\rvert = \lvert A(f, w)\rvert$	set $\tilde{g}(w) = f(v)$
$\lvert A(g, w)\rvert < \lvert A(f, w)\rvert$	set $\tilde{g}(w) = Unif(A(g, w))$

Exercise 4.10 *Check that the above update rule chooses a color for $\tilde{g}(w)$ uniformly from $A(g, w)$.*

Note that the probability that the two configurations do not update to the same color is at most $1/|A(f, w)|$, which is bounded above by $1/(q - \Delta)$.

Given two colorings f and g which are at unit distance, we have constructed a coupling $(\tilde{f}, \tilde{g})$ of $\text{P}(f, \cdot)$ and $\text{P}(g, \cdot)$. The distance $d(\tilde{f}, \tilde{g})$ increases from 1 only in the case where a neighbor of v is updated and the updates are different in the two configurations. Also, the distance decreases to zero when v is selected to be updated. In all other cases the distance is unchanged. This shows that

$$d_K(\delta_f \text{P}, \delta_g \text{P}) \leq \mathbb{E} d(\tilde{f}, \tilde{g}) = 1 - \frac{1}{n} + \frac{\deg(v)}{n} \mathbb{E}\left(\frac{1}{|A(f, w)|}\right). \tag{4.16}$$

[The expectation is needed on the right because w is chosen at random.] This is bounded above by

$$1 - \frac{1}{n} + \frac{\Delta}{n} \cdot \frac{1}{q - \Delta} \tag{4.17}$$

and is less than 1 provided that $\frac{\Delta}{q-\Delta} < 1$, or $q > 2\Delta$. If this condition holds, then holding Δ and q constant, we obtain $\gamma = O(\frac{1}{n})$ and hence $\tau_1 = O(n \log n)$.

Let us emphasize that we created coupling for all possible pairs of adjacent colorings, not only the proper ones and that the distance is defined for any two elements of Ω. This is necessary since the path between two proper colorings f, g realizing the Hamming distance $d(f, g)$ may pass through colorings that are not proper. However, once a general coupling is constructed, we can apply it to proper colorings. We can assume that a chain starts at some proper coloring (which can be constructed in linear time for $q > \Delta + 2$); or we can also consider the more complex case with the chain starting at any colorings.

4.4 The Ising Model

Let $G = (V, E)$ be a graph of maximal degree Δ. To each vertex in V assign a spin from $\{+1, -1\}$. Then $\Omega = \{-1, 1\}^V$ is a state space of all spin configurations. Define the probability $\pi(\sigma)$ of a spin configuration $\sigma \in \Omega$ by

$$\pi(\sigma) = \frac{1}{Z(\beta)} \exp(\beta \sum_{u \sim v} \sigma(u)\sigma(v)) \tag{4.18}$$

where β is a parameter called inverse temperature, and $Z(\beta)$ is a normalization constant. The above is also called the *Gibbs distribution.* For this distribution, configurations with neighboring spins aligned are favored.

Glauber dynamics (a Markov chain with state space Ω) for the Ising Model are defined as follows. Given the current state σ, pick $w \in V$ uniformly at random and update the spin at w to $\tilde{\sigma}(w)$ according to the conditional probabilities for the Gibbs distribution given the spins at all other sides. We can easily obtain transition probabilities:

$$\mathbf{P}(\tilde{\sigma}(w) = 1) = \frac{e^{\beta s}}{e^{\beta s} + e^{-\beta s}} \tag{4.19}$$

where $s = s(w) = \sum_{u \sim w} \sigma(u)$.

It is easy to check that the Glauber dynamics for the Ising Model define a reversible Markov chain and that the Gibbs distribution is stationary for this chain.

Just as before, we want to find constraints on the parameters β, Δ of the model which will guarantee fast mixing. Consider two neighboring spin configurations $\sigma, \tau \in \Omega$ which differ at a single vertex v: $\sigma(v) = -1, \tau(v) = +1$. We will couple these configurations in the following fashion. If a vertex w picked on a next step of Glauber dynamics is different from v or $\Gamma(v) := \{w \sim v\}$, then update both chains to the same spin picked according to Gibbs distribution. If $v = w$, do the same. If $w \in \Gamma(v)$, probability distributions of new values of a spin at w are no longer the same for σ and τ. We can couple an update at w by choosing a uniform $U \in (0, 1)$ and setting $\tilde{\sigma}(w) = 1$ iff $U < \mathbf{P}(\tilde{\sigma}(w) = 1)$ and $\tilde{\tau}(w) = 1$ iff $U < \mathbf{P}(\tilde{\tau}(w) = 1)$. Noting that $s(\tau) = s(\sigma) + 2 = s + 2$ we obtain

$$\mathbf{P}(\tilde{\sigma}(v) \neq \tilde{\tau}(v)) = \frac{e^{\beta(s+2)}}{e^{\beta(s+2)} + e^{-\beta(s+2)}} - \frac{e^{\beta s}}{e^{\beta s} + e^{-\beta s}} = \frac{1}{2}(\tanh(\beta(s+2)) - \tanh(\beta s)) \leq \tanh(\beta) \tag{4.20}$$

where the last inequality follows by maximizing the expression $\tanh(\beta(s+2)) - \tanh(\beta s)$ as a function of s (the maximum occurs at $s = -1$).

Hence if we define the following metric on configurations in Ω: $d(\sigma, \tau) = \frac{1}{2} \sum_{u \in V} |\sigma(u) - \tau(u)|$ (normalized so that distance between neighboring configurations is 1), we obtain that if $d(\sigma, \tau) = 1$, then $\mathbb{E}d(\tilde{\sigma}, \tilde{\tau}) \leq 1 - \frac{1}{n} + \frac{\Delta}{n} \tanh(\beta)$. Theorem 4.8 tells us that if $\Delta \tanh(\beta) < 1$, the mixing time is $O(n \log n)$ as $\mathrm{Diam}(\Omega) = n$ and $\gamma = O(\frac{1}{n})$ when β and Δ are treated as constants. The above condition can be rewritten as $\beta < \frac{1}{2} \log(\frac{\Delta+1}{\Delta-1})$.

In the high temperature region, we can make approximation $\tanh(\beta) \sim \beta$ hence giving us simple condition for rapid mixing: $\beta\Delta < 1$.

Lecture 5: The Ising Model and the Bottleneck Ratio

5.1 Cycle Identity for Reversible Chains

Remember that in the reversible chains, we have:

Lemma 5.1
$$\mathbb{E}_a(\tau_b) + \mathbb{E}_b(\tau_c) + \mathbb{E}_c(\tau_a) = \mathbb{E}_c(\tau_b) + \mathbb{E}_b(\tau_a) + \mathbb{E}(\tau_c)$$

Proof: We can reword this lemma as

$$\mathbb{E}_a(\tau_{bca}) = \mathbb{E}_a(\tau_{cba}). \tag{5.1}$$

Let π be the stationary distribution. It turns out that it is much easier to start at stationarity, since it allows us to use reversibility easily. Define

$$\mathbb{E}_\pi(\tau_a) = \sum_x \pi(x)\mathbb{E}_x(\tau_a).$$

Adding $\mathbb{E}_\pi(\tau_a)$ to both sides of (5.1), we find it is enough to show that

$$\mathbb{E}_\pi(\tau_{abca}) = \mathbb{E}_\pi(\tau_{acba}).$$

In fact, we will show equality in distribution, not just expectation. Suppose s and t are finite strings with bits in Ω, say $s \in \Omega^m, t \in \Omega^n$ with $m \leq n$. We say that $s \leq t$ iff s sits inside t as a subsequence; that is there exist indices $1 \leq i_1, \ldots, i_m \leq n$ with $s(k) = t(i_k)$ for all $1 \leq k \leq m$. We have

$$\begin{aligned}
\mathbf{P}_\pi(\tau_{abca} > k) &= \mathbf{P}_\pi(abca \not\leq X_0 \ldots X_k) \\
&= \mathbf{P}_\pi(abca \not\leq X_k \ldots X_0) \\
&= \mathbf{P}_\pi(acba \not\leq X_0 \ldots X_k) \\
&= \mathbf{P}_\pi(\tau_{acba} > k).
\end{aligned}$$

■

Note: The same proof works for non-reversible chains. Again all we need is to check that $E_\pi(\tau_a) = \hat{E}_\pi(\tau_a)$ and that $\mathbf{P}_\pi(\tau_a > k) = \hat{\mathbf{P}}_\pi(\tau_a > k)$.

5.2 Path Coupling for Ising Examples

Return to the Ising model. For any configuration $\sigma \in \{-1, 1\}^V$, the Ising distribution is

$$\pi(\sigma) = \frac{1}{Z(\beta)} e^{\beta \sum_{u \sim v} J_{uv}\sigma(u)\sigma(v)} \tag{5.2}$$

The parameter $J_{uv} \geq 0$ and $\beta \geq 0$. Usually, $J_{uv} \equiv J$ for some constant J, so we can replace βJ by β.

Example 5.2 Ising model on $\mathbb{Z}_n$

Recall from (4.20), if we start at configuration σ and τ which differs only at a vertex w. And if $v \sim w$, then,

$$\mathbf{P}(\tilde{\sigma}(v)) \neq \tilde{\tau}(v) = \frac{1}{2}(\tanh(\beta(s+2)) - \tanh(\beta s)) \leq \tanh(\beta), \tag{5.3}$$

where $s = \sum_{u \sim v} \sigma(u)$. Hence our analysis yielded that

$$d_K(\delta_\sigma P, \delta_\tau P) \leq \mathbb{E}d(\tilde{\sigma}, \tilde{\tau}) \leq 1 - \frac{1}{n} + \frac{\Delta}{n}\tanh(\beta).$$

Recall that the inequality in (5.3) arises when we maximize and take $s = -1$. In the case of $\mathbb{Z}_n$, however, the only possible values for s are 0 and ± 2. In this case the maximum occurs at $s \in \{0, -2\}$. Hence we obtain instead that:

$$d_K(\delta_\sigma P, \delta_\tau P) \leq 1 - \frac{1}{n} + \frac{\tanh(2\beta)}{n} \leq 1 - \frac{C(\beta)}{n}. \tag{5.4}$$

By this we find the mixing time $\tau_1 \leq C(\beta) n \log(n)$.

Example 5.3 Ising model on K_n, the complete graph without loops.

We take $J = \frac{1}{n}$, then $\pi(\sigma) = \frac{e^{\frac{\beta}{n}\sum_{u \sim v} \sigma_u \sigma_v}}{Z(\beta)}$. So that $\Delta = n-1$, $\beta \to \frac{\beta}{n}$, and we obtain that:

$$d_K(\delta_\sigma P, \delta_\tau P) \leq 1 - \frac{1}{n} + \frac{n-1}{n}\tanh(\frac{\beta}{n}) \tag{5.5}$$

So taking $\beta < 1$ and a Taylor expansion of tanh, we find that the mixing time $\tau_1 \leq C(\beta) n \log(n)$.

5.3 Bottleneck Ratio, Conductance, Cheeger Constant

As usual, we work in the setting of a irreducible and aperiodic Markov Chain on a finite state space Ω, with transition probabilities P and stationary distribution π. We define the *edge measure* Q via:

$$Q(x, y) = \pi(x)p(x, y), \qquad Q(A, B) = \sum_{x \in A, y \in B} Q(x, y) \tag{5.6}$$

In particular, $Q(S, S^c)$ gives the probability of moving from S to S^c in one step starting from stationarity.

Exercise 5.4 *Show that for any $S \subset \Omega$, we have*

$$Q(S, S^c) = Q(S^c, S).$$

The result is trivial in the reversible case, but true in general.

The *bottleneck ratio* of the set S is given by

$$\Phi(S) = \frac{Q(S, S^c)}{\pi(S)}. \tag{5.7}$$

The bottleneck ratio of the whole chain is defined by:

$$\Phi_* = \min_{\pi(S) \le \frac{1}{2}} \Phi(S). \tag{5.8}$$

To see how this is connected to mixing, consider the measure

$$\mu = \pi(\cdot|S), \qquad \mu(A) = \frac{\pi_S(A)}{\pi(S)} = \frac{\pi(A \cap S)}{\pi(S)}$$

From a version of the definition of the total variation norm, we have

$$\pi(S)\|\mu\mathrm{P} - \mu\|_{\mathrm{TV}} = \sum_{y:\pi_S\mathrm{P}(y) \ge \pi_S(y)} (\pi_S\mathrm{P})(y) - \pi_S(y) \tag{5.9}$$

Since π_S vanishes on S^c, the difference in (5.9) is nonnegative on S^c. Moreover, for $y \in S$ we have

$$(\pi_S\mathrm{P})(y) = \sum_x \pi_S(x)\mathrm{P}(x,y) = \sum_{x \in S} \pi(x)\mathrm{P}(x,y) \le \sum_x \pi(x)\mathrm{P}(x,y) = \pi(y) = \pi_S(y),$$

so the difference is nonpositive on S. Thus the sum in (5.9) is taken over $y \in S^c$. Hence

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

It follows that

$$\|\mu\mathrm{P} - \mu\|_{\mathrm{TV}} \le \Phi(S).$$

Recalling that convolution decreases the TV norm, we have for any $t \ge 0$

$$\|\mu\mathrm{P}^t - \mu\mathrm{P}^{t+1}\|_{\mathrm{TV}} \le \Phi(S),$$

so

$$\|\mu\mathrm{P}^t - \mu\|_{\mathrm{TV}} \le t\Phi(S).$$

Now assume that $\pi(S) \le \frac{1}{2}$, so that $\|\mu - \pi\|_{TV} \ge \frac{1}{2}$. Taking $t = \tau_1(\frac{1}{4})$, we have by the definition of τ_1

$$\frac{1}{2} \le \|\mu - \pi\|_{TV} \le \frac{1}{4} + t\Phi(S).$$

Minimizing over S, we have proved a lower bound on the mixing time:

Lemma 5.5

$$\tau_1(\frac{1}{4}) \ge \frac{1}{4\Phi_*} \tag{5.10}$$

5.4 Checking for Bottlenecks

Example 5.6 We return to the Ising Model on the complete graph K_n. Still, we take $J = \frac{1}{n}$. We want to take $n \to \infty$ and check for bottlenecks. Take $k \sim \alpha n$ and consider the set

$$S_k = \{\sigma \mid \#\{v|\sigma_v = 1\} = k\}.$$

By counting we obtain that:

$$\pi(S_k) = \frac{1}{Z(\beta)}\binom{n}{k}\exp[\frac{\beta}{n}[\binom{k}{2} + \binom{n-k}{2} - k(n-k)]] =: \frac{A_k}{Z(\beta)}.$$

Taking a log and applying Stirling's formula, we obtain

$$\log A_k \sim nh(\alpha) + ng(\alpha) \tag{5.11}$$

where

$$h(\alpha) = \alpha\log(\frac{1}{\alpha}) + (1-\alpha)\log(\frac{1}{1-\alpha});$$
$$g(\alpha) = \beta(\frac{(1-2\alpha)^2}{2}).$$

If we do some calculus we find that $h'(\frac{1}{2}) = g'(\frac{1}{2}) = 0$, $h''(\frac{1}{2}) = -4, g''(\frac{1}{2}) = 4\beta$. Hence $\alpha = \frac{1}{2}$ is a critical point of $h + g$, whereby it is a (local) maximum or minimum depending on the value of β. Let us take the case $\beta > 1$, so we have a minimum at $\alpha = \frac{1}{2}$. In this case we have a bottleneck. Define

$$S = \{\sigma \mid \sum_u \sigma_u < 0\}$$

By symmetry, $\pi(S) \leq \frac{1}{2}$. For simplicity think of $k = \lfloor\frac{n}{2}\rfloor$. Observe that the only way to get from S to S^c is through S_k, since we are only allowed to change one spin at a time. Thus $\mathbb{Q}(S, S^c) = \lceil\frac{n}{2}\rceil/n\pi(S_k)$ and $\pi(S) = \sum_{j<k}\pi(S_j)$. We recall that at $\alpha = \frac{1}{2}$, we didn't have a maximum since we took $\beta > 1$, so after clearing the logs we get a negative exponential:

$$\Phi_* \leq e^{-nC(\beta)} \tag{5.12}$$

By Lemma 5.5 we conclude that the mixing time is exponential.

Note: When $\beta \in (0, 1)$, we get a lower bound $n\log n$ for the mixing time. The rate of mixing time goes from this to exponential at the critical point $\beta = 1$.

Lecture 6: Introduction to block dynamics

6.1 Expectation of hitting times

We first revisit Exercises 4.28 and 4.29.

Proposition 6.1 *There exists a transitive graph G and a pair of vertices $x,\ y \in V(G)$, for which there is no automorphism ψ of G satisfying $\psi(x) = y$ and $\psi(y) = x$.*

Proof: The simplest example we know, suggested to us by Luigi, is a tetrahedron with each corner replaced by a triangle. (If multiple edges are allowed, one can simply take a hexagon with three nonadjacent edges doubled). An example where the vertices x, y that cannot be flipped are adjacent was suggested by Ander Holroyd: The snub cube.

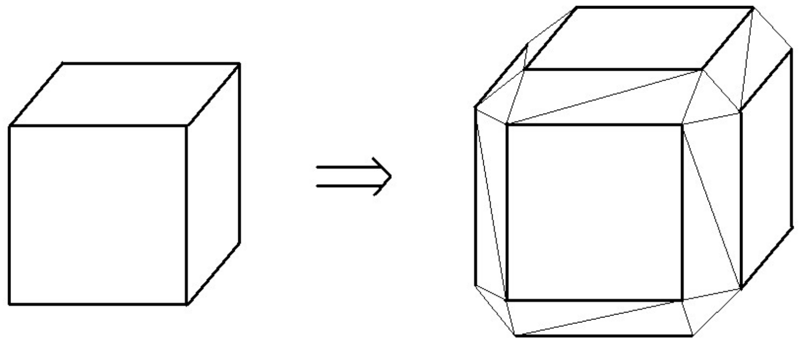

Figure 6.1: Construction of the snub cube.

We describe how to construct it. Start from a cube, and detach its six faces. So we have six separate faces with 24 vertices. We add some other lines between these 24 vertices. Now we have a polyhedron with 6 squares and 32 triangular faces. This figure is transitive . But for any two neighbors x and y, there is no automorphism ψ of G such that $\psi(x) = y$ and $\psi(y) = x$.

∎

Proposition 6.2 *For a simple random walk on a transitive (connected) graph G, for any vertices $a,\ b \in V(G)$ we have*

$$\mathbb{E}_a(\tau_b) = \mathbb{E}_b(\tau_a) \tag{6.1}$$

Proof: Let ψ be an automorphism such that $\psi(a) = b$. Let $a_0 = a$, $a_j = \psi^j(a_0)$ for $j \geq 1$, where ψ^j denotes the j-th iterate of ψ. The sequence $a_0, a_1, \dots$ will return to a_0 eventually, say $a_m = a_0$, $m > 0$. Because the automorphism ψ^j takes a, b to a_j, a_{j+1}, so for any j:

$$\mathbb{E}_{a_j}(\tau_{a_{j+1}}) = \mathbb{E}_a(\tau_b), \tag{6.2}$$

Summing over j from 0 to $m - 1$ we obtain

$$\mathbb{E}_{a_0}(\tau_{a_1 a_2 \dots a_{m-1} a_0}) = m\mathbb{E}_a(\tau_b) \tag{6.3}$$

For the same reason yield:

$$\mathbb{E}_{a_0}(\tau_{a_{m-1}a_{m-2}\dots a_1a_0}) = m\mathbb{E}_b(\tau_a) \tag{6.4}$$

From the same proof of equation (5.1), we get the left hand side of equation (6.3) and (6.4) are the same. So we have proved

$$\mathbb{E}_a(\tau_b) = \mathbb{E}_b(\tau_a) \tag{6.5}$$

■

In fact, more is true: τ_a when the chain starts at b has the same distribution as τ_b when the chain starts at a.

Exercise 6.3 *Show that* $\mathbf{P}_a(\tau_b > k) = \mathbf{P}_b(\tau_a > k)$*,* $k = 0, 1, \dots$ *for simple random walk on a transitive graph.*

Exercise 6.4 *On an* $n \times n$ *square, with edges inherited from* $\mathbb{Z}^2$*, let* a *be the lower-left corner and* b *the upper-right corner. Give a simple proof that*

$$\mathbb{E}_a(\tau_b) = O(n^2 \log n) \tag{6.6}$$

In order to solve this problem and the following mixing time problems, we now introduce a method called block dynamics.

Recall we have shown that the mixing time for the Ising model on the cycle of length n is

$$\tau_1(1/4) = O_\beta(n \log n), \tag{6.7}$$

i.e. the constant in the O may depend on β.

Note that lazy simple random walk on the d-cube $\{0,1\}^d$ can be regarded as Glauber dynamics on a graph with no edges and d vertices. In such graphs, the ising distribution is uniform. Generally, on any graph of d vertices, the Ising distribution is

$$\pi(\sigma) = \frac{1}{Z(\beta)} e^{\beta \sum_{u \sim v} \sigma_u \sigma_v} \tag{6.8}$$

If we let $\beta \downarrow 0$, then the Glauber dynamics degenerate into simple random walk on the d-cube. Thus the Ising model at infinite temperature is just like simple random walk on the d-cube.

We now consider Ising model on a ladder graph. To establish the mixing time $O_\beta(n \log n)$, the contraction method we learned in lecture 5 won't work anymore. We will use block dynamics to get these kinds of mixing times.

In the ladder example, we regard each pair of 2-spins in a column as a 4-spin, so the whole graph may be regarded as a 4-spin system on a one-dimensional line,

Our discussion below applies to general one-dimensional systems with local interactions. For such systems, we have exponential decay of spatial correlations, which will get us the mixing times. The exponential decay of spatial correlations means that there exists $0 < \theta < 1$, such that for any functions f, g and any $\ell \geq 0$, we have:

$$\mathrm{Cov}(f(s_j, j \leq k), g(s_j, j > k + \ell)) \leq \theta^\ell \|f\|_2 \|g\|_2 \tag{6.9}$$

where s_j is the spin at the site j.

6.2 Block dynamics

Consider a one-dimensional system with configuration $\{\sigma_j\}_{j=0}^{N-1}$. Fix a large integer $b = b(\beta)$, the *block size*. Chose uniformly at random an integer $w \in [-b, N-1]$ and *update* the block $\{\sigma_j \ : \ w \le j \le w+b\}$ by erasing the spins within the block and replacing them with a configuration chosen according to the conditional probability determined by the spins at the neighboring sites σ_{w-1} and σ_{w+b+1}. (For $w < 0$ we only update the interval $[0, w+b]$. The same applied to $w > N-b-1$.) We call this a *heat-bath block update.*

Theorem 6.5 *If b is large enough, the block dynamics give a contraction of the Hamming metric.*

We will prove this for *monotone systems* like the Ising model. A monotone system is a Markov chain on a partially ordered set with the property that for any pair of states $x \le y$ there exist random variables $X_1 \le Y_1$ such that for every state z

$$\mathbf{P}[X_1 = z] = p(x,z), \quad \mathbf{P}[Y_1 = z] = p(y,z).$$

In words, if two copies of the chain are started from states $x \le y$, we can couple them so that the one which started in the lower state always remains in a lower state. Last time, we checked that the single site dynamics for the Ising model are a monotone system under the coordinatewise partial ordering: $\sigma \le \tau$ if $\sigma_x \le \tau_x$ for every x.

Lemma 6.6 *If the single-site dynamics within each block are irreducible and aperiodic, then the block dynamics for the Ising model are also a monotone system.*

Proof: For each block, the distribution in the block conditioned on the boundary values is stationary for single-site dynamics within the block. Since these dynamics are irreducible and aperiodic, they converge to the block update. Since the single-site dynamics are monotone, we can couple them so that the limiting distribution of (X_t, Y_t) is supported on $\{(x,y) \ : \ x \le y\}$. Therefore the block dynamics are also monotone. ∎

Proof:[Proof of Theorem 6.5] Recall that it is enough to check the contraction for configurations σ and τ differing only on a single site w. There are $N+b$ blocks altogether, of which $b+1$ contain w. If we update one of these $b+1$ blocks, we can couple the updates to remove the defect at w. Moreover if we update a block not containing w and not adjacent to w, we can couple the updates so as not to introduce any additional defects. The only situation in which we might introduce a new defect is when we're updating one of the two blocks adjacent to w. But in this case the exponential decay of spacial correlations (6.9) implies that the expected number of new defects created is bounded independent of the block size. Hence

$$d_K(\delta_\sigma \mathrm{P}_B, \delta_\tau \mathrm{P}_B) = 1 - \frac{b+1}{N+b} + \frac{2C(\beta)}{N+b}.$$

Taking b sufficiently large we can ensure that this is $\le 1 - \frac{\tilde{C}(\beta)}{N}$, which gives $O_\beta(n \log n)$ mixing. ∎

6.3 Strong Spacial Mixing

In higher-dimensional systems, the condition (6.9) does not always hold, and in order to prove fast mixing we impose it as a hypothesis; this is the hypothesis of *strong spacial mixing.* Under strong spacial mixing, the block dynamics for the Ising model on a subset V of $\mathbb{Z}^d$ mix in time $O(|V|\log|V|)$. Here the blocks are boxes $[w, w+b-1]^d$. Take $V = [0,n]^d$ as an example. For configurations σ and τ differing at a single vertex v on the boundary of a box Λ, we have by strong spacial mixing

$$\left|\mathbb{E}_{\sigma|\partial\Lambda}\sigma(u) - \mathbb{E}_{\tau|\partial\Lambda}\tau(u)\right| \leq c_1 e^{-c_2(\beta)||u-v||}.$$

The effect of a block update on the Kantorovich distance is therefore

$$d_K(\delta_\sigma \mathrm{P}_B, \delta_\tau \mathrm{P}_B) = 1 - \frac{b^d}{n^d} + \frac{b^{d-1}c(\beta)}{n^d},$$

which for sufficiently large b can be taken $\leq 1 - \frac{\tilde{c}(\beta)}{n^d}$, giving mixing time $O_\beta(n^d \log n)$.

Lecture 7: Cut-off and Eigenvalues of Reversible Chains

7.1 Cut-off for the Hypercube

We'll consider the hypercube $\Omega = \{-1, 1\}^d$. We used $\{0, 1\}^d$ before, but this representation is better suited for the following calculations. The random walk on the hypercube was explained in detail in Example 2.12. Recall, from $(X_1, \ldots, X_d)$ one chooses a coordinate uniformly and replaces it by ± 1 with equal probability $\frac{1}{2}$. In Example 4.9, we found an upper bound on the mixing time $\tau_1(\epsilon)$ of order $d \log d$. Now we'll look for a lower bound. By the definition of the mixing time, we have

$$\tau_1(\epsilon) = \max_{x \in \Omega} \min\{t : \|\mathrm{P}^t(x, \cdot) - \pi(\cdot)\|_{TV} \le \epsilon\} \ge \min\{t : |\mathrm{P}^t(x, A) - \pi(A)| \le \epsilon\} \tag{7.1}$$

for all $A \subset \Omega$ and all $x \in \Omega$. A lower bound hence can be achieved by fixing an initial state x and finding a set A such that if we start X at x, start Y under the stationary distribution π, and run both chains until time t, the resulting distributions differ on A with probability greater than ϵ:

Definition 7.1 *Let $\mu_t = \delta_{\mathbf{1}} \mathrm{P}^t$, i.e. μ_t is the distribution of our Markov chain if we start with uniform configuration $\mathbf{1}$ and run it t steps.*

Recall that the stationary distribution of the random walk on the hypercube is uniform (use for example that P is symmetric). We'll look for a set A which will make $|\mathrm{P}^t(x, A) - \pi(A)|$ as big as possible and thus provide us with a lower bound on the total variation distance. The random walk on the hypercube is identifiable with the Glauber dynamics for the Ising model at infinite temperature (i.e. $\beta = 0$). As above, let X_i be the ith configuration. Let

$$S_d = \sum_{i=1}^{d} X_i. \tag{7.2}$$

Since for each μ_t, the distribution only depend on the sum of the coordinates. So we only have to look for sets given in terms of S_d. Let's try to get a first idea of the distribution of S_d under the measures π and μ_t. One can easily calculate

$$\mathbb{E}_\pi(S_d) = 0 \tag{7.3}$$

$$\mathbf{V}ar_\pi(S_d) = \mathbb{E}_\pi S_d^2 = \mathbb{E}_\pi\Big(\sum_{i=1}^{d} X_i^2 + \sum_{i,j=1, i \ne j}^{d} X_i X_j\Big) = d + 0 = d. \tag{7.4}$$

As the probability that we haven't chosen the i-th coordinate up to time t is $(1 - \frac{1}{d})^t$ we obtain (as we started with configuration $+1$ in each coordinate)

$$\begin{aligned} \int X_i d\mu_t &= 1 \cdot \Big(1 - \frac{1}{d}\Big)^t + \sum_{j=1}^{t} \mathbf{P}_{\mu_t}(\text{“coordinate } i \text{ is first chosen at time } j\text{”}) \cdot \Big(\frac{1}{2}(+1) + \frac{1}{2}(-1)\Big) \\ &= \Big(1 - \frac{1}{d}\Big)^t \end{aligned} \tag{7.5}$$

and therefore

$$\int S_d d\mu_t = \Big(1 - \frac{1}{d}\Big)^t d. \tag{7.6}$$

As we'll see later, $\tau_1 \sim \frac{1}{2}d\log d$. So we set $t = \frac{1-\epsilon}{2}d\log d$, and try to find the lower bound for $\|\mu_t - \pi\|_{TV}$. Using $(1-\frac{1}{d})^t \sim e^{-\frac{t}{d}}$, we obtain:

$$\int S_d d\mu_t = \mathbb{E}_{\mu_t} S_d \sim d^{\frac{1+\epsilon}{2}}. \tag{7.7}$$

Moreover, the variables X_i are negatively correlated under μ_t due to similar calculations

$$\int X_i X_j d\mu_t = \Big(1 - \frac{2}{d}\Big)^t \text{ for } i \neq j \Rightarrow \mathrm{cov}_{\mu_t}(X_i, X_j) < 0, \tag{7.8}$$

where the first equation results from considering the case that we haven't touched either of the variables i and j so far. Having chosen the representation $\Omega = \{-1,1\}^d$ for the hypercube we directly obtain $\int X_i^2 d\mu_t = 1$ and thus by the negative correlation of the coordinates

$$\mathbf{V}ar_{\mu_t}(S_d) \leq \sum_{i=1}^{d} \mathbf{V}ar_{\mu_t}(X_i) \leq d. \tag{7.9}$$

The idea how to choose A now is to separate both measures, i.e. in terms of the following figure to choose it in such a way that the area over this set becomes small under the first graph while it becomes big under the second one: The left resp. right graph represents the distribution of S_d under π resp. μ_t:

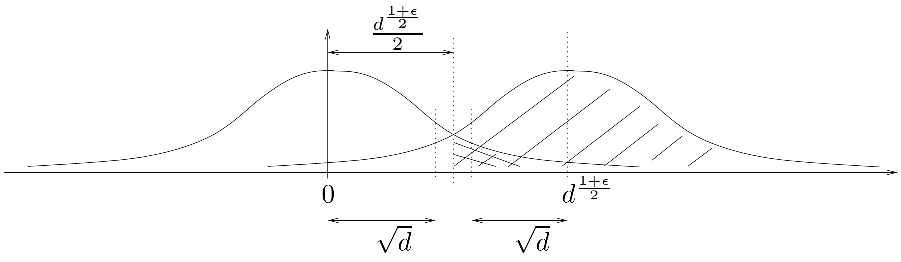

Figure 7.1: Separating measures for lower bounds on the mixing time

For these reasons we'll choose

$$A = \Big\{x \mid S_d \geq \frac{d^{\frac{1+\epsilon}{2}}}{2}\Big\}. \tag{7.10}$$

Chebyshev's inequality gives

$$\pi(A) \leq \pi\Big(|S_d| \geq \frac{d^{\frac{1+\epsilon}{2}}}{2}\Big) \leq 4d^{-\epsilon}. \tag{7.11}$$

Further calculations, leading to

$$\mu_t(A^c) \leq 4d^{-\epsilon} \tag{7.12}$$

finally result in

$$\|\mu_t - \pi\|_{TV} \geq \mu_t(A) - \pi(A) \geq 1 - 8d^{-\epsilon}[1 + o(1)]. \tag{7.13}$$

We add this $o(1)$ because we did some approximations in the computation. But this $o(1)$ could be skipped.

Exercise 7.2 *Check that the approximations don't matter.*

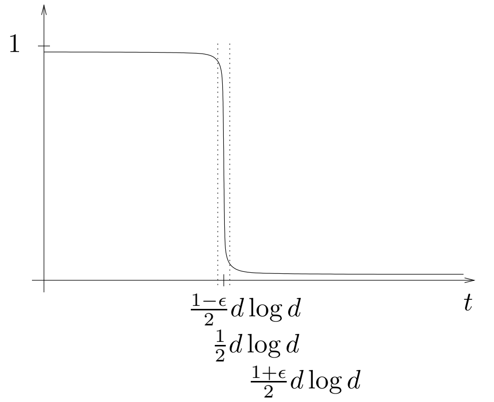

Figure 7.2: Cut-off

Remark 7.3 *In fact, $\frac{1}{2}d\log d$ is what you need to mix. When t get a little smaller than $\frac{1}{2}d\log d$, the total variation distance $\|\mu_t - \pi\|_{TV}$ get quickly close to 1. This behaviour follows from our preceding calculations taking ϵ small and d big.*

This phenomenon is known under the name "cut-off". It appears for instance when shuffling cards, where a deck suddenly turns from being mostly ordered to random (see Diaconis' book for more). It is still an open question when exactly this cut-off phenomenon occurs. For example it doesn't happen for the cycle (here the distribution of the chain approaches the stationary distribution a bit further in every time-step).

Remark 7.4 *Equations (7.1) and (7.13) yield $\tau_1(\epsilon) \geq \frac{1}{2}d\log d$. Together with earlier results for the hypercube we therefore obtain that $\tau_1(\epsilon) = \Theta(d\log d)$ is the right order of magnitude for the mixing time of random walk on the hypercube.*

Remark 7.5 *For the lazy random walk on transitive graphs, Yuval Peres conjectures that*

$$[1-\lambda_2^{(n)}]\tau_1^{(n)} \to \infty \text{ for } n \to \infty \tag{7.14}$$

is necessary (easy to check) and sufficient for cut-off, where $\lambda_2^{(n)}$ is the 2^{nd} largest eigenvalue of the chain on n vertices. One can show in general that for irreducible chains (such as the lazy random walk) the eigenvalues can be ordered as follows $\lambda_1 = 1 > \lambda_2 > \lambda_3 > \ldots > 0$. The difference $1-\lambda_2$ (resp. more generally $1-\max\limits_{i\geq 2}\{|\lambda_i|\}$) is called spectral gap.

7.2 Eigenvalues of Reversible Chains

In this section we'll examine the eigenvalue spectrum of the transition matrix P of a reversible chain, and establish a first connection between the spectral gap and the mixing time of such a chain. Recall that a reversible chain is one satisfying $\pi(x)\mathrm{P}(x,y) = \pi(y)\mathrm{P}(y,x)$ for all states $x,y \in \Omega$. Note that we agreed only to look at irreducible and aperiodic chains. The irreducibility together with the reversibility now imply $\pi(x) > 0$ for all $x \in \Omega$ (or use the fact that the entries of the stationary distribution of an irreducible aperiodic chain, run on a finite space, are all positive due to the positive recurrence of the chain).

Instead of directly considering P, let us first consider the symmetric matrix

$$A(x,y) = \sqrt{\frac{\pi(x)}{\pi(y)}}\mathrm{P}(x,y).$$

The fact that $A(x,y) = A(y,x)$ follows directly from the reversibility. As A is symmetric we can take advantage of the spectral theorem, which yields an orthonormal basis of real eigenvectors φ_j with real eigenvalues λ_j.

As one can directly check, $\varphi_1 = \sqrt{\pi(\cdot)}$ is an eigenvector of A with corresponding eigenvalue $\lambda_1 = 1$. Let (with a slight abuse of notation) π also denote the diagonal matrix with entries $\pi(x)$. Then

$$A = \pi^{\frac{1}{2}}\mathrm{P}\pi^{-\frac{1}{2}}$$

Setting

$$f_j = \pi^{-\frac{1}{2}}\varphi_j \tag{7.15}$$

we compute

$$\mathrm{P}f_j = \mathrm{P}\pi^{-\frac{1}{2}}\varphi_j = \pi^{-\frac{1}{2}}A\varphi_j = \pi^{-\frac{1}{2}}\lambda_j\varphi_j = \lambda_j f_j.$$

Thus P has eigenvectors f_j and eigenvalues λ_j. The disadvantage of this representation is that the eigenfunctions are not necessarily orthonormal. Therefore we will introduce a new inner product under which the eigenfunctions $f_j = \pi^{-\frac{1}{2}}\varphi_j$ will be orthonormal again. Let $< \cdot, \cdot >$ denote the inner product we seek and $< \cdot, \cdot >_{\mathbf{R}^{|\Omega|}}$ the usual scalar product on $\mathbf{R}^{|\Omega|}$. We have $\delta_{ij} =< \varphi_i, \varphi_j >_{\mathbf{R}^{|\Omega|}} =< \pi^{\frac{1}{2}}f_i, \pi^{\frac{1}{2}}f_j >_{\mathbf{R}^{|\Omega|}} =< f_i, \pi f_j >_{\mathbf{R}^{|\Omega|}}$. Hence by introducing

$$< f, g >= \int f(x)g(x)d\pi(x) = \sum_{x\in\Omega} f(x)g(x)\pi(x) =< f, \pi g >_{\mathbf{R}^{|\Omega|}} \tag{7.16}$$

we obtain a new inner product on $\mathbf{R}^{|\Omega|}$ under which the f_j form a basis of orthonormal eigenvectors of P. Note that (7.16) proves, together with the fact that π is a symmetric positive definite matrix, that $< \cdot, \cdot >$ defines indeed a new inner product on $\mathbf{R}^{|\Omega|}$. The transition matrix P is self-adjoint for this inner product, i.e. $< \mathrm{P}f, g >=< f, \mathrm{P}g >$, as can easily be shown by checking it for the basis of eigenfunctions.
Considering $(\mathbf{R}^{|\Omega|}, < \cdot, \cdot >)$ with its orthonormal basis of eigenfunctions $\{f_j\}_{j\in\{1,\dots,|\Omega|\}}$ we can write δ_y via basis-decomposition as

$$\delta_y = \sum_j < \delta_y, f_j > f_j = \sum_j f_j(y)\pi(y)\cdot f_j \tag{7.17}$$

and noting that $\mathrm{P}^t(x,y) = (\mathrm{P}^t\delta_y)(x)$ this yields

$$P^t(x,y) = \sum_j f_j(y)\pi(y)\cdot\lambda_j^t f_j(x) \tag{7.18}$$

$$\iff \frac{\mathrm{P}^t(x,y)}{\pi(y)} = \sum_{j\geq 1} f_j(x)f_j(y)\lambda_j^t = 1 + \sum_{j\geq 2} f_j(x)f_j(y)\lambda_j^t, \tag{7.19}$$

where we used that $\mathrm{P}\mathbf{1} = \mathbf{1}$ and $f_1 = \mathbf{1}$ as $f_1 = \pi^{-\frac{1}{2}}\varphi_1 = \mathbf{1}$ (see (7.15) and before) with corresponding eigenvalue $\lambda_1 = 1$.

In general (i.e. even without assuming reversibility) we have $|\lambda_j| \leq 1$:
Indeed, for all functions f we have $\| \mathrm{P}f \|_\infty = \max_x |\sum_y p(x,y)f(y)| \leq \| f \|_\infty$. Hence if $\mathrm{P}f = \lambda f$ we obtain $\| \mathrm{P}f \|_\infty = |\lambda| \, \| f \|_\infty \leq \| f \|_\infty$ i.e. $|\lambda| \leq 1$.

Exercise 7.6 *For an irreducible aperiodic chain we have $|\lambda_j| < 1$ for all $j > 1$.*
Hint: See it directly or use convergence theorem.

Continuing our calculation by setting

$$|\lambda_*| = \max_{j\geq 2} |\lambda_j|$$

we obtain

$$\Big|\frac{\mathrm{P}^t(x,y)}{\pi(y)} - 1\Big| \leq \sum_{j\geq 2} |f_j(x)f_j(y)||\lambda_*|^t \leq \sqrt{\sum_{j\geq 2} f_j^2(x) \sum_{j\geq 2} f_j^2(y)}|\lambda_*|^t, \tag{7.20}$$

where the last equation follows by Cauchy-Schwarz. Using the definition of $< \cdot, \cdot >$ and (7.17) we also get

$$\pi(x) =< \delta_x, \delta_x >=< \sum_j f_j(x)\pi(x) \cdot f_j, \sum_j f_j(x)\pi(x) \cdot f_j >$$

and together with the orthonormality of the f_j this gives

$$\pi(x) = \pi(x)^2 \sum_{j\geq 1} f_j(x)^2 \Rightarrow \sum_{j\leq 2} f_j(x)^2 \leq \frac{1}{\pi(x)}.$$

Inserting this in (7.20) we obtain

$$\Big|\frac{\mathrm{P}^t(x,y)}{\pi(y)} - 1\Big| \leq \frac{1}{\sqrt{\pi(x)\pi(y)}}|\lambda_*|^t \leq \frac{1}{\pi_{min}}|\lambda_*|^t$$

with $\pi_{min} = \min_{x\in\Omega} \pi(x)$. Hence we obtain for the weighted average that if

$$|\lambda_*|^t \leq \epsilon \cdot \pi_{min} \Rightarrow \Big|\frac{\mathrm{P}^t(x,y)}{\pi(y)} - 1\Big| \leq \epsilon$$

and for the total variation distance

$$\|\mathrm{P}^t(x,\cdot) - \pi(\cdot)\|_{TV} = \frac{1}{2}\sum_{y\in\Omega} |P^t(x,y) - \pi(y)| \leq \frac{1}{2}\sum_{y\in\Omega} \epsilon\pi(y) < \epsilon.$$

This finally gives an estimate for the mixing time of a reversible Markov chain in terms of its stationary distribution π and the eigenvalues of its transition matrix P:

Theorem 7.7 *For a reversible Markov chain we have*

$$\tau_1(\epsilon) \leq \frac{\log(\epsilon \cdot \pi_{min})}{\log(|\lambda_*|)} = \frac{\log\left(\frac{1}{\epsilon\cdot\pi_{min}}\right)}{\log\left(\frac{1}{|\lambda_*|}\right)} \tag{7.21}$$

where $\pi_{min} = \min_{x\in\Omega} \pi(x)$ and $|\lambda_*| = \max_{j\geq 2} |\lambda_j|$.

Remark 7.8 *Observe that when $\lambda_* \to 1$ then $\log\left(\frac{1}{|\lambda_*|}\right) \simeq 1 - |\lambda_*|$ which gives an upper bound on the mixímg time in terms of the spectral gap $1 - |\lambda_*|$. In fact, applications show that this relation strongly depends on the model under consideration.*

Exercise 7.9 *Write the eigenvalues and the eigenfunctions for the Lazy SRW on the cycle $\mathbb{Z}_n$ explicitly. Check that $1 - \lambda_2 \asymp \frac{1}{n^2}$ ($\asymp$ signifies that the ratio is bounded by positive constants).*
Hint: A solution can be found in Feller's book, using symmetric functions.

Example 7.10 We'll consider the hypercube $\{-1,1\}^d$ again. In this representation the eigenfunctions are the following functions as we'll prove in an instant:

$$f_{\mathcal{S}}(x) = \prod_{j\in\mathcal{S}} x_j, \qquad \mathcal{S} \subset \{1,\dots,d\},$$

where $f_\emptyset = \mathbf{1}$. Indeed, let $\mathrm{P} f_{\mathcal{S}}(x) = \sum_y p(x,y) f_{\mathcal{S}}(y) = \mathbb{E}_x f_{\mathcal{S}}(\tilde{X})$, where $\tilde{X}$ is the state of the chain, started at x and run one step. As one step consists in replacing a uniformly chosen coordinate with equal probability via ± 1, we get (observing that $f_{\mathcal{S}}(x)$ only changes to $-f_{\mathcal{S}}(x)$ if a coordinate of the set $\mathcal{S}$ is chosen and changed with probability $\frac{1}{2}$)

$$\begin{aligned}
\mathbb{E}_x f_{\mathcal{S}}(\tilde{X}) &= \mathbf{P}_x(\text{ “no coord. out of } \mathcal{S} \text{ is chosen” }) f_{\mathcal{S}}(x) \\
&\quad + \mathbf{P}_x(\text{ “a coord. out of } \mathcal{S} \text{ is chosen” }) \Big(\frac{1}{2} f_{\mathcal{S}}(x) + \frac{1}{2}(-1) f_{\mathcal{S}}(x)\Big) \\
&= \Big(1 - \frac{|\mathcal{S}|}{d}\Big) f_{\mathcal{S}}(x). \qquad (7.22)
\end{aligned}$$

This provides us with all eigenvalues and eigenfunctions of P. In particular, $\mathcal{S} = \{1,\dots,d\}$ has 0 as corresponding eigenvalue, $\mathcal{S} = \emptyset$ has $\lambda_1 = 1$ and the $\mathcal{S}$ with $|\mathcal{S}| = 1$ have $\lambda_2 = 1 - \frac{1}{d}$, i.e. the spectral gap is $\frac{1}{d}$.

Hence theorem 7.7 gives $\tau_1(\epsilon) \leq \frac{\log\left(\frac{1}{\epsilon\cdot\pi_{min}}\right)}{\log\left(\frac{1}{1-\frac{1}{d}}\right)}$. As $\pi(\cdot) = \frac{1}{2^d}$ we have $\pi_{min} = \frac{1}{2^d}$ and therefore $\tau_1(\epsilon) \leq \frac{\log\left(\frac{2^d}{\epsilon}\right)}{\log\left(\frac{1}{1-\frac{1}{d}}\right)} = \frac{d\log 2 - \log\epsilon}{-\log(1-\frac{1}{d})} = \frac{d\log 2-\log\epsilon}{\frac{1}{d}(1+O(\frac{1}{d}))} = O(d^2)$, which in this case is an overestimation comparing it with previous results.

7.3 Bounds on the Spectrum via Contractions

Definition 7.11 *The Lipschitz constant of a function f on a metric space (Ω, d) is defined as*

$$\mathrm{Lip}(f) = \sup_{x\neq y} \frac{|f(x)-f(y)|}{d(x,y)}.$$

Suppose there exists a constant $\theta < 1$ and a coupling such that

$$\mathbb{E} d(X_1, Y_1) \leq \theta d(x,y), \text{ for all } x, y, \qquad (7.23)$$

where X_1 resp. Y_1 are the states of the chain started at x resp. y and run one step.

Theorem 7.12 *If the chain is irreducible aperiodic, and (7.23) holds, then the eigenvalues of the chain satisfy $|\lambda_j| \leq \theta$ for $j \geq 2$ (reversibility is not necessary for proving this theorem).*

Proof: (M. F. Chen) Let f satisfy $\mathrm{P}f = \lambda f$. First estimate:

$$|\mathrm{P}f(x) - \mathrm{P}f(y)| = |\mathbb{E}(f(X_1) - f(Y_1))| \leq \mathbb{E}|f(X_1) - f(Y_1)|.$$

The definition of the Lipschitz constant for $|f(X_1) - f(Y_1)|$ and the hypothesis (7.23) yield

$$|\mathrm{P}f(x) - \mathrm{P}f(y)| \leq \mathrm{Lip}(f)\mathbb{E}d(X_1,Y_1) \leq \theta\,\mathrm{Lip}(f) d(x,y).$$

We now obtain

$$\operatorname{Lip}(\mathrm{P}f) \leq \theta \operatorname{Lip}(f)$$

involving no restrictions on behalf of the considered Markov chain so far. If f is constant we have $\lambda = 1 = \lambda_1$, see exercise 7.6. Let f_j be one of the non-constant eigenfunctions for $j \geq 2$, then we obtain

$$|\lambda_j| \operatorname{Lip}(f_j) = \operatorname{Lip}(\lambda_j f_j) = \operatorname{Lip}(\mathrm{P}f_j) \leq \theta \operatorname{Lip}(f_j)$$

for $j \geq 2$ which proves the claim. ■

Remark 7.13 *For the hypercube with contraction $1 - \frac{1}{d}$ the inequality of λ_2 is sharp: Let us mix as described in example 4.9, then we obtain:*

$$\mathbb{E}d(X_1, Y_1) \leq \Big(1 - \frac{d(x,y)}{d}\Big)d(x,y) + \frac{d(x,y)}{d}(d(x,y) - 1) = \Big(1 - \frac{1}{d}\Big)d(x,y),$$

i.e. we have found $\theta = 1 - \frac{1}{d}$. Recalling our result in example 7.10, we obtain $\theta = \lambda_2$.

7.4 A new Interpretation of the Kantorovich Distance

Definition 7.14 *Let $\tilde{d}_K(\mu, \nu) = \sup\limits_{\operatorname{Lip}(f) \leq 1} |\int f d\mu - \int f d\nu|$.*

It is easy to see that $\tilde{d}_K \leq d_K$, where d_K denotes the Kantorovich distance as usual: If $\operatorname{Lip}(f) \leq 1$ and (X, Y) is a coupling of μ and ν realizing the Kantorovich distance, then

$$\left| \int f d\mu - \int f d\nu \right| = |\mathbb{E}(f(X) - f(Y))| \leq \mathbb{E}d(X, Y) = d_K(\mu, \nu),$$

where we used $\operatorname{Lip}(f) \leq 1$ for the inequality and optimal coupling for the last equality.

The following theorem provides us with the other direction:

Theorem 7.15 *Kantorovich-Rubinstein (1958): $d_K = \tilde{d}_K$.*

Remark 7.16 *This theorem is more generally valid on compact metric spaces. The proof uses a form of duality.*

Lecture 8: Expander Graphs

8.1 Definition of Expander Graphs

A sequence of graphs $\{G_n\}$ is called an *expander* family if there exists $\Theta > 0$ such that for all n and for all $A \subset V(G_n)$ such that $\sum_{x \in A} \deg(x) \le |E(G_n)|$ we have

$$|\partial A| > \Theta \sum_{x \in A} \deg(x),$$

where ∂A denotes the set of edges connecting A to its complement A^c.

Consider the simple random walk on a graph. The stationary measure is $\pi(x) = \frac{deg(x)}{2|E|}$, and thus in the Cheeger constant terminology, a sequence of graphs $\{G_n\}$ is an expander family if there exists $\Theta > 0$ such that the Cheeger constant of the simple random walk satisfies

$$\Phi_*(G_n) \ge \Theta \quad \forall n.$$

8.2 A Random Construction of Expander Graphs

We now construct a family of 3-regular expander graphs. This is the first construction of an expander family and it is due to Pinsker (1973). Let $G = (V, E)$ be a bipartite graph with equal sides, A and B, each with n vertices. Denote $A, B = \{1, \ldots, n\}$. Draw uniformly at random two permutations $\pi_1, \pi_2 \in S_n$, and set the edge set to be $E = \{(i,i), (i, \pi_1(i)), (i, \pi_2(i)) : 1 \le i \le n\}$.

Theorem 8.1 *With positive probability bounded below from* 0*,* G *has a positive Cheeger constant, i.e., there exists* $\delta > 0$ *such that for any* $S \subset V$ *with* $|S| \le n$ *we have*

$$\frac{\#\{\textit{edges between } S \textit{ and } S^c\}}{\#\{\textit{edges in } S\}} > \delta\,.$$

Proof: It is enough to prove that any $S \subset A$ of size $k \le \frac{n}{2}$ has at least $(1+\delta)k$ neighbors in B. This is because for any $S \subset V$ simply consider the side in which S has more vertices, and if this side has more than $\frac{n}{2}$ vertices, just look at an arbitrary subset of size exactly $\frac{n}{2}$ vertices. Let $S \subset A$ be a set of size $k \le \frac{n}{2}$, and denote by $N(S)$ the neighborhood of S. We wish to bound the probability that $|N(S)| \le (1+\delta)k$. Since (i,i) is an edge for any $1 \le i \le k$, we get immediately that $|N(S)| \ge k$. So all we have to enumerate is the surplus δk vertices that a set which contains $N(S)$ will have, and to make sure both $\pi_1(S)$ and $\pi_2(S)$ fall within that set. This argument gives

$$\mathbf{P}\Big[|N(S)| \le (1+\delta)k\Big] \le \frac{\binom{n}{\delta k}\binom{(1+\delta)k}{k}^2}{\binom{n}{k}^2}\,,$$

so

$$\mathbf{P}\Big[\exists S\,, |S| \le \frac{n}{2}\,, |N(S)| \le (1+\delta)k\Big] \le \sum_{k=1}^{\frac{n}{2}} \binom{n}{k} \frac{\binom{n}{\delta k}\binom{(1+\delta)k}{\delta k}^2}{\binom{n}{k}^2}\,.$$

To conclude the proof we need to show that there exists $\delta > 0$ such that the above sum will be strictly less than 1 uniformly in n. We bound $\binom{n}{\delta k} \leq \frac{n^{\delta k}}{(\delta k)!}$, similarly $\binom{(1+\delta)k}{\delta k}$ and $\binom{n}{k} \geq \frac{n^k}{k^k}$. This gives

$$\sum_{k=1}^{\frac{n}{2}} \frac{\binom{n}{\delta k}\binom{(1+\delta)k}{\delta k}^2}{\binom{n}{k}} \leq \sum_{k=1}^{\frac{n}{2}} \frac{n^{\delta k}((1+\delta)k)^{2\delta k} k^k}{(\delta k)!^3 n^k}.$$

Recall that for any integer ℓ we have $\ell! > (\ell/e)^\ell$, and bound $(\delta k)!$ by this. We get

$$\sum_{k=1}^{\frac{n}{2}} \frac{\binom{n}{\delta k}\binom{(1+\delta)k}{\delta k}^2}{\binom{n}{k}} \leq \sum_{k=1}^{\log n} \left(\frac{\log n}{n}\right)^{(1-\delta)k} \left[\frac{e^3(1+\delta)^2}{\delta^3}\right]^{\delta k} + \sum_{k=\log n}^{\frac{n}{2}} \left(\frac{k}{n}\right)^{(1-\delta)k} \left[\frac{e^3(1+\delta)^2}{\delta^3}\right]^{\delta k}.$$

The first sum clearly tends to 0 as n tends to ∞, for any $\delta \in (0,1)$, and since $\frac{k}{n} \leq \frac{1}{2}$ and $\frac{1}{2}^{(1-\delta)}\left[\frac{e^3(1+\delta)^2}{\delta^3}\right]^\delta < 0.9$ for $\delta < 0.01$, for any such δ the second sum tends to 0 as n tends to ∞. ∎

8.3 Mixing Time for Random Walk on Expander Graphs

Let $|\lambda_*| = \max_{i\geq 2}|\lambda_i|$ where $\{\lambda_i\}$ are the eigenvalues of a reversible Markov chain. Also let $g_* = 1 - \lambda_*$ and $\pi_{\min} = \min_{x\in V}\pi(x)$. We previously proved using spectral decomposition that if P is a transition matrix of a reversible Markov chain, then for any two states x, y we have

$$\left|\frac{P^t(x,y)}{\pi(y)} - 1\right| \leq \frac{e^{-g_* t}}{\pi_{\min}},$$

hence the mixing time $\tau_1(\epsilon)$ satisfies

$$\tau_1(\epsilon) \leq \frac{1}{g_*}\log\frac{1}{\epsilon\pi_{\min}}.$$

The following theorem connects the spectral gap and the Cheeger constant.

Theorem 8.2 *Assume the walk is lazy, i.e. $p(x,x) \geq \frac{1}{2}$ for all x. Then*

$$\frac{\Phi_*^2}{2} \leq g_* \leq 2\Phi_*.$$

By the theorem and the above discussion we find that for a uniformly bounded degree expander family the mixing time of the simple random walk satisfies

$$\tau_1(1/4) \leq O(\log(V_n)),$$

where we denote $V_n = |V(G_n)|$. This is indeed optimal. Let Δ denote the maximum degree of the expander family. Clearly, $P^t(x,\cdot)$ is supported on at most Δ^t vertices, so if $\Delta^t < V_n/2\Delta$ then

$$||P^t(x,\cdot) - \pi||_{\mathrm{TV}} \geq \frac{1}{2},$$

and so $\tau_1(1/4) \geq \Omega(\log(V_n))$.

Lecture 9: The Comparison Method

9.1 The Dirichlet Form

Definition 9.1 *The* Dirichlet form *for P is given by*

$$\mathcal{E}(f,h) = \mathrm{Re}\langle (I-P)f, h\rangle_{\ell^2(\pi)}.$$

$\mathcal{E}(f) = \mathcal{E}(f,f)$ has several equivalent formulations:

$$\begin{aligned}\mathcal{E}(f) &= \frac{1}{2}\mathbb{E}_\pi\left[f(X_0) - f(X_1)\right]^2 \\ &= \mathbb{E}_\pi\left[f(X_0)^2 - f(X_0)f(X_1)\right]\end{aligned}$$

by stationarity of π;

$$\begin{aligned}\mathcal{E}(f) &= \frac{1}{2}\sum_{x,y}\pi(x)p(x,y)[f(x)-f(y)]^2 \\ &= \frac{1}{2}\sum_{x,y}Q(x,y)[f(x)-f(y)]^2;\end{aligned}$$

and, if P is reversible, then it is self-adjoint and therefore

$$\mathcal{E}(f) = \langle (I-P)f, f\rangle_{\ell^2(\pi)}.$$

Definition 9.2 *The* spectral gap *of P is*

$$\begin{aligned}g &= \min_f\left(\frac{\mathcal{E}(f)}{\mathrm{Var}(f)}1[Var(f)\neq 0]\right) \\ &= \min_f\left(\frac{\mathcal{E}(f)}{\mathrm{Var}(f)}1[f\perp \mathbf{1}, f\neq \mathbf{0}]\right)\end{aligned}$$

Recall that if P is irreducible and aperiodic, then, the eigenvalues of P can be written as, $1=\lambda_1 > \lambda_2 > \cdots > \lambda_N > -1$. If P is reversible, $f\perp\mathbf{1}$ where $\mathbf{1}$ is the vector $(1,\ldots,1)$, and $\|f\|_2$=1, then $f=\sum_{j\geq 2}a_jf_j$, where f_j are eigenfunctions of P and $\sum_{j=2}^n a_j^2 = 1$. Thus,

$$\langle (I-P)f, f\rangle = \sum_{j\geq 2}a_j^2(1-\lambda_j) \geq 1-\lambda_2,$$

implying that $g = 1-\lambda_2$.[1]

[1] For historic reasons, $g_* := 1-|\lambda_2|$ is also called spectral gap. We denote it by g_* to distinguish it from g which we defined here.

9.2 A Lower Bound

Let P be reversible and have eigenfunction f: $Pf = \lambda f$, with $\lambda \neq 1$. Since the chain is reversible, eigenfunctions for distinct eigenvalues are orthogonal, and so $< \mathbf{1}, f >= \sum_y \pi(y)f(y) = 0$. Then

$$\left|\lambda^t f(x)\right| = \left|\left(P^t f\right)(x)\right| = \left|\sum_y \left[p^t(x,y)f(y) - \pi(y)f(y)\right]\right| \leq \|f\|_\infty 2d(t),$$

where $d(t) = \max_x \|p^t(x,\cdot) - \pi\|_{TV}$. With this inequality, we can obtain a lower bound on the mixing time. Taking x with $|f(x)| = \|f\|_\infty$ yields $|\lambda|^{\tau_1(\epsilon)} \leq 2\epsilon$, and so

$$\tau_1(\epsilon)\left(\frac{1}{|\lambda|} - 1\right) \geq \tau_1(\epsilon)\log\left(\frac{1}{|\lambda|}\right) \geq \log\left(\frac{1}{2\epsilon}\right).$$

Recall that $|\lambda_*| := \max_{j\geq 2}|\lambda_j| < 1$ since P is irreducible and aperiodic, and that we defined $g_* = 1 - |\lambda_*|$. Rewriting the above, we have

$$\tau_1(\epsilon) \geq \left(\frac{1}{g_*} - 1\right)\log\left(\frac{1}{2\epsilon}\right).$$

Key Fact If g_* is small because the smallest eigenvalue λ_N is near -1, the slow mixing sugessted by this lower bound can be rectified by passing to a lazy or continuous time chain to make eigenvalues positive. However, if the largest eigenvalue λ_2 is near 1, then the mixing may be very slow indeed. Therefore, we are mainly concerned with g, not g_*

9.3 The Comparison Theorem

Recall that for lazy simple random walk on the d-dimensional torus $\mathbb{Z}_n^d$, we used coupling to show that $\tau_1 \leq C_d n^2$ and $\frac{1}{g} \leq K_d n^2$ for constants C_d and K_d. Now, suppose we remove some edges from the graph (e.g. some subset of the horizontal edges at even heights). Then coupling cannot be applied, due to the irregular pattern. The following theorem—proved in various forms by Jerrum and Sinclair (1989), Diaconis and Stroock (1991), Quastel (1992), Saloff-Coste (1993), and in the form presented here by Diaconis and Saloff-Coste, allows one to compare the behavior of similar chains to achieve bounds on the mixing time in general.

Theorem 9.3 (The Comparison Theorem) *Let π, P and $\tilde{\pi}, \tilde{P}$ be two Markov chains on Ω. Write $E = \{(z,w) : P(z,w) > 0\}$, and similarly with $\tilde{E}$. Assume that for all $(x,y) \in \tilde{E}$, there is a path $(e_1, e_2, \ldots, e_m)$ contained in E from x to y. For every pair (x,y), we choose one out of these paths and denote it by γ_{xy}. Define the* congestion ratio *to be*

$$B = \max_{e=(z,w)}\left(\frac{1}{Q(z,w)}\sum_{\gamma_{xy}\ni e}\tilde{Q}(x,y)\,|\gamma_{xy}|\right).$$

Then $\tilde{\mathcal{E}}(f) \leq B\mathcal{E}(f)$ for all f.

Remark 9.4 *In the reversible case, it follows from the variational formula for the spectral gap that $\tilde{g} \leq Bg$. An important special sub-case occurs when $\tilde{\pi} = \pi$ and $\tilde{p}(x,y) = \pi(y)$. Then $\tilde{\mathcal{E}}(f) =$*

$\mathrm{Var}_\pi(f) = \frac{1}{2}\sum_{x,y}\pi(x)\pi(y)[f(x)-f(y)]^2$, *and so* $\mathcal{E}(f) \geq \frac{1}{B}\mathrm{Var}(f)$, *giving* $g \geq \frac{1}{B}$. *Thus gives the following bound:*

$$\frac{1}{g} \leq B = \max_{e=(z,w)} \sum_{\gamma_{xy}\ni e} \pi(x)\pi(y)|\gamma_{xy}| \tag{9.1}$$

Proof:

$$\begin{aligned}
2\tilde{\mathcal{E}}(f) &= \sum_{(x,y)\in\tilde{E}} \tilde{Q}(x,y)[f(x)-f(y)]^2 \\
&= \sum_{x,y} \tilde{Q}(x,y)\left[\sum_{e\in\gamma_{x,y}} df(e)\right]^2
\end{aligned}$$

where for an edge $e=(z,w)$, we write $df(e) = f(w) - f(z)$. Thus,

$$\begin{aligned}
2\tilde{\mathcal{E}}(f) &\leq \sum_{x.y} \tilde{Q}(x,y)\,|\gamma_{xy}| \sum_{e\in\gamma_{x,y}} (df(e))^2 \\
&= \sum_{e\in E}\left[\sum_{\gamma_{xy}\ni e} \tilde{Q}(x,y)\,|\gamma_{xy}|\right](df(e))^2 \\
&\leq \sum_{(z,w)\in E} BQ(z,w)[f(w)-f(z)]^2 \\
&= 2B\mathcal{E}(f).
\end{aligned}$$

■

Example 9.5 (Comparison for Simple Random Walks on Graphs) *If two graphs have the same vertex set but different edge sets* E *and* $\tilde{E}$, *then*

$$Q(x,y) = \frac{1}{2|E|}, \quad \textit{and} \quad \tilde{Q}(x,y) = \frac{1}{2|\tilde{E}|},$$

since the vertex degrees cancel. Therefore, the congestion ratio is simply

$$B = \left(\max_{e\in E}\sum_{\gamma_{xy}\ni e}|\gamma_{xy}|\right)\frac{|E|}{|\tilde{E}|}.$$

In our motivating example, we only removed horizontal edges at even heights from the torus. Since all odd-height edges remain, we can take $|\gamma_{xy}| \leq 3$ *since we can traverse any missing edge in the torus by moving upwards, then across the edge of odd height, and then downwards. The horizontal edge in this path would then be used by at most 3 paths* γ *(including the edge itself). Since we removed at most one quarter of the edges, we get that* $B \leq 12$, *and therefore, we can compare the spectral gap of the irregular graph* g *to that of the complete torus* $\tilde{g}$, *yielding that the relaxation time for the irregular graph satisfies* $\frac{1}{g} \leq 12\frac{1}{\tilde{g}} \leq C_d n^2$.

Exercise 9.6 *Show that for lazy simple random walk on the box* $[1,\ldots,n]^d$, *where opposite faces are not identified, the relaxation time still satisfies* $\frac{1}{g} \leq C_d n^2$.

9.4 An ℓ^2 Bound

If P is reversible, then

$$\left\| \frac{p^t(x,y)}{\pi(y)} - 1 \right\|^2_{\ell^2(\pi\times\pi)} \leq \left\| \sum_{j=2}^{N} \lambda_j^t f_j(x) f_j(y) \right\|^2_{\ell^2(\pi\times\pi)} \leq \sum_{j=2}^{N} \lambda_j^{2t}$$

by orthogonality. If the chain is lazy simple random walk on a transitive graph,

$$2\left\| p^t(x,\cdot) - \pi \right\|_{TV} = \left\| \frac{p^t(x,\cdot)}{\pi} - 1 \right\|_{\ell^1(\pi)} \leq \left\| \frac{p^t(x,\cdot)}{\pi} - 1 \right\|_{\ell^2(\pi)} \leq \sqrt{\sum_{j=2}^{N} \lambda_j^{2t}}.$$

For lazy simple random walk on the hypercube $\{0,1\}^n$,

$$\left\| p^t(x,\cdot) - \pi \right\|^2 \leq \sum_{k=1}^{n} \left(1 - \frac{k}{m}\right)^{2t} \binom{n}{k} \leq \sum_{k=1}^{n} e^{-2tk/n} \binom{n}{k} = \left(1 + e^{-2t/n}\right)^n - 1.$$

If we take $t = \frac{1}{2} n \log n + An$, then

$$\left\| p^t(x,\cdot) - \pi \right\|^2 \leq \left(1 + \frac{1}{n} e^{-2A}\right)^n - 1 \leq e^{e^{-2A}} - 1,$$

suggesting a cutoff phenomenon at $\frac{1}{2} n \log n$.

Lecture 10: Spectral gap for the Ising model

10.1 Open questions

We start with a few open questions.

Open Question: For the Ising model on any graph with n vertices, show that $\tau_1(1/4) \geq Cn\log n$ where C is a universal constant.

Hayes and Sinclair have shown that this result holds with $C = C(\text{max degree})$, but their constant tends to zero as the degree of the graph increases.

Open Question: For the Ising model on a general n vertex graph, show that $\tau_1^{\beta}(1/4)$ increases in β.

Open Question: For the Ising model on a general n vertex graph, show that $\frac{1}{g(\beta)}$ increases in β and in the edge set.

10.2 Bounding $1/g$ for Ising model

Consider the Ising model on $G = (V, E)$ with n vertices with $\beta = 1/T$. Recall that the probability of observing a configuration $\sigma \in \{-1, 1\}^V$ is given by $\pi(\sigma) = \frac{1}{Z} e^{\beta H(\sigma)}$, where $H(\sigma) = \sum_{u \sim v} \sigma_u \sigma_v + \sum_u h_u \sigma_u$. In order to state our result, we must first introduce some terminology.

Definition 10.1 *The* cutwidth $W(G)$ *of a graph G is obtained as follows. Order the vertices of G arbitrarily: $v_1, v_2, \ldots, v_n$ and let S_n denote the set of permutations of n elements. Then*

$$W(G) = \min_{\tau \in S_n} \max_k \# \left\{ \textit{edges from } \left\{ v_{\tau(i)} \right\}_{i=1}^{k} \textit{ to } \left\{ v_{\tau(j)} \right\}_{j>k} \right\}$$

Example 10.2 The cycle has cutwidth 2, an $\ell \times m$ lattice with $\ell \leq m$ has cutwidth $\ell + 1$.

Example 10.3 The d-dimensional lattice $[0, n]^d$ has cutwidth $W([0, n]^d) \sim c_d n^{d-1}$.

Example 10.4 The binary tree with height ℓ, (so the number of vertices $n \sim 2^{\ell}$), has cutwidth $W(\text{binary tree}) \sim \log n$.

Out main result, using the cutwidth of the graph to bound the spectral gap, is presented as the following theorem.

Theorem 10.5 *For the Ising model on a graph G with n vertices,*

$$\frac{1}{g(\beta)} \leq n^2 e^{4\beta W(G)}.$$

Proof: The proof uses the comparison theorem introduced in lecture 9, it is based on a combinatorial method introduced by Jerrum-Sinclair for studying monomer-dimer systems. Order the vertices of G using a total order $<$. Suppose σ and η are two spin configurations differing at the vertices $x^1, x^2, \ldots, x^\ell$. Let $\gamma(\sigma,\eta) = (\sigma^0, \sigma^1, \ldots, \sigma^\ell)$ be the path in configution space defined by

$$\sigma^j(x) = \begin{cases} \eta(x) & \text{if } x \le x^j \\ \sigma(x) & \text{if } x > x^j \end{cases}.$$

Now, consider an edge $e = (\xi, \xi^x)$, where $\xi(y) = \xi^x(y)$ for all $y \neq x$, in the path $\gamma(\sigma,\eta)$. Suppose without loss of generality that $\pi(\xi^x) \ge \pi(\xi)$ (otherwise flip σ and η). Define $Q(e) = \pi(\xi)\mathrm{P}(\xi,\xi^x)$ and note that $Q(e) \ge \frac{\pi(\xi)}{2n}$.

Set $\Gamma(e) = \{\gamma(\sigma',\eta') : e \in \gamma(\sigma',\eta')\}$. Define $\Phi_e : \Gamma(e) \to \{-1,1\}^V$ as follows,

$$\Phi_e(\gamma(\sigma',\eta'))(y) = \begin{cases} \sigma'(y) & \text{if } y < x \\ \eta'(y) & \text{if } y \ge x \end{cases}.$$

Since for every γ, $\Phi_e(\gamma)(x) = \eta'(x) = \xi^x(x)$. So the image set of Φ_e sits in $S := \{\zeta \in \{-1,1\}^V : \zeta(x) = \xi^x(x)\}$. On the other hand, for any $\Phi_e(\gamma(\sigma',\eta')) \in S$, we could reconstruct σ' as

$$\sigma'(y) = \begin{cases} \Phi_e(\gamma(\sigma',\eta'))(y) & \text{if } y < x \\ \xi(y) & \text{if } y \ge x \end{cases}$$

and similarly η' may be reconstructed from e and $\Phi_e(\gamma(\sigma',\eta'))$ by the same way, so Φ_e is a bijection between $\Gamma(e)$ and S. Now observe that under the optimal ordering of the vertices of V, the following inequality holds

$$H(\sigma') + H(\eta') - H(\xi) - H(\Phi_e(\gamma(\sigma',\eta'))) \le 4W(G) \tag{10.1}$$

This is true because any edge that does not go across from $\{y : y > x\}$ to $\{y : y \le x\}$ will not contribute to the left hand side. Exponentiating (10.1) we obtain

$$\begin{aligned} \pi(\sigma')\pi(\eta') &\le e^{4\beta W(G)}\pi(\xi)\pi(\Phi_e(\gamma(\sigma',\eta'))) \\ &\le 2ne^{4\beta W(G)}Q(e)\pi(\Phi_e(\gamma(\sigma',\eta'))). \end{aligned}$$

Now, summing over all the paths $\gamma(\sigma',\eta')$ which contain the edge e, we obtain

$$\begin{aligned} \sum_{\sigma',\eta' : e \in \gamma(\sigma',\eta')} \pi(\sigma')\pi(\eta')|\gamma(\sigma',\eta')| &\le 2n^2 e^{4\beta W(G)} Q(e) \sum_{\gamma \in \Gamma(e)} \pi(\Phi_e(\gamma)) \\ &= n^2 e^{4\beta W(G)} Q(e). \end{aligned}$$

The last equality is because $\sum_{\gamma \in \Gamma(e)} \pi(\Phi_e(\gamma)) = \sum_{\zeta \in \{-1,1\}^V : \zeta(x) = \xi^x(x)} \pi(\zeta) = \frac{1}{2}$. It follows from the comparison theorem (9.1) that $\frac{1}{g} \le n^2 e^{4\beta W(G)}$, as claimed. ■

Remark 10.6 This result is "sharp" for boxes in $\mathbb{Z}^2$ provided that β is large. More precisely, one can show that for an $m \times m$ box in $\mathbb{Z}^2$ we have $\frac{1}{g} \ge e^{c\beta m}$. This lower bound can be obtained by showing that the configurations σ for which $\sum_j \sigma_j = 0$ (assuming m even) form a bottleneck which inhibits the mixing.

Lecture 11: Mixing Times for Card Shuffling and Lattice Paths

This lecture, delivered by David B. Wilson of Microsoft Research, presents results from two papers by the speaker, *Mixing Times of Lozenge Tilings and Card Shuffling Markov Chains* (2001) and *Mixing Time of the Rudvalis Shuffle* (2002). Both are available on the arXiv. Several useful illustrations are found in (Wilson, 2001); see pp. 1, 5, 10, 12, and 14.

11.1 Introduction; random adjacent transpositions

In computing mixing times, we wish to find an event which is likely to occur after t steps of the Markov chain, but unlikely to occur at stationarity. In this lecture, the event will be that a certain eigenfunction is sufficiently large.

As a specific instance of this method, consider a deck of n cards laid out left to right. At each time t, shuffle according to the following random adjacent transposition rule. Pick a pair of adjacent cards from among the $n-1$ possible pairs, uniformly at random. With probability $1/2$, swap this pair of cards, and otherwise do nothing.

This chain had been previously studied and was known to converge in $O(n^3 \log n)$ time. The best known lower bound was about n^3 time; but it was strongly suspected that $n^3 \log n$ was the correct order of magnitude. In this lecture, upper and lower bounds that are within a constant factor of each other will be proved.

Our plan is to find a eigenfunction Φ on the state space such that $E(\Phi(\sigma_{t+1}) \,|\, \sigma_t) = \lambda \Phi(\sigma_t)$, where $\lambda < 1$. If we can, it follows that $E(\Phi(\sigma)) = 0$ at stationarity; so as long as $\Phi(\sigma_t)$ remains large enough, the size of $\Phi(\sigma)$ can be used to show that mixing has not yet occured.

First, some notation. σ_t is the state of the chain at time t. A typical card is denoted $\diamondsuit$ (the diamond suit symbol); $\sigma_t(\diamondsuit)$ is the position of $\diamondsuit$ at time t. For technical reasons, it is convenient to label the possible positions of a card as $x = \frac{1}{2}, \frac{3}{2}, \ldots, n - \frac{1}{2}$ (rather than as $0, 1, \ldots, n-1$, for instance). That is, $\sigma_t(\diamondsuit) \in \{\frac{1}{2}, \frac{3}{2}, \ldots, n - \frac{1}{2}\}$.

Fix a card $\diamondsuit$ and follow it. According to the dynamics of the whole chain, $\diamondsuit$ performs a lazy random walk on $\{\frac{1}{2}, \frac{3}{2}, \ldots, n - \frac{1}{2}\}$, with probability $\dfrac{1}{2(n-1)}$ of moving to the left and the same probability of moving to the right (except at the boundaries, where the probability of an illegal move is replaced by 0). Temporarily call this random walk's transition matrix $\bar{P}$. It is easy to verify that the left eigenvalues of $\bar{P}$ are $\{u_k\}_{k=0}^{n-1}$ where

$$u_k(x) = \cos \frac{\pi k x}{n}$$

with corresponding eigenvalues

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

(Use the addition and subtraction formulas for cosine.) These eigenvectors give rise to eigenfunctions

$\Phi_{k,\diamondsuit}(\sigma) = u_k(\sigma(\diamondsuit))$ which satisfy

$$\begin{aligned} E(\Phi_{k,\diamondsuit}(\sigma_{t+1}) \,|\, \sigma_t) &= \sum_x u_k(x)\bar{P}(\sigma_t(\diamondsuit), x) \\ &= (u_k P)(\sigma_t(\diamondsuit)) \\ &= \lambda_k u_k(\sigma_t(\diamondsuit)) = \lambda_k \Phi_{k,\diamondsuit}(\sigma_t) \end{aligned}$$

as desired. Now, we are interested in the second-largest eigenvalue, which here corresponds to $k = 1$. So define

$$\Phi_\diamondsuit(\sigma) = \cos \frac{\pi\sigma(\diamondsuit)}{n}.$$

In principle, we might think of studying all the $\Phi_\diamondsuit$'s as $\diamondsuit$ ranges over the cards of the deck. With an eye to future developments, however, a better idea is to fix an arbitrary subset Red $\subset$ deck and define

$$\Phi(\sigma) = \sum_{\diamondsuit \in \mathrm{Red}} \Phi_\diamondsuit(\sigma)$$

where the dependence on the choice of red cards is implicit. Both Φ and $\Phi_\diamondsuit$ are eigenfunctions, with

$$E(\Phi(\sigma_{t+1}) \,|\, \sigma_t) = \lambda \Phi(\sigma_t) \quad \text{and}$$

$$\lambda = \lambda_1 = 1 - \frac{1}{n-1}\left(1 - \cos\frac{\pi}{n}\right) \approx 1 - \frac{\pi^2}{2n^3}.$$

With Φ now defined, we are ready to state the theorem which converts Φ into a lower bound on the mixing time.

11.2 The lower bounds

Theorem 11.1 *Suppose Φ is a function on the state space of a Markov chain with the property that $E(\Phi(X_{t+1}) \,|\, X_t) = (1-\gamma)\Phi(X_t)$, with $E((\Delta\Phi)^2 \,|\, X_t) = E((\Phi(X_{t+1}) - \Phi(X_t))^2 \,|\, X_t) \le R$ and $0 < \gamma \le 2 - \sqrt{2}$. Then the mixing time is at least*

$$\frac{\log \Phi_{max} + \frac{1}{2}\log\frac{\gamma\epsilon}{4R}}{-\log(1-\gamma)}$$

in the sense that for values of t up to that value, the variation distance from stationarity is at least $1 - \epsilon$.

This theorem actually follows (except for a slightly different bound on γ) from a more general theorem which is relevant to the Rudvalis shuffle:

Theorem 11.2 *Suppose a Markov chain X_t has a lifting (X_t, Y_t), and that Ψ is a function on the lifted state space such that $E(\Psi(X_t, Y_t) \,|\, (X_t, Y_t)) = \lambda\Psi(X_t, Y_t)$, and such that $|\,\Psi(x, y)\,|$ is a function of x alone; and also $|\,\lambda\,| < 1$, $\gamma = 1 - \Re(\lambda) \le 1/2$, and $E(|\,\Psi(X_{t+1}, Y_{t+1}) - \Psi(X_t, Y_t)\,|^2 \,|\, (X_t, Y_t)) \le R$. Then the mixing time is at least*

$$\frac{\log \Psi_{max} + \frac{1}{2}\log\frac{\gamma\epsilon}{4R}}{-\log(1-\gamma)}$$

in the same sense as before.

Proof: Write $\Psi_t = \Psi(X_t, Y_t)$, $\Delta\Psi = \Psi_{t+1} - \Psi_t$, and Ψ_t^* for the complex conjugate. By induction, it follows that $E(\Psi_t) = \lambda^t \Psi_0$; since $|\,\lambda\,| < 1$, $E(\Psi) = 0$ at stationarity. Also $E(\Delta\Psi\,|\,(X_t, Y_t)) = (\lambda - 1)\Psi_t$ and, writing

$$\begin{aligned}
\Psi_{t+1}\Psi_{t+1}^* &= \Psi_t\Psi_t^* + \Psi_t\Delta\Psi^* + \Psi_t^*\Delta\Psi + |\,\Delta\Psi\,|^2\,, \\
E(\Psi_{t+1}\Psi_{t+1}^*\,|\,(X_t, Y_t)) &= \Psi_t\Psi_t^*[1 + (\lambda-1)^* + (\lambda-1)] + E(|\,\Delta\Psi\,|^2\ |\, X_t) \\
&\le \Psi_t\Psi_t^*[2\Re(\lambda) - 1] + R.
\end{aligned}$$

Induction gives

$$E(\Psi_t\Psi_t^*) \le \Psi_0\Psi_0^*[2\Re(\lambda) - 1]^t + \frac{R}{2 - 2\Re(\lambda)} \qquad \text{and}$$

$$\operatorname{Var}(\Psi_t) = E(\Psi_t\Psi_t^*) - E(\Psi_t)E(\Psi_t)^* \le \Psi_0\Psi_0^*\left[[2\Re(\lambda) - 1]^t - (\lambda\lambda^*)^t\right] + \frac{R}{2 - 2\Re(\lambda)}.$$

Our assumption on $\Re(\lambda)$ means the first term is negative (because: $(1-\lambda)(1-\lambda^*) \ge 0$ leads to $\lambda\lambda^* \ge 2\Re(\lambda) - 1 \ge 0$ and hence $(\lambda\lambda^*)^t \ge [2\Re(\lambda) - 1]^t$ for all t) so that

$$\operatorname{Var}(\Psi_t) \le \frac{R}{2 - 2\Re(\lambda)} = \frac{R}{2\gamma}$$

and so $\operatorname{Var}(\Psi) \le \frac{R}{2\gamma}$ at stationarity also.

We are now in a position to use Chebyshev's inequality. To acheive a bound of $1 - \epsilon$ on the variation distance, the appropriate estimates are: at stationarity,

$$P\left(|\,\Psi\,| \ge \sqrt{\frac{R}{\gamma\epsilon}}\right) \le \frac{\epsilon}{2}, \tag{11.1}$$

whereas

$$P\left(|\,\Psi_t - E\Psi_t\,| \ge \sqrt{\frac{R}{\gamma\epsilon}}\right) \le \frac{\epsilon}{2}$$

and so by the triangle inequality

$$P\left(|\,\Psi_t\,| \le \sqrt{\frac{R}{\gamma\epsilon}}\right) \le \frac{\epsilon}{2} \tag{11.2}$$

provided that

$$|\,E(\Psi_t)\,| \ge 2\sqrt{\frac{R}{\gamma\epsilon}}. \tag{11.3}$$

If we can arrange (11.3), then (11.2) together with (11.1) will show that the variation distance from stationarity is at least $1 - \epsilon$. But we know $|\,E(\Psi_t)\,| = |\,\lambda\,|^t\,|\,\Psi_0\,| \ge |\,\Psi_0\,|\,(\Re(\lambda))^t = |\,\Psi_0\,|\,(1-\gamma)^t$. The initial conditions are arbitrary so choose them so that $|\,\Psi_0\,| = \Psi_{\max}$ is maximized. Finally, substituting in (11.3) and solving for t shows that mixing has not occured as long as

$$t \le \frac{-\log\left(2\sqrt{\frac{R}{\gamma\epsilon}}/\Psi_{\max}\right)}{-\log(1-\gamma)} = \frac{\log\Psi_{\max} + \frac{1}{2}\log\frac{\gamma\epsilon}{4R}}{-\log(1-\gamma)}$$

as claimed. ■

Let us apply this theorem to the random adjacent transposition example. Recall that

$$\Phi(\sigma) = \sum_{\Diamond\in\text{Red}} \cos\frac{\pi\sigma(\Diamond)}{n}$$

where we must select Red $\subset$ deck. In principle, different choices of Red (to make $\Phi_{\max}$ bigger) might have the side effect of making R bigger as well, which might worsen the estimate. In actual fact, though, whatever the choice of Red, we will have $|\,\Delta\Phi\,| \le \frac{\pi}{n}$ and hence $R = \frac{\pi^2}{n^2}$ (since the only transposition that alters Φ is the swap of a red card and a black card, and this only changes one term in the sum, by at most $\frac{\pi}{n}$). So our best estimate will come from simply making $\Phi_{\max}$ as large as possible, in other words choosing as red cards every card which contributes a positive term $\cos\frac{\pi x}{n}$ to Φ.

Once this is done, we find that $\Phi_{\max} \sim \mathit{const}\cdot n$. Recall also that $\gamma = 1-\lambda = \frac{1}{n-1}\left(1-\cos\frac{\pi}{n}\right) \approx \frac{\pi^2}{2n^3}$; so overall, the mixing time is

$$\sim \frac{\log n - \mathit{const} + \frac{1}{2}\log(\epsilon/8n)}{\pi^2/2n^3} \sim \frac{1}{\pi^2}n^3\log n.$$

It is believed that the constant $\frac{1}{\pi^2}$ is in fact the correct one.

11.3 An upper bound; lattice paths

Eigenfunction techniques can also be used to find an upper bound. To do so, it is helpful to reformulate the permutation problem in terms of lattice paths.

Our lattice paths will be on the diagonally oriented lattice, i.e., $\mathbb{Z}^2$ rotated by forty-five degrees, moving from left to right. Thus, the two possible moves at each step are north-east and south-east. Given a permutation, we can encode it as a collection of lattice paths as follows. Fix a subset Red. At each position x, if the card at position x is a red card, move up (north-east); otherwise move down (south-east). As an even more abbreviated form, record the moves as a sequence of bits (threshold functions), with 1 for up and 0 for down. See David Wilson's illustration in (Wilson, 2001, p. 10).

For a fixed choice of red cards, the permutation obviously cannot be recovered from the lattice path moves. However, number the cards from 1 to n, and successively choose the red cards to be: the highest card only; the two highest cards; $\ldots$; all but card 1; all the cards. Then the permutation can be recovered from the collection of all n lattice paths. (Indeed, in the bit interpretation, the value of the card at position x is simply the number of bits at position x which are 1.) Hence "mixing to within ϵ" can be acheived by requiring that every lattice path has "mixed to within ϵ/n".

In lattice path terminology, the Markov chain dynamics are the following: Pick three adjacent points of the path uniformly at random. If the middle point is a local maximum or minimum ($\vee$ or $\wedge$), switch it with probability 1/2; if it is not a local extremum, do nothing. In the bit version, choose a pair of adjacent bits and randomly either sort them or reverse-sort them. In this context, we consider the state space to be all lattice paths have a fixed number a of up moves and $b = n - a$ down moves, so that the chain is irreducible. (So for the permutation case, a will range from 1 to n.)

Here, the eigenfunction is (after adjusting the coordinates in an appropriate way)

$$\Phi(h) = \sum_{x=-n/2}^{n/2} h(x)\cos\frac{\pi x}{n}$$

with the same eigenvalue as before,

$$\lambda = 1 - \frac{1}{n-1}\left(1-\cos\frac{\pi}{n}\right).$$

This time, to get an upper bound on the mixing time, the key observation is that Φ is discrete, so

$$P(\Phi \neq 0) \leq \frac{1}{\Phi_{\min}} E(\Phi_t)$$

and the right-hand side can be computed as before. We apply this to the height difference between two coupled paths to show that they must coalesce with probability $1 - \epsilon$ within

$$\frac{2 + o(1)}{\pi^2} n^3 \log \frac{ab}{\epsilon}$$

steps.

Returning to card permutations, we can acheive coupling of the permutations with probability $1 - \epsilon$ as soon as every lattice path has coupled with probability $1 - \epsilon/n$. As a crude bound, $ab \leq n^2$ uniformly, so it takes at most about

$$\frac{2}{\pi^2} n^3 \log \frac{n^2}{\epsilon/n} = \frac{6}{\pi^2} n^3 \log n$$

steps to ensure mixing. (The constant 6 is not optimal and (Wilson, 2001) improves it to 2.)

11.4 Extensions

In the theorem proved here, from Wilson (2002), the Markov chain could be lifted to a larger chain (X_t, Y_t), an innovation we did not need for random adjacent transpositions. It is relevant for other shuffles, notably the Rudvalis shuffle and variants, which will be very briefly described. See (Wilson, 2002) for further details.

In the original Rudvalis shuffle, at each stage the top card is randomly either swapped with the bottom card or not, and then the entire deck is rotated by putting the (possibly new) top card on the bottom. In other words, either "swap and shift-left" or "swap". Two variants change the possible moves to (1) "swap" or "shift-left" or (2) "swap", "shift-left", "shift-right", or "hold" (do nothing).

In all cases, by time t there has been a net shift of Y_t cards to the left, where $Y_t = t$ for the Rudvalis shuffle and Y_t is random for the variants. In all cases, the presence of these shifts would tend to make the direct eigenfunction approach fail: the eigenvalues are complex, and the eigenfunction values tend towards a shrinking, narrow annulus – rather than concentrating at a point, with the result that the variance becomes too large. However, retaining the value of Y_t as part of the state space (i.e., lifting the chain) allows us to introduce a phase factor correction in the eigenfunction which corrects for this effect. Eigenfunctions for the lifted shuffles take the form

$$\Psi_\diamondsuit(X_t, Y_t) = v(X_t(\diamondsuit)) e^{2\pi i Z_t(\diamondsuit)/n}$$

where $|\,\Psi_\diamondsuit(X_t)\,| = v(X_t(\diamondsuit))$ depends only on the original chain, and $Z_t(\diamondsuit) = X_t(\diamondsuit) - X_0(\diamondsuit) + Y_t \pmod n$.

Another alteration is to relax the requirement of finding an actual eigenfunction. (For example, $|\,\Psi\,|$ above is no longer necessarily an eigenfunction, but it is the function that provides the mixing time.) For shuffling on the $\sqrt{n} \times \sqrt{n}$ grid, the boundary conditions make finding an exact eigenvector computationally difficult. But an approximate eigenvector can be good enough.

Lattice paths also have application in lozenge tilings. Like permutations, lozenge tilings can be understood in terms of a family of lattice paths; now, however, it is necessary to consider the interactions between distinct lattice paths. (Wilson, 2001) also discusses this and other Markov chains.

Lecture 12: Shuffing by semi-random transpositions and Ising model on Trees

12.1 Shuffling by random and semi-random transpositions.

Shuffling by random transpositions is one of the simplest random walks on the symmetric group: given n cards in a row, at each step two cards are picked uniformly at random and exchanged. Diaconis and Shahshahani, in the 1980's, showed that the cutoff for mixing in this random walk is $\frac{1}{2}n\log n$. Broder, using a strong uniform time argument, has given the mixing time

$$\tau_1 = O(n\log n).$$

Open Question: Is there a coupling proof for this mixing time?

There are a few facts about this random walk. By the time $\frac{\theta}{2}n\log n$, with $\theta < 1$, the number of cards untouched is with high probability bigger than $\frac{n^{1-\theta}}{2}$. While at time t, the expectation of the number of untouched cards is $n(1-\frac{1}{n})^{2t}$. This shuffle was precisely analyzed in 1981, see [12].

Let $\{L_t\}_{t=1}^{\infty}$ be a sequence of random variables taking values in $[n] = \{0, 1, \ldots, n-1\}$ and let $\{R_t\}_{t=1}^{\infty}$ be a sequence of i.i.d. cards chosen uniformly from $[n]$. The **semi-random transposition shuffle** generated by $\{L_t\}$ is a stochastic process $\{\sigma_t^*\}_{t=0}^{\infty}$ on the symmetric group S_n, defined as follows. Fix the initial permutation σ_0^*. The permutation σ_t^* at time t is obtained from σ_{t-1}^* by transposing the cards at locations L_t and R_t.

One example is the **cyclic-to-random shuffle**, in which the sequence L_t is given by $L_t = t \mod n$.

Mironov showed that Broder's method can be adapted to yield a uniform upper bound of $O(n\log n)$ on the mixing time of all semi-random transpositions. Mossel, Peres and Sinclair proved a lower bound of $\Omega(n\log n)$ for the mixing time of the cyclic-to-random shuffle. They also proved a general upper bound of $O(n\log n)$ on the mixing time of *any* semi-random transposition shuffle.

Open Question: Is there a universal constant $c > 0$ such that, for any semi-random transposition shuffle on n cards, the mixing time is at least $cn\log n$?

For this lower bound question there is no obvious reduction to the case where the sequence $\{L_t\}$ is deterministic, so conceivably the question could have different answers for deterministic $\{L_t\}$ and random $\{L_t\}$. One hard case is the following.

For each $k \geq 0$, let $\{L_{kn+r}\}_{r=1}^{n}$ be a uniform random permutation of $\{0, \ldots, n-1\}$, where these permutations are independent.

Another two cases may give some illumination on the lower bound.

Case 1: At time t, let $L_t = t \mod n$, R_t is choosing uniformly from $\{t, t+1, \ldots, n\}$. Note that this situation is not a semi-random transposition shuffle as we defined. This random walk gives stationary distribution π at the n-th step.

Case 2: At time t, let $L_t = t \mod n$, R_t is choosing uniformly from $\{1, 2, \ldots, n\}$. This is a semi-random transposition shuffle.

Exercise 12.1 *Prove that this random walk does not give the stationary distribution after n steps.*

12.2 Ising model on trees

Consider the Ising model on the b-ary tree $T_r = T_r^{(b)}$ with r levels. The number of vertices $n = \sum_{j=0}^{r} b^j$. The cut-width is

$$W = rb.$$

Consider the Ising model at temperature β. Using graph geometry to bound $\frac{1}{g}$ as we learned in the last lecture, we get:

$$\frac{1}{g} \leq n^2 e^{4\beta W} \leq n^{O(\beta)}$$

where g is the spectral gap. This means $\frac{1}{g}$ is bounded by a polynomial of n.

This Ising model have some other versions of descriptions in applications. For example, the mutation model in biology and the noisy broadcast model.

In the mutation model, we imagine that the child's configuration is with $1-\epsilon$ probability to be the same as its parent, and with ϵ probability to be different. We set spins to a b-ary tree T_r from its root. σ_{root} is uniformly chosen from $\{+,-\}$. Then, scan the tree top-down, assigning vertex v a spin equal to the spin of its parent with probability $1-\epsilon$ and opposite with probability ϵ. This construction gives us a distribution on $\{-1,1\}^{T_r}$. We prove that this is an Ising distribution.

Suppose we flip a subtree of σ, and get $\widetilde{\sigma}$ as shown in figure 12.4. If $0 < \epsilon < 1/2$, then

$$\frac{\mathbf{P}(\widetilde{\sigma})}{\mathbf{P}(\sigma)} = \frac{\epsilon}{1-\epsilon}$$

This coordinates with the ratio of $\mathbf{P}(\widetilde{\sigma})$ and $\mathbf{P}(\sigma)$ in the Ising model, which says

$$\frac{\mathbf{P}(\widetilde{\sigma})}{\mathbf{P}(\sigma)} = e^{-2\beta}$$

If we set β as

$$e^{-2\beta} = \frac{\epsilon}{1-\epsilon}, \tag{12.1}$$

since for any two different configurations σ and $\widetilde{\sigma}$, we can always do several flips to turn σ into $\widetilde{\sigma}$, the ratio of $\mathbf{P}(\sigma)$ and $\mathbf{P}(\widetilde{\sigma})$ given in the mutation model is always the same as given in the Ising distribution. This proves that the mutation model gives us the Ising distribution. Sometimes we

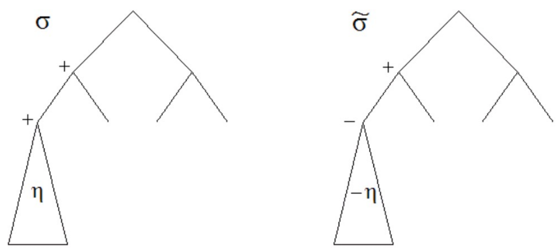

Figure 12.1: Flipping a subtree

use another parameter $\theta = 1-2\epsilon$ instead of ϵ. By equation (12.1), we get

$$\theta = \frac{e^{2\beta}-1}{e^{2\beta}+1} = \tanh(\beta)$$

One thing we care about in the mutation model is how to predict the configuration of the ancestors given the information of the children. Even more, can we predict the structure of the tree? There are some results addressing these questions.

When $1-2\epsilon < 1/b$, the expected value of the spin at the root σ_ρ given any boundary conditions $\sigma_{\partial T_r^{(b)}}$ decays exponentially in r. In the intermediate regime, where $1/b < 1-2\epsilon < 1/\sqrt{b}$, this exponential decay still holds for typical boundary conditions, but not for certain exceptional boundary conditions, such as the all $+$ boundary. In the low temperature regime, where $1-2\epsilon > 1/\sqrt{b}$, typical boundary conditions impose bias on the expected value of the spin at the root σ_ρ. At this low temperature regime, the reconstruction of the configuration becomes possible.

We now introduce another representation for the Ising model, called the percolation representation. The mutation model can be represented as

$$\sigma_v = \begin{cases} \sigma_w, & \text{with probability } 1-2\epsilon = \theta; \\ \text{uniformly from } \{1,-1\}, & \text{with probability } 2\epsilon. \end{cases} \tag{12.2}$$

If at some edges, the uniform distributed situation has been chosen, then the children thereafter are distributed uniformly at random. Thus those children can not do any help to predict the ancestor. So it is the same as the following model.

In the tree T_r, we "cross out" each edge independently with probability $1-\theta$. After this, if the root is still connected to the leaves, then the spins at the leaves help to predict the spin at the root. So the reconstruction fails if $\mathbf{P}$(root connected to the k-th level) $\to 0$, as $k \to \infty$. A necessary condition for this is $b\theta \leq 1$, which means $1-2\epsilon \leq \frac{1}{b}$. Intuitionally, the condition for possible reconstruction should be $(b\theta)^k > \sqrt{b^k}$, which gives us a hint for the results of the thresholds of root reconstructions.

12.3 Lower Bounds at low temperatures

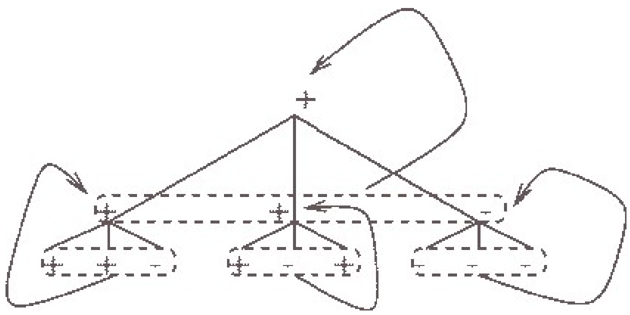

Figure 12.2: The recursive majority function

In order to find the lower bound of $1/g$ for low temperatures, we apply recursive majority to the boundary spins. For simplicity we consider first the ternary tree T, see Figure 12.2. Recursive majority is defined on the configuration space as follows. Given a configuration σ, denote the recursive majority value at v as m_v. For leaves v, $m_v = \sigma_v$. If $\{m_k\}$ is defined for all of the children u of w. Then, define m_w by the majority of $\{m_u\}$, where u take values in all of w's children.

Lemma 12.2 *If u and w are children of the same parent v, then* $\mathbf{P}[m_u \neq m_w] \leq 2\epsilon + 8\epsilon^2$.

Proof: We have

$$\mathbf{P}[m_u \neq m_w] \leq \mathbf{P}[\sigma_u \neq m_u] + \mathbf{P}[\sigma_w \neq m_w] + \mathbf{P}[\sigma_u \neq \sigma_v] + \mathbf{P}[\sigma_w \neq \sigma_v].$$

We will show that recursive majority is highly correlated with spin; in particular, if ϵ is small enough, then $\mathbf{P}[m_v \neq \sigma_v] \leq 4\epsilon^2$.

The proof is by induction on the distance ℓ from v to the boundary of the tree. For a vertex v at distance ℓ from the boundary of the tree, write $p_\ell = \mathbf{P}[m_v \neq \sigma_v]$. By definition $p_0 = 0 \leq 4\epsilon^2$.

For the induction step, note that if $\sigma_v \neq m_v$ then one of the following events hold:

- At least 2 of the children of v, have different σ value than that of σ_v, or
- One of the children of v has a spin different from the spin at v, and for some other child w we have $m_w \neq \sigma_w$, or
- For at least 2 of the children of v, we have $\sigma_w \neq m_w$.

Summing up the probabilities of these events, we see that $p_\ell \leq 3\epsilon^2 + 6\epsilon p_{\ell-1} + 3p_{\ell-1}^2$. It follows that $p_\ell \leq 4\epsilon^2$, hence the Lemma. ■

Let $m = m_{root}$. Then by symmetry $\mathbb{E}[m] = 0$, and $\mathbb{E}[m^2] = 1$. Recall that

$$g = \min_{f \neq const} \frac{\mathcal{E}(f)}{Var(f)}$$

If we plug in $f = m$, then we get

$$\begin{aligned} g &\leq \frac{\mathcal{E}(m)}{Var(m)} \\ &= \mathcal{E}(m) \\ &= \frac{1}{2}\mathbf{P}(m(\sigma) \neq m(\widetilde{\sigma})) \cdot 4 \end{aligned} \tag{12.3}$$

Where σ has the Ising distribution and $\widetilde{\sigma}$ is one step glauber move from σ.

From Lemma 12.2, we know that if u and v are siblings, then $\mathbf{P}[m_u \neq m_w] \leq 2\epsilon + 8\epsilon^2$. Since $m(\sigma) \neq m(\widetilde{\sigma})$ only when we update a leaf, and all the ancestors of this leaf has two siblings with different values under function m. So, for a ternary tree with height k, we have

$$\begin{aligned} \mathbf{P}(m(\sigma) \neq m(\widetilde{\sigma})) &\leq (2\epsilon + 8\epsilon^2)^{k-1} \\ &\leq (3e^{-2\beta})^{k-1} \\ &\leq n^{-c\beta} \end{aligned} \tag{12.4}$$

Sum up equation (12.3) and (12.4), we get the polynomial lower bound for $\frac{1}{g}$ in the low temperature case.

Note that the proof above easily extends to the d-regular tree for $d > 3$. A similar proof also applies to the binary tree T, where the majority function m is defined as follows. Look at vertices at distance r for even r. For the boundary vertices define $m_v = \sigma_v$. For each vertex v at distance 2 from the boundary, choose three leaves on the boundary below it v_1, v_2, v_3 (e.g. the first three) and let m_v be the majority of the values m_{v_i}. Now continue recursively.

Repeating the above proof, and letting $p_\ell = P[m_v \neq \sigma_v]$ for a vertex at distance 2ℓ from the boundary, we derive the following recursion: $p_\ell \leq 3\epsilon + 6(2\epsilon)p_{\ell-1} + 3p_{\ell-1}^2$. By induction, we get $p_\ell \leq 4\epsilon$ (suppose ϵ is small enought). If u and v are with same even distance from the boundary, and have the same ancestor at distance two above them. Then, $\mathbf{P}[m_u \neq m_v] \leq 4\epsilon + 2(4\epsilon) = 12\epsilon$. It follows:

$$\begin{aligned}
\mathbf{P}(m(\sigma) \neq m(\widetilde{\sigma})) &\leq (12\epsilon)^{\lfloor \frac{k}{2} \rfloor} \\
&\leq (12e^{-2\beta})^{\lfloor \frac{k}{2} \rfloor} \\
&\leq n^{-c\beta}
\end{aligned}$$

where k is the height of the binary tree. This, as the same as the ternary tree, gives the polynomial lower bound for $\frac{1}{g}$.

Lecture 13: Evolving sets

13.1 Introduction

It is well known that the absence of "bottlenecks" in the state space of a Markov chain implies rapid mixing. Precise formulations of this principle, related to Cheeger's inequality in differential geometry, have been proved by algebraic and combinatorial techniques [3, 19, 17, 23, 14, 20]. They have been used to approximate permanents, to sample from the lattice points in a convex set, to estimate volumes, and to analyze a random walk on a percolation cluster in a box.

In this lecture, we show that a new probabilistic technique, introduced in [25], yields the sharpest bounds obtained to date on mixing times in terms of bottlenecks.

Let $\{\mathrm{P}(x,y)\}$ be transition probabilities for an irreducible Markov chain on a countable state space V, with stationary distribution π. For $x, y \in V$, let $Q(x,y) = \pi(x)\mathrm{P}(x,y)$, and for $S, A \subset V$, define $Q(S,A) = \sum_{s\in S, a\in A} Q(s,a)$. For $S \subset V$, the "boundary size" of S is measured by $|\partial S| = Q(S, S^c)$. Following [17], we call $\Phi_S := \frac{|\partial S|}{\pi(S)}$ the *conductance* of S. Write $\pi_* := \min_{x\in V} \pi(x)$ and define $\Phi(r)$ for $r \in [\pi_*, 1/2]$ by

$$\Phi(r) = \inf\left\{\Phi_S : \pi(S) \le r\right\} . \tag{13.1}$$

For $r > 1/2$, let $\Phi(r) = \Phi_* = \Phi(1/2)$. Define the ϵ-*uniform mixing time* by

$$\tau_u(\epsilon) = \tau_{unif}(\epsilon) := \min\left\{n : \left|\frac{\mathrm{P}^n(x,y) - \pi(y)}{\pi(y)}\right| \le \epsilon \ \ \forall x, y \in V\right\}.$$

Jerrum and Sinclair [17] considered chains that are reversible ($Q(x,y) = Q(y,x)$ for all $x, y \in V$) and also satisfy

$$\mathrm{P}(x,x) \ge 1/2 \text{ for all } x \in V . \tag{13.2}$$

They estimated the second eigenvalue of P in terms of conductance, and derived the bound

$$\tau_{unif}(\epsilon) \le 2\Phi_*^{-2}\left(\log\frac{1}{\pi_*} + \log\frac{1}{\epsilon}\right) . \tag{13.3}$$

We will prove (13.3) in the next lecture. Algorithmic applications of (13.3) are described in [30]. Extensions of (13.3) to non-reversible chains were obtained by Mihail [23] and Fill [14]. A striking new idea was introduced by Lovász and Kannan [20], who realized that in geometric examples, small sets often have larger conductance, and discovered a way to exploit this. Let $\|\mu - \nu\| = \frac{1}{2}\sum_{y\in V} |\mu(y) - \nu(y)|$ be the total variation distance, and denote by

$$\tau_1(\epsilon) := \min\left\{n : \|p^n(x,\cdot) - \pi\| \le \epsilon \text{ for all } x \in V\right\} \tag{13.4}$$

the ϵ-mixing time in total variation. (This can be considerably smaller than the uniform mixing time $\tau_u(\epsilon)$, see the lamplighter walk discussed at the end of this section, or §13.6, Remark 1.) For reversible chains that satisfy (13.2), Lovász and Kannan proved that

$$\tau_1(1/4) \le 2000 \int_{\pi_*}^{3/4} \frac{du}{u\Phi^2(u)}, \tag{13.5}$$

Note that in general, $\tau_1(\epsilon) \le \tau_1(1/4)\log_2(1/\epsilon)$. Therefore, ignoring constant factors, the bound in (13.5) is tighter than the bound of (13.3), but at the cost of employing a weaker notion of mixing.

The main result sharpens (13.5) to a bound on the uniform mixing time. See Theorem 13.9 for a version that relaxes the assumption (13.2). We use the notation $\alpha \wedge \beta := \min\{\alpha, \beta\}$.

Theorem 13.1 *Assume (13.2). Then the ϵ-uniform mixing time satisfies*

$$\tau_u(\epsilon) \le 1 + \int_{4\pi_*}^{4/\epsilon} \frac{4du}{u\Phi^2(u)}. \tag{13.6}$$

More precisely, if

$$n \ge 1 + \int_{4(\pi(x)\wedge\pi(y))}^{4/\epsilon} \frac{4du}{u\Phi^2(u)}, \tag{13.7}$$

then

$$\left|\frac{p^n(x,y) - \pi(y)}{\pi(y)}\right| \le \epsilon. \tag{13.8}$$

(Recall that $\Phi(r)$ is constant for $r \ge \frac{1}{2}$.) This result has several advantages over (13.5):

- The uniformity in (13.6).
- It yields a better bound when the approximation parameter ϵ is small.
- It applies to non-reversible chains.
- It yields an improvement of the upper bound on the time to achieve (13.8) when $\pi(x), \pi(y)$ are larger than π_*.
- The improved constant factors make the bound (13.6) potentially applicable as a stopping time in simulations.

Other ways to measure bottlenecks can yield sharper bounds. One approach, based on "blocking conductance functions" and restricted to the mixing time in total variation τ_1, is presented in [18, Theorem 3].

Another boundary gauge ψ is defined in §13.2. For the n-dimensional unit hypercube, this gauge (applied to the right class of sets, see [26]) gives a bound of the right order $\tau_u(1/e) = O(n \log n)$ for the uniform mixing time. Previous methods of measuring bottlenecks did not yield the right order of magnitude for the uniform mixing time in this benchmark example.

Theorem 13.1 is related to another line of research, namely the derivation of heat kernel estimates for Markov chains using Nash and Sobolev inequalities. For finite Markov chains, such estimates were obtained by Chung and Yau [8], and by Diaconis and Saloff-Coste [13]. In particular, for the special case where Φ is a power law, the conclusion of Theorem 13.1 can be obtained by combining Theorems 2.3.1 and 3.3.11 of Saloff-Coste [29]. For infinite Markov chains, Nash inequalities have been developed for general isoperimetric profiles; see Varopoulos [31], the survey by Pittet and Saloff Coste [27], the book [32], and especially the work of Coulhon [10, 11]. Even in this highly developed subject, our probabilistic technique yields improved estimates when the stationary measure is not uniform. Suppose that π is an infinite stationary measure on V for the transition kernel p. As before, we define

$$Q(x,y) = \pi(x)p(x,y); \qquad |\partial S| = Q(S, S^c); \qquad \Phi_S := \frac{|\partial S|}{\pi(S)}.$$

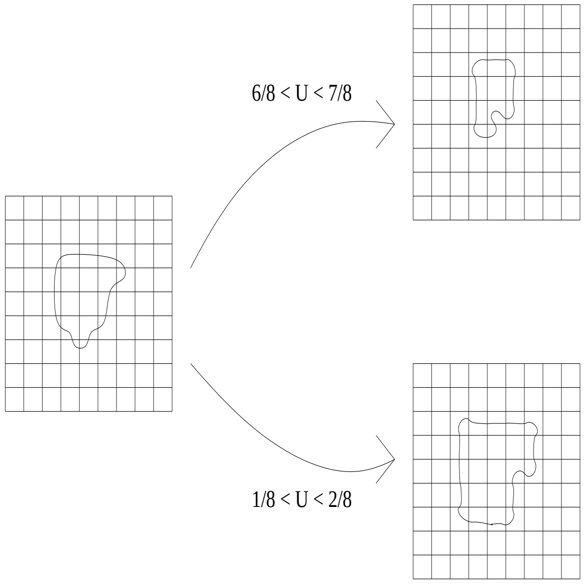

Figure 13.1: One step of the evolving set process.

Define $\Phi(r)$ for $r \in [\pi_*, \infty)$ by

$$\Phi(r) = \inf \{\Phi_S : \pi(S) \le r\} . \tag{13.9}$$

For the rest of the introduction, we focus on the case of finite stationary measure.

Definition 13.2 Evolving sets. *Given V, π and Q as above, consider the Markov chain $\{S_n\}$ on subsets of V with the following transition rule. If the current state S_n is $S \subset V$, choose U uniformly from $[0,1]$ and let the next state S_{n+1} be*

$$\widetilde{S} = \{y : Q(S,y) \ge U\pi(y)\} .$$

Consequently,

$$\mathbf{P}(y \in \widetilde{S}) = \mathbf{P}\Big(Q(S,y) \ge U\pi(y)\Big) = \frac{Q(S,y)}{\pi(y)} . \tag{13.10}$$

Figure 13.2 illustrates one step of the evolving set process when the original Markov chain is a random walk in a box (with a holding probability of $\frac{1}{2}$). Since π is the stationary distribution, $\emptyset$ and V are absorbing states for the evolving set process.

Write $\mathbf{P}_S\Big(\,\cdot\,\Big) := \mathbf{P}\Big(\,\cdot\,\Big|\, S_0 = S\Big)$ and similarly for $\mathbf{E}_S\Big(\,\cdot\,\Big)$. The utility of evolving sets stems from the relation

$$\mathrm{P}^n(x,y) = \frac{\pi(y)}{\pi(x)}\, \mathbf{P}_{\{x\}}(y \in S_n)$$

(see Proposition 13.11). Their connection to mixing is indicated by the inequality

$$\|\mu_n - \pi\| \le \frac{1}{\pi(x)} \mathbf{E}_{\{x\}} \sqrt{\pi(S_n) \wedge \pi(S_n^c)} ,$$

where $\mu_n := p^n(x, \cdot\,)$; see [26] for a sharper form of this. The connection of evolving sets to conductance can be seen in Lemma 13.7 below.

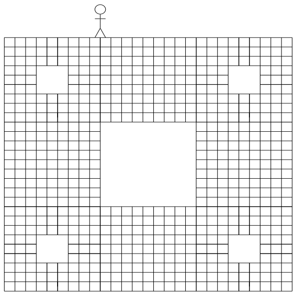

Figure 13.2: A box with holes.

Example 13.3 (Random Walk in a Box): Consider a simple random walk in an $n \times n$ box. To guarantee condition (13.2) we add a holding probability of $\frac{1}{2}$ to each state (i.e., with probability $\frac{1}{2}$ do nothing, else move as above). When $1/2 \le u \le 1$, the conductance profile satisfies

$$\Phi(u) \ge \frac{a}{n\sqrt{u}},$$

where a is a constant. Thus our bound implies that the ϵ uniform mixing time is at most

$$C_\epsilon + 4\int_{1/n^2}^{1/2} \frac{1}{u\left(\frac{a}{n\sqrt{u}}\right)^2}\, du = O(n^2),$$

which is the correct order of magnitude. Of course, other techniques such as coupling or spectral methods would give the correct-order bound of $O(n^2)$ in this case. However, these techniques are not robust under small perturbations of the problem, whereas the conductance method is.

Example 13.4 (Box with Holes): For a random walk in a box with holes (see Figure 13.4), it is considerably harder to apply coupling or spectral methods. However, it is clear that the conductance profile for the random walk is unchanged (up to a constant factor), and hence the mixing time is still $O(n^2)$.

Example 13.5 (Random Walk in a Percolation Cluster): In fact, the conductance method is robust enough to handle an even more extreme variant: Suppose that each edge in the box is deleted with probability $1-p$, where $p > \frac{1}{2}$. Then with high probability there is a connected component that contains a constant fraction of the original edges. Benjamini and Mossel [5] showed that for the random walk in the big component the conductance profile is sufficiently close (with high probability) to that of the box and deduced that the mixing time is still $O(n^2)$. (See [22] for analogous results in higher dimensions.) By our result, this also applies to the uniform mixing times.

Example 13.6 (Random Walk on a Lamplighter Group): The following natural chain mixes more rapidly in the sense of total variation than in the uniform sense. A state of this chain consists of n lamps arrayed in a circle, each lamp either *on* (1) or *off* (0), and a *lamplighter* located next to one of the lamps. In one "active" step of the chain, the lamplighter either switches the current lamp or moves at random to one of the two adjacent lamps. We consider the lazy chain that stays put with probability 1/2 and makes an active step with probability 1/2. The path of the lamplighter is a delayed simple random walk on a cycle, and this implies that $\tau_1(1/4) = \Theta(n^2)$, see [15]. However, by considering the possibility that the lamplighter stays in one half of the cycle

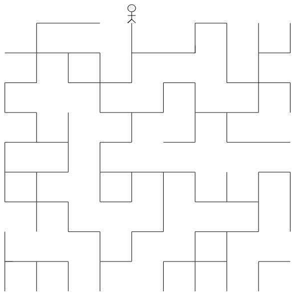

Figure 13.3: Random walk in a percolation cluster.

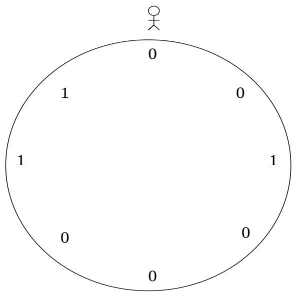

Figure 13.4: Random walk on a lamplighter group

for a long time, one easily verifies that $\tau_u(1/4) \geq c_1 n^3$ for some constant $c_1 > 0$. Using the general estimate $\tau_u(\epsilon) = O(\tau_1(\epsilon)\log(1/\pi_*))$ gives a matching upper bound $\tau_u(1/4) = O(n^3)$.

13.2 Further results and proof of Theorem 13.1

We will actually prove a stronger form of Theorem 13.1, using the boundary gauge

$$\psi(S) := 1 - \mathbf{E}_S\sqrt{\frac{\pi(\widetilde{S})}{\pi(S)}}$$

instead of the conductance Φ_S. The next lemma relates these quantities.

Lemma 13.7 *Let $\emptyset \neq S \subset V$. If (13.2) holds, then $\psi(S) \geq \Phi_S^2/2$. More generally, if $0 < \gamma \leq \frac{1}{2}$ and $p(x,x) \geq \gamma$ for all $x \in V$, then $\psi(S) \geq \frac{\gamma^2}{2(1-\gamma)^2}\Phi_S^2$.*

See §13.4 for the proof. In fact, $\psi(S)$ is often much larger than Φ_S^2.

Define the *root profile* $\psi(r)$ for $r \in [\pi_*, 1/2]$ by

$$\psi(r) = \inf\{\psi(S) : \pi(S) \leq r\}, \tag{13.11}$$

and for $r > 1/2$, let $\psi(r) := \psi_* = \psi(\frac{1}{2})$. Observe that the root profile ψ is (weakly) decreasing on $[\pi_*, \infty)$.

For a measure μ on V, write

$$\chi^2(\mu,\pi) := \sum_{y\in V} \pi(y)\Big(\frac{\mu(y)}{\pi(y)}-1\Big)^2 = \Big(\sum_{y\in V}\frac{\mu(y)^2}{\pi(y)}\Big) - 1\,. \tag{13.12}$$

By Cauchy-Schwarz,

$$2\|\mu-\pi\| = \Big\|\frac{\mu(\,\cdot\,)}{\pi(\,\cdot\,)}-1\Big\|_{L^1(\pi)} \le \Big\|\frac{\mu(\,\cdot\,)}{\pi(\,\cdot\,)}-1\Big\|_{L^2(\pi)} = \chi(\mu,\pi)\,. \tag{13.13}$$

We can now state our key result relating evolving sets to mixing.

Theorem 13.8 *Denote $\mu_n = p^n(x,\,\cdot\,)$. Then $\chi^2(\mu_n,\pi)\le\epsilon$ for all*

$$n \ge \int_{4\pi(x)}^{4/\epsilon} \frac{du}{u\psi(u)}\,.$$

See §13.5 for the proof.

Derivation of Theorem 13.1 from Lemma 13.7 and Theorem 13.8:

The *time-reversal* of a Markov chain on V with stationary distribution π and transition matrix $p(x,y)$, is another Markov chain with stationary distribution π, and transition matrix $\hat p(\cdot,\cdot)$ that satisfies $\pi(y)p(y,z) = \pi(z)\hat p(z,y)$ for all $y,z\in V$. Summing over intermediate states gives $\pi(z)\hat p^m(z,y) = \pi(y)p^m(y,z)$ for all $z,y\in V$ and $m\ge 1$.

Since $p^{n+m}(x,z) = \sum_{y\in V} p^n(x,y)p^m(y,z)$, stationarity of π gives

$$p^{n+m}(x,z)-\pi(z) = \sum_{y\in V}\Big(p^n(x,y)-\pi(y)\Big)\Big(p^m(y,z)-\pi(z)\Big) \tag{13.14}$$

whence

$$\Big|\frac{p^{n+m}(x,z)-\pi(z)}{\pi(z)}\Big| \tag{13.15}$$

$$= \Big|\sum_{y\in V}\pi(y)\Big(\frac{p^n(x,y)}{\pi(y)}-1\Big)\Big(\frac{\hat p^m(z,y)}{\pi(y)}-1\Big)\Big| \tag{13.16}$$

$$\le \chi\Big(p^n(x,\cdot),\pi\Big)\chi\Big(\hat p^m(z,\cdot),\pi\Big) \tag{13.17}$$

by Cauchy-Schwarz.

The quantity $Q(S,S^c)$ represents, for any $S\subset V$, the asymptotic frequency of transitions from S to S^c in the stationary Markov chain with transition matrix $p(\cdot,\cdot)$ and hence $Q(S,S^c)=Q(S^c,S)$. It follows that the time-reversed chain has the same conductance profile $\Phi(\cdot)$ as the original Markov chain. Hence, Lemma 13.7 and Theorem 13.8 imply that if

$$m,\ell \ge \int_{4(\pi(x)\wedge\pi(y))}^{4/\epsilon}\frac{2du}{u\Phi^2(u)},$$

and (13.2) holds, then

$$\chi\Big(p^\ell(x,\cdot),\pi\Big) \le \sqrt{\epsilon} \ \text{ and } \ \chi\Big(\hat{p}^m(z,\cdot),\pi\Big) \le \sqrt{\epsilon}\,.$$

Thus by (13.17),

$$\Big|\frac{p^{\ell+m}(x,z)-\pi(z)}{\pi(z)}\Big| \le \epsilon\,,$$

and Theorem 13.1 is established.

In fact, the argument above yields the following more general statement.

Theorem 13.9 *Suppose that* $0<\gamma\le\frac{1}{2}$ *and* $p(x,x)\ge\gamma$ *for all* $x\in V$*. If*

$$n \ge 1 + \frac{(1-\gamma)^2}{\gamma^2}\int_{4(\pi(x)\wedge\pi(y))}^{4/\epsilon} \frac{4du}{u\Phi^2(u)}\,, \tag{13.18}$$

then (13.8) holds.

To complete the proof of Theorems 13.1 and 13.9, it suffices to prove Lemma 13.7 and Theorem 13.8. This is done in §13.4 and §13.5, respectively.

13.3 Properties of Evolving Sets

Lemma 13.10 *The sequence* $\{\pi(S_n)\}_{n\ge 0}$ *forms a martingale.*

Proof: By (13.10), we have

$$\begin{aligned}
\mathbf{E}\left(\pi(S_{n+1})\Big|S_n\right) &= \sum_{y\in V}\pi(y)\,\mathbf{P}\left(y\in S_{n+1}\Big|S_n\right)\\
&= \sum_{y\in V} Q(S_n,y) \;=\; \pi(S_n)\,.
\end{aligned}$$

■

The following proposition relates the nth order transition probabilities of the original chain to the evolving set process.

Proposition 13.11 *For all* $n\ge 0$ *and* $x,y\in V$ *we have*

$$p^n(x,y) = \frac{\pi(y)}{\pi(x)}\,\mathbf{P}_{\{x\}}(y\in S_n)\,.$$

Proof: The proof is by induction on n. The case $n = 0$ is trivial. Fix $n > 0$ and suppose that the result holds for $n-1$. Let U be the uniform random variable used to generate S_n from S_{n-1}. Then

$$\begin{aligned}
p^n(x,y) &= \sum_{z\in V} p^{n-1}(x,z)p(z,y) \\
&= \sum_{z\in V} \mathbf{P}_{\{x\}}(z \in S_{n-1})\,\frac{\pi(z)}{\pi(x)}\,p(z,y) \\
&= \frac{\pi(y)}{\pi(x)}\mathbf{E}_{\{x\}}\left(\frac{1}{\pi(y)}Q(S_{n-1},y)\right) \\
&= \frac{\pi(y)}{\pi(x)}\,\mathbf{P}_{\{x\}}(y \in S_n).
\end{aligned}$$

■

We will also use the following duality property of evolving sets.

Lemma 13.12 *Suppose that $\{S_n\}_{n\geq 0}$ is an evolving set process. Then the sequence of complements $\{S_n^c\}_{n\geq 0}$ is also an evolving set process, with the same transition probabilities.*

Proof: Fix n and let U be the uniform random variable used to generate S_{n+1} from S_n. Note that $Q(S_n,y) + Q(S_n^c,y) = Q(V,y) = \pi(y)$. Therefore, with probability 1,

$$\begin{aligned}
S_{n+1}^c &= \Big\{y : Q(S_n,y) < U\pi(y)\Big\} \\
&= \Big\{y : Q(S_n^c,y) \geq (1-U)\pi(y)\Big\}.
\end{aligned}$$

Thus, $\{S_n^c\}$ has the same transition probabilities as $\{S_n\}$, since $1-U$ is uniform. ■

Next, we write the χ^2 distance between $\mu_n := p^n(x,\,\cdot\,)$ and π in terms of evolving sets. Let $\{S_n\}_{n\geq 0}$ and $\{\Lambda_n\}_{n\geq 0}$ be two independent replicas of the evolving set process, with $S_0 = \Lambda_0 = \{x\}$. Then by (13.12) and Proposition 13.11, $\chi^2(\mu_n,\pi)$ equals

$$\sum_{y\in V}\pi(y)\frac{\mathbf{P}_{\{x\}}(y\in S_n)^2}{\pi(x)^2} - 1 \tag{13.19}$$

$$= \frac{1}{\pi(x)^2}\Big[\sum_{y\in V}\pi(y)\mathbf{P}_{\{x\}}\Big(\{y\in S_n\}\cap\{y\in\Lambda_n\}\Big) - \pi(x)^2\Big] \tag{13.20}$$

$$= \frac{1}{\pi(x)^2}\mathbf{E}_{\{x\}}\Big(\pi(S_n\cap\Lambda_n) - \pi(S_n)\pi(\Lambda_n)\Big)\,, \tag{13.21}$$

where the last equation uses the relation $\pi(x) = \mathbf{E}_{\{x\}}\pi(S_n) = \mathbf{E}_{\{x\}}\pi(\Lambda_n)$. Define

$$S^\sharp := \begin{cases} S & \text{if } \pi(S) \leq \frac{1}{2}; \\ S^c & \text{otherwise,} \end{cases}$$

Lemma 13.13 *For any two sets $S,\Lambda \subset V$,*

$$|\pi(S\cap\Lambda) - \pi(S)\pi(\Lambda)| \leq \sqrt{\pi(S^\sharp)\pi(\Lambda^\sharp)}$$

Proof:

$$\pi(S \cap \Lambda) + \pi(S^c \cap \Lambda) = \pi(\Lambda) = \pi(S)\pi(\Lambda) + \pi(S^c)\pi(\Lambda),$$

and hence

$$|\pi(S \cap \Lambda) - \pi(S)\pi(\Lambda)| = |\pi(S^c \cap \Lambda) - \pi(S^c)\pi(\Lambda)|.$$

Similarly, this expression doesn't change if we replace Λ by Λ^c. Thus,

$$\begin{aligned} |\pi(S \cap \Lambda) - \pi(S)\pi(\Lambda)| &= |\pi(S^\sharp \cap \Lambda^\sharp) - \pi(S^\sharp)\pi(\Lambda^\sharp)| \\ &\leq |\pi(S^\sharp) \wedge \pi(\Lambda^\sharp)| \\ &\leq \sqrt{\pi(S^\sharp)\pi(\Lambda^\sharp)}\,. \end{aligned}$$

■

Apply this lemma into (13.21), we obtain

$$\chi^2(\mu_n, \pi) \leq \frac{1}{\pi(x)^2}\mathbf{E}\sqrt{\pi(S_n^\sharp)\pi(\Lambda_n^\sharp)}\,,$$

whence

$$2\|\mu_n - \pi\| \leq \chi(\mu_n, \pi) \leq \frac{1}{\pi(x)}\mathbf{E}\sqrt{\pi(S_n^\sharp)}\,. \tag{13.22}$$

13.4 Evolving sets and conductance profile: proof of Lemma 3

Lemma 13.14 *For every real number $\beta \in [-\frac{1}{2}, \frac{1}{2}]$, we have*

$$\frac{\sqrt{1+2\beta} + \sqrt{1-2\beta}}{2} \leq \sqrt{1-\beta^2} \leq 1 - \beta^2/2.$$

Proof: Squaring gives the second inequality and converts the first inequality into

$$1 + 2\beta + 1 - 2\beta + 2\sqrt{1-4\beta^2} \leq 4(1-\beta^2)$$

or equivalently, after halving both sides,

$$\sqrt{1-4\beta^2} \leq 1 - 2\beta^2\,,$$

which is verified by squaring again. ■

Lemma 13.15 *Let*

$$\varphi_S := \frac{1}{2\pi(S)} \sum_{y \in V} \Big(Q(S, y) \wedge Q(S^c, y) \Big)\,. \tag{13.23}$$

Then

$$1 - \psi(S) \leq \frac{\sqrt{1+2\varphi_S} + \sqrt{1-2\varphi_S}}{2} \leq 1 - \varphi_S^2/2\,. \tag{13.24}$$

Proof: The second inequality in (13.24) follows immediately from Lemma 13.14. To see the first inequality, let U be the uniform random variable used to generate $\widetilde{S}$ from S. Then

$$\mathbf{P}_S\Big(y \in \widetilde{S} \,\Big|\, U < \tfrac{1}{2}\Big) = 1 \wedge \frac{2Q(S, y)}{\pi(y)}\,.$$

Consequently,

$$\pi(y)\mathbf{P}_S(y \in \widetilde{S} \,|\, U < \tfrac{1}{2}) = Q(S,y) + \Big(Q(S^c,y) \wedge Q(S,y)\Big)\,.$$

Summing over $y \in V$, we infer that

$$\mathbf{E}_S\Big(\pi(\widetilde{S})\,\Big|\,U < \tfrac{1}{2}\Big) = \pi(S) + 2\pi(S)\varphi_S\,. \tag{13.25}$$

Therefore, $R := \pi(\widetilde{S})/\pi(S)$ satisfies $\mathbf{E}_S(R|U < \frac{1}{2}) = 1 + 2\varphi_S$. Since $\mathbf{E}_S R = 1$, it follows that

$$\mathbf{E}_S(R\,|\,U \geq \tfrac{1}{2}) = 1 - 2\varphi_S\,.$$

Thus

$$\begin{aligned} 1 - \psi(S) &= \mathbf{E}(\sqrt{R}) \\ &= \frac{\mathbf{E}(\sqrt{R}|U < \frac{1}{2}) + \mathbf{E}(\sqrt{R}|U \geq \frac{1}{2})}{2} \\ &\leq \frac{\sqrt{\mathbf{E}(R|U < \frac{1}{2})} + \sqrt{\mathbf{E}(R|U \geq \frac{1}{2})}}{2}, \end{aligned}$$

by Jensen's inequality (or by Cauchy-Schwarz). This completes the proof. ■

Proof: (Proof of Lemma 13.7) If $p(y,y) \geq 1/2$ $\forall y \in V$, then it is easy to check directly that $\varphi_S = \Phi_S$ for all $S \subset V$.

If we are only given that $p(y,y) \geq \gamma$ $\forall y \in V$, where $0 < \gamma \leq \frac{1}{2}$, we can still conclude that for $y \in S$,

$$Q(S,y) \wedge Q(S^c,y) \geq \gamma\pi(y) \wedge Q(S^c,y) \geq \frac{\gamma}{1-\gamma} Q(S^c,y)\,.$$

Similarly, for $y \in S^c$ we have $Q(S,y) \wedge Q(S^c,y) \geq \frac{\gamma}{1-\gamma} Q(S,y)$. Therefore

$$\sum_{y \in V} [Q(S,y) \wedge Q(S^c,y)] \geq \frac{2\gamma}{1-\gamma} Q(S,S^c)\,,$$

whence $\varphi_S \geq \frac{\gamma}{1-\gamma}\Phi_S$. This inequality, in conjunction with Lemma 13.15, yields Lemma 13.7. ■

13.5 Proof of Theorem 13.8

Denote by $K(S,A) = \mathbf{P}_S(\widetilde{S} = A)$ the transition kernel for the evolving set process. In this section we will use another Markov chain on sets with transition kernel

$$\widehat{K}(S,A) = \frac{\pi(A)}{\pi(S)} K(S,A). \tag{13.26}$$

This is the Doob transform of $K(\cdot,\cdot)$. As pointed out by J. Fill (Lecture at Amer. Inst. Math. 2004), the process defined by $\widehat{K}$ can be identified with one of the "strong stationary duals" constructed in [9].

The martingale property of the evolving set process, Lemma 13.10, implies that $\sum_A \widehat{K}(S,A) = 1$ for all $S \subset V$. The chain with kernel (13.26) represents the evolving set process conditioned to absorb in V; we will not use this fact explicitly.

Note that induction from equation (13.26) gives

$$\widehat{K}^n(S,A) = \frac{\pi(A)}{\pi(S)} K^n(S,A)$$

for every n, since

$$\begin{aligned}
\widehat{K}^{n+1}(S,B) &= \sum_A \widehat{K}^n(S,A)\widehat{K}(A,B) \\
&= \sum_A \frac{\pi(B)}{\pi(S)} K^n(S,A)K(A,B) \\
&= \frac{\pi(B)}{\pi(S)} K^{n+1}(S,B)
\end{aligned}$$

for every n and $B \subset V$. Therefore, for any function f,

$$\widehat{\mathbf{E}}_S f(S_n) = \mathbf{E}_S \left[\frac{\pi(S_n)}{\pi(S)} f(S_n) \right], \tag{13.27}$$

where we write $\widehat{\mathbf{E}}$ for the expectation when $\{S_n\}$ has transition kernel $\widehat{K}$. Define

$$Z_n = \frac{\sqrt{\pi(S_n^\sharp)}}{\pi(S_n)},$$

and note that $\pi(S_n) = Z_n^{-2}$ when $Z_n \geq \sqrt{2}$, that is, when $\pi(S_n) \leq \frac{1}{2}$. Then by equations (13.27) and (13.22), $\chi(\mu_n, \pi) \leq \widehat{\mathbf{E}}_{\{x\}}(Z_n)$ and

$$\begin{aligned}
\widehat{\mathbf{E}}\left(\frac{Z_{n+1}}{Z_n} \Big| S_n \right) &= \mathbf{E}\left(\frac{\pi(S_{n+1})}{\pi(S_n)} \cdot \frac{Z_{n+1}}{Z_n} \Big| S_n \right) \\
&= \mathbf{E}\left(\frac{\sqrt{\pi(S_{n+1}^\sharp)}}{\sqrt{\pi(S_n^\sharp)}} \Big| S_n \right) &(13.28) \\
&\leq 1 - \psi(\pi(S_n)) = 1 - f_0(Z_n), &(13.29)
\end{aligned}$$

where $f_0(z) := \psi(1/z^2)$ is nondecreasing. (Recall that we defined $\psi(x) = \psi_*$ for all real numbers $x \geq \frac{1}{2}$.) Let $L_0 = Z_0 = \pi(x)^{-1/2}$. Next, observe that $\widehat{\mathbf{E}}(\cdot)$ is just the expectation operator with respect to a modified distribution, so we can apply Lemma 13.16 below, with $\widehat{\mathbf{E}}$ in place of $\mathbf{E}$. By part (iii) of that lemma (with $\delta = \sqrt{\epsilon}$), for all

$$n \geq \int_\delta^{L_0} \frac{2dz}{z f_0(z/2)} = \int_\delta^{L_0} \frac{2dz}{z\psi(4/z^2)}, \tag{13.30}$$

we have $\chi(\mu_n, \pi) \leq \widehat{\mathbf{E}}_{\{x\}}(Z_n) \leq \delta$. The change of variable $u = 4/z^2$ shows the integral (13.30) equals

$$\int_{4\pi(x)}^{4/\delta^2} \frac{du}{u\psi(u)} \leq \int_{4\pi(x)}^{4/\epsilon} \frac{du}{u\psi(u)}.$$

This establishes Theorem 13.8.

Lemma 13.16 *Let $f, f_0 : [0,\infty) \to [0,1]$ be increasing functions. Suppose that $\{Z_n\}_{n\geq 0}$ are non-negative random variables with $Z_0 = L_0$. Denote $L_n = \mathbf{E}(Z_n)$.*

(i) *If* $L_n - L_{n+1} \geq L_n f(L_n)$ *for all* n*, then for every* $n \geq \int_\delta^{L_0} \frac{dz}{zf(z)}$*, we have* $L_n \leq \delta$*.*

(ii) *If* $\mathbf{E}(Z_{n+1}|Z_n) \leq Z_n(1 - f(Z_n))$ *for all* n *and the function* $u \mapsto uf(u)$ *is convex on* $(0, \infty)$*, then the conclusion of (i) holds.*

(iii) *If* $\mathbf{E}(Z_{n+1}|Z_n) \leq Z_n(1 - f_0(Z_n))$ *for all* n *and* $f(z) = f_0(z/2)/2$*, then the conclusion of (i) holds.*

Proof: (i) It suffices to show that for every n we have

$$\int_{L_n}^{L_0} \frac{dz}{z\,f(z)} \geq n. \tag{13.31}$$

Note that for all $k \geq 0$ we have

$$L_{k+1} \leq L_k\Big[1 - f(L_k)\Big] \leq L_k e^{-f(L_k)}\,,$$

whence

$$\int_{L_{k+1}}^{L_k} \frac{dz}{zf(z)} \geq \frac{1}{f(L_k)} \int_{L_{k+1}}^{L_k} \frac{dz}{z} = \frac{1}{f(L_k)} \log \frac{L_k}{L_{k+1}} \geq 1.$$

Summing this over $k \in \{0, 1, \ldots, n-1\}$ gives (13.31).

(ii) This is immediate from Jensen's inequality and (i).

(iii) Fix $n \geq 0$. We have

$$\mathbf{E}\left(Z_n - Z_{n+1}\right) \geq \mathbf{E}\left[2Z_n f(2Z_n)\right] \geq L_n f(L_n)\,, \tag{13.32}$$

by Lemma 13.17 below. This yields the hypothesis of (i). ■

The following simple fact was used in the proof of Lemma 13.16.

Lemma 13.17 *Suppose that* $Z \geq 0$ *is a nonnegative random variable and* f *is a nonnegative increasing function. Then*

$$\mathbf{E}\Big(Zf(2Z)\Big) \geq \frac{\mathbf{E}Z}{2} \cdot f(\mathbf{E}Z).$$

Proof: Let A be the event $\{Z \geq \mathbf{E}Z/2\}$. Then $\mathbf{E}(Z\mathbf{1}_{A^c}) \leq \mathbf{E}Z/2$, so $\mathbf{E}(Z\mathbf{1}_A) \geq \mathbf{E}Z/2$. Therefore,

$$\mathbf{E}\Big(Zf(2Z)\Big) \geq \mathbf{E}\Big(Z\mathbf{1}_A \cdot f(\mathbf{E}Z)\Big) \geq \frac{\mathbf{E}Z}{2} f(\mathbf{E}Z)\,.$$

■

13.6 Concluding remarks

1. The example of the lamplighter group in the introduction shows that $\tau_1(1/4)$, the mixing time in total variation on the left-hand side of (13.5), can be considerably smaller than the corresponding uniform mixing time $\tau_u(1/4)$ (so an upper bound for $\tau_u(\cdot)$ is strictly stronger). We note that there are simpler examples of this phenomenon. For lazy random walk on a clique of n vertices, $\tau_1(1/4) = \Theta(1)$ while $\tau_u(1/4) = \Theta(\log n)$. To see a simple example with bounded degree, consider a graph consisting of two expanders of cardinality n and 2^n, respectively, joined by a single edge. In this case $\tau_1(1/4)$ is of order $\Theta(n)$, while $\tau_u(1/4) = \Theta(n^2)$.

2. Let X_n be a finite, reversible chain with transition matrix P. Write $\mu_n^x := p^n(x,\cdot)$. Equation (13.22) gives
$$\chi(\mu_n^x,\pi) \le \frac{1}{\pi(x)}\mathbf{E}\sqrt{\pi(S_n^\sharp)} \le \frac{1}{\sqrt{\pi(x)}}(1-\psi_*)^n\,. \tag{13.33}$$
Let $f_2 : V \to \mathbf{R}$ be the second eigenfunction of P and λ_2 the second eigenvalue, so that $Pf_2 = \lambda_2 f_2$. For $x \in V$, define $f_x : V \to \mathbf{R}$ by $f_x(y) = \delta_x(y) - \pi(y)$, where δ is the Dirac delta function. We can write $f_2 = \sum_{x\in V}\alpha_x f_x$. Hence
$$\begin{aligned}
\Big\|\frac{P^n f(\cdot)}{\pi(\cdot)}\Big\|_{L^2(\pi)} &\le \sum_x \alpha_x \Big\|\frac{P^n f_x(\cdot)}{\pi(\cdot)}\Big\|_{L^2(\pi)} && (13.34)\\
&= \sum_x \alpha_x \chi(\mu_n^x,\pi) && (13.35)\\
&\le \text{const}\cdot\max_x \chi(\mu_n^x,\pi) && (13.36)\\
&\le \text{const}\cdot(1-\psi_*)^n, && (13.37)
\end{aligned}$$
where the first line is subadditivity of a norm and the last line follows from (13.33). But
$$\Big\|\frac{P^n f(\cdot)}{\pi(\cdot)}\Big\|_{L^2(\pi)} \ge \Big\|\frac{P^n f(\cdot)}{\pi(\cdot)}\Big\|_{L^1(\pi)} = \sum_x |P^n f_2(x)| = \lambda_2^n \sum_x |f_2(x)|. \tag{13.38}$$
Combining (13.37) and (13.38) gives $\lambda_2^n \le c\cdot(1-\psi_*)^n$ for a constant c. Since this is true for all n, we must have $\lambda_2 \le 1-\psi_*$, so ψ_* is a lower bound for the spectral gap.

Lecture 14: Evolving sets and strong stationary times

14.1 Evolving sets

Recall the definition of uniform mixing time, $\tau_{unif}(\epsilon)$ introduced in the previous lecture:

$$\tau_{unif}(\epsilon) := \min\left\{t : \left|\frac{p^t(x,y)}{\pi(y)} - 1\right| < \epsilon \ \ \forall x, y \in \Omega\right\}.$$

We use the method of evolving sets to finish the proof of the following upper bound for the uniform mixing time.

Theorem 14.1 *If the chain is reversible and lazy ($\mathbf{P}(x,x) \geq \frac{1}{2}$ for all $x \in \Omega$), then*

$$\tau_{unif}(\epsilon) \leq \frac{2}{\Phi_*^2} \log \frac{1}{\epsilon \pi_{\min}} \tag{14.1}$$

Proof: Let $\{S_n\}$ be an evolving set process started at $\{x\}$ corresponding to the Markov chain with transition matrix P, and let $\{\Lambda_m\}$ be an independent evolving set process started at $\{z\}$ corresponding the reverse Markov chain with transition matrix $\hat{\mathrm{P}}$. From the easy fact

$$\mathrm{P}^{n+m}(x,z) = \sum_{y \in \Omega} \mathrm{P}^n(x,y)\mathrm{P}^m(y,z)$$

and the detailed balance equations, we deduce

$$\frac{\mathrm{P}^{n+m}(x,z)}{\pi(z)} = \sum_{y \in \Omega} \frac{\mathrm{P}^n(x,y)}{\pi(y)} \frac{\hat{\mathrm{P}}^m(z,y)}{\pi(y)} \pi(y).$$

Now, from the previous lecture, we know that $\mathbf{P}(y \in S_n) = \frac{\pi(x)\mathrm{P}^n(x,y)}{\pi(y)}$, and it follows that

$$\begin{aligned}
\frac{\mathrm{P}^{n+m}(x,z)}{\pi(z)} &= \sum_{y \in \Omega} \frac{\mathbf{P}(y \in S_n)}{\pi(x)} \frac{\mathbf{P}(y \in \Lambda_m)}{\pi(z)} \cdot \pi(y) \\
&= \frac{1}{\pi(x)\pi(z)} \mathbb{E} \sum_{y \in \Omega} \pi(y) \mathbf{1}_{\{y \in S_n\}} \mathbf{1}_{\{y \in \Lambda_m\}} \\
&= \frac{1}{\pi(x)\pi(z)} \mathbb{E}[\pi(S_n \cap \Lambda_m)].
\end{aligned}$$

Subtracting 1 from each side of the above equation (recalling that $\pi(S_n)$ is a martingale) and taking absolute values, we obtain:

$$\left|\frac{\mathrm{P}^{n+m}(x,z) - \pi(z)}{\pi(z)}\right| = \left|\frac{1}{\pi(x)\pi(z)} \mathbb{E}[\pi(S_n \cap \Lambda_m) - \pi(S_n)\pi(\Lambda_m)]\right| \tag{14.2}$$

Following from lemma 13.13, we get

$$\left|\frac{\mathrm{P}^{n+m}(x,z) - \pi(z)}{\pi(z)}\right| \leq \frac{1}{\pi(x)\pi(z)} \mathbb{E}\sqrt{\pi(S_n^\sharp)\pi(\Lambda_m^\sharp)}.$$

Theorem 13.7 implies that

$$\frac{\mathbb{E}_{\{x\}}\sqrt{\pi(S_n^\sharp)}}{\sqrt{\pi(x)}} \leq (1-\frac{\Phi_*^2}{2})^n \leq e^{-n\Phi_*^2/2}$$

The same inequality applies for $\frac{\mathbb{E}_{\{x\}}\sqrt{\pi(S_n^\sharp)}}{\sqrt{\pi(x)}}$. Using this as well as the independence of S_n and Λ_m, we obtain:

$$\begin{aligned}
\left|\frac{\mathrm{P}^{n+m}(x,z)-\pi(z)}{\pi(z)}\right| &\leq \frac{\mathbb{E}_{\{x\}}\sqrt{\pi(S_n^\sharp)}}{\pi(x)}\frac{\mathbb{E}_{\{z\}}\sqrt{\pi(\Lambda_m^\sharp)}}{\pi(z)} \\
&\leq \frac{e^{-n\Phi_*^2/2}}{\sqrt{\pi(x)}}\frac{e^{-m\Phi_*^2/2}}{\sqrt{\pi(z)}} \\
&\leq \frac{-e^{(n+m)\Phi_*^2/2}}{\pi_{\min}},
\end{aligned}$$

from which the theorem follows. ■

14.2 Stationary stopping times

In this section, we define stationary stopping times and strong stationary stopping times, and give several examples.

Definition 14.2 *A random stopping time T for a chain X_i is a* stationary stopping time *(starting at x) if*

$$\mathbf{P}_x(X_T \in A) = \pi(A)$$

for all $A \subset \Omega$ and all $x \in \Omega$. The stopping time T is said to be a strong stationary stopping time *(starting at x) if*

$$\mathbf{P}_x(X_T \in A, T = k) = \pi(A)\mathbf{P}_x(T = k)$$

for all A, x and k.

To see the connection between strong stationary times and mixing, note that if T is a strong stationary stopping time, then

$$\left|1-\frac{\mathrm{P}^n(x,y)}{\pi(y)}\right| \leq \mathbf{P}(T > n).$$

Averaging over the states $y \in \Omega$ weighted according to the stationary distribution, we find

$$||\pi - \mathrm{P}^n(x,\cdot)||_{TV} \leq \mathbf{P}(T > n).$$

We now give some examples of strong stationary times.

Example 14.3 Top to random insertion
Consider a deck of n cards. After successive time intervals, remove the top card from the deck and place it in a uniform random position in the deck. Let T be the first time the original bottom card in the deck reaches the top and is then randomized. Since all the cards below the original bottom card are always in a uniform random order, it is easy to see that T is a strong stationary time. By

considering the expected time for the original bottom card to move up a position once it is in the k^{th} postion from the bottom, and then summing over k, we obtain:

$$\mathbb{E}T = \sum_{k=0}^{n-1} \frac{n}{k+1} \sim n \log n.$$

The following lemma and proof from [1, lemma 2] shows that $\lim_{n\to\infty} \mathbf{P}(T > n \log n + cn) \leq e^{-c}$.

Lemma 14.4 *Sample uniformly with replacement from an urn with n balls. Let V be the number of draws required until each ball has been selected at least once. Then $\mathbf{P}(V > n \log n + cn) \leq e^{-c}$, where $c \geq 0$ and $n \geq 1$.*

Proof: Let $m = n \log n + cn$. For each ball b, let A_b be the event that the b^{th} ball is not selected in the first m draws. Then

$$\mathbf{P}(V > m) = \mathbf{P}\left(\cup_b A_b\right) \leq \sum_b \mathbf{P}(A_b) = n\left(1 - \frac{1}{n}\right)^m$$
$$\leq ne^{-m/n} = e^{-c}.$$

■

To see that there is a cutoff at $n \log n$, consider the events $A_j(t)$ that the bottom j cards are in their original order at time t. The probability that the j^{th} card from the bottom has reached the top in time $t_\epsilon = (1-\epsilon)n \log n$ is very small, so $\mathbf{P}^t(A_j(t_\epsilon)) - \pi(A_j) \sim 1 - \frac{1}{j!}$. For the proof of this fact, we follow [1]. Define T_ℓ to be the first time when the card originally in the ℓ^{th} position (counting from the top) is placed underneath the original bottom card. Observe that $\mathbf{P}(A_j(t_\epsilon)) \geq \mathbf{P}(T - T_{j-1} > t_\epsilon)$, since $T - T_{j-1}$ is distributed as the time for the card initially j^{th} from the bottom to reach the top and be randomized. We prove that $\mathbf{P}(T - T_{j-1} \leq t_\epsilon) \to 0$ as $n \to \infty$, where j is fixed. Observe that $\mathbb{E}(T_{i+1} - T_i) = \frac{n}{i+1}$ and $\mathrm{Var}(T_{i+1} - T_i) = \left(\frac{n}{i+1}\right)^2 \left(1 - \frac{i+1}{n}\right)$ and sum over i to obtain $\mathbb{E}(T - T_j) = n \log n + O(n)$ and $\mathrm{Var}(T - T_j) = O(n^2)$. The claim now follows from Chebyshev's inequality.

Example 14.5 Riffle shuffle (Gilbert, Shannon, Reeds)
Break a deck of n cards into two piles of size $B(n, 1/2)$ and $n - B(n, 1/2)$. Then merge them uniformly, preserving the order of the cards within each respective pile. If the cards are labelled with 1's and 0's according to the cut, the resulting ordering gives a uniform sequence of binary bits. The reverse shuffle, which yields the same probability distribution, has the following simple discription: assign all the cards uniform random bits, and then move the cards with 0's to the top of the deck preserving their order. After this process has been repeated k times, each card has been assigned k uniform binary bits. It is easy to see that the relative order of two cards with distinct binary sequences is uniquely determined, and the first time T that each card has been assigned a unique binary sequence is a strong stationary time. Since $\mathbf{P}(T > t) \leq \frac{n^2}{2} \cdot 2^{-t}$, it follows that $\tau_{unif}(\epsilon) \leq 2 \log_2(n/\epsilon)$. A lower lower bound of the same order can be obtained by computing the probability that the resulting permutation contains an increasing sequence of length $10\sqrt{n}$.

Remark 14.6 Consider the simple random walk on the cycle $\mathbb{Z}_n$. The *cover time*, τ_c is defined as follows

$$\tau_c := \min\left\{t : \{X_0, \ldots, X_t\} = \mathbb{Z}_n\right\}.$$

The cycle has the property that for any starting state the distribution of X_{τ_c} is uniform off the starting position. This property is trivially true for the complete graph as well, and a remarkable theorem due to Lovász and Winkler [21] establishes that these are the only two graphs with this property. Note that there are other graphs that have this property for *some* starting states (e.g. a star).

Lecture 15: Hitting and cover times for irreducible MC and Lamplighter graphs

15.1 Relations of hitting time and cover time

In a Markov chain on a finite state space, we define $t_* := \max_{a,b} \mathbb{E}_a(\tau_b)$ to be the *maximal hitting time*. Define $t_\pi = t_\pi(a) := \sum_b \mathbb{E}_a(\tau_b)\pi(b)$, where π is the stationary distribution.

Theorem 15.1 *t_π is independent of its parameter a.*

For the proof, see the "Random target lemma" (Lemma 29 in Chapter 2) in [2], or following the exercise.

Exercise 15.2 *Check that t_π is a harmonic function of a, which means*

$$t_\pi(a) = \sum_z \mathrm{P}(a,z) t_\pi(z).$$

Since the only harmonic functions are the constant functions, t_π is independent of a.

Let $\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda n$ be the eigenvalues of the transition matrix P. If the chain is reversible, then $\lambda_i < 1$ for $i \geq 2$.

Lemma 15.3

$$t_\pi = \sum_{i>1} \frac{1}{1-\lambda_i}$$

See the "Eigentime identity" in [2].

By lemma 15.3, we have:

$$t_\pi = \sum_{i>1} \frac{1}{1-\lambda_i} \geq \frac{1}{1-\lambda_2} \tag{15.1}$$

In other words, t_π is at least the relaxation time.

Consider the *cover time* $e_n := \max_a \tau_a$ on an n-state chain, i.e. the time required to visit every state. The following result bounds the expected cover time.

Theorem 15.4 *(Matthews bound)*

$$\mathbb{E}_a e_n \leq t_*(1 + \frac{1}{2} + \cdots + \frac{1}{n}) \sim t_* \log n$$

Proof: Start the chain at a. Let $\{J_1, \ldots, J_n\}$ be a uniform random permutation of the states. Let L_k be the last state of $J_1, \ldots, J_k$ to be visited by the chain, and write $T_k = \tau_{L_k}$ for the hitting time of L_k. Then $T_n = e_n$.

Consider $\mathbb{E}_a(T_k - T_{k-1}|T_{k-1}, X_1, \ldots, X_{T_{k-1}})$. The event $\{J_k = L_k\}$ belongs to $\sigma(T_{k-1}, X_1, \ldots, X_{T_{k-1}})$. On the event $\{J_k \neq L_k\}$, we have $T_k - T_{k-1} = 0$; and on the event $\{J_k = T_k\}$, we have $T_k - T_{k-1} = \mathbb{E}_{L_{k-1}}(\tau_{J_k})$. So

$$\begin{aligned}\mathbb{E}_a(T_k - T_{k-1}|T_{k-1}, X_1, \ldots, X_{T_{k-1}}) &= \mathbf{1}_{\{J_k = L_k\}} \mathbb{E}_{L_{k-1}}(\tau_{J_k}) \\ &\leq \mathbf{1}_{\{J_k = L_k\}} \cdot t_*\end{aligned}$$

Take expectations, we get:

$$\mathbb{E}_a(T_k - T_{k-1}) \leq t_* \mathbf{P}(J_k = L_k) = \frac{1}{k} t_* \tag{15.2}$$

Summing over k gives the desired inequality. ■

On the torus $\mathbb{Z}_n^d$ for $d \geq 3$, this upper bound for $\mathbb{E}e$ is also the asymptotic formula for $\mathbb{E}e$. See Corollary 24 in Chapter 7 of [2]. For $d = 2$, Aldous-Lawler proved that for $n \times n$ torus $\mathbb{Z}_n^2$, the expectation is bounded by:

$$\frac{2}{\pi} n^2 \log^2 n \leq \mathbb{E}e \leq \frac{4}{\pi} n^2 \log^2 n.$$

Dembo, Peres, Rosen and Zeitouni later proved that the upper bound is the asymptotic order:

$$\mathbb{E}e \sim \frac{4}{\pi} n^2 \log^2 n$$

15.2 Lamplighter graphs

Given a finite graph $G = (V, E)$, the wreath product $G^* = \{0,1\}^V \times V$ is the graph whose vertices are ordered pairs (f, x), where $x \in V$ and $f \in \{0,1\}^V$. There is an edge between (f, x) and (h, y) in the graph G^* if x, y are adjacent in G and $f(z) = h(z)$ for $z \notin \{x, y\}$. These wreath products are called *lamplighter graphs* because of the following interpretation: place a lamp at each vertex of G; then a vertex of G^* consists of a configuration f indicating which lamps are on, and a lamplighter located at a vertex $x \in V$.

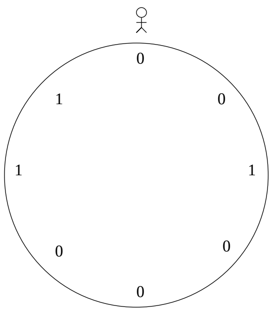

Figure 15.1: Lamplighter graphs

The random walk we analyze on G^* is constructed from a random walk on G as follows. Let p denote the transition probabilities in the wreath product and q the transition probabilities in G. For $a \neq b$, $p[(f, a), (h, b)] = q(a, b)/4$ if f and h agree outside of $\{a, b\}$, and when $a = b$, $p[(f, a), (h, a)] = q(a, a)/2$ if f and h agree off of $\{a\}$. A more intuitive description of this is to say that at each time step, the current lamp is randomized, the lamplighter moves, and then the new lamp is also randomized. The second lamp at b is randomized in order to make the chain reversible.

To avoid periodicity problems, we will assume that the underlying random walk on G is already aperiodic.

In the following paragraphs, we give bounds for the mixing time $\tau_1(\epsilon)$, the relaxation time τ_{rel} and the uniform mixing time $\tau_u(\epsilon)$ on this lamplighter graph G^*. Recall the definition of these three mixing times:

$$\tau_{rel} = \max_{i:|\lambda_i|<1} \frac{1}{1-|\lambda_i|}; \tag{15.3}$$

$$\tau_1(\epsilon) = \min\left\{t : \frac{1}{2}\sum_y |p^t(x,y) - \mu(y)| \le \epsilon \ \forall\ x \in G\right\}; \tag{15.4}$$

$$\tau_u(\epsilon) = \min\left\{t : \left|\frac{p^t(x,y) - \mu(y)}{\mu(y)}\right| \le \epsilon \ \forall\ x, y \in G\right\}. \tag{15.5}$$

They satisfy the relations

$$\tau_{rel} \le \tau_1(\epsilon) \le \tau_u(\epsilon).$$

Let $\{G_n\}$ be a sequence of transitive graphs and let G_n^* be the lamplighter graph of G_n. Suppose $t_*(G_n) = o(\mathbb{E}e(G_n))$ as $n \to \infty$. An example is $G_n = \mathbb{Z}_n^d$ with $d \ge 2$. The following three theorems are given by Peres and Revelle.

Theorem 15.5 *With the definitions above, as $|G_n|$ goes to infinity,*

$$\frac{1}{8\log 2} \le \frac{\tau_{rel}(G_n^*)}{t_*(G_n)} \le \frac{2}{\log 2} + o(1). \tag{15.6}$$

Theorem 15.6 *Let $\{G_n\}$ be a sequence of vertex transitive graphs with $|G_n| \to \infty$, and e_n denote the cover time for simple random walk on G_n. For any $\epsilon > 0$, there exist constants c_1, c_2 depending on ϵ such that the total variation mixing time satisfies*

$$[c_1 + o(1)]\mathbb{E}e_n \le \tau_1(\epsilon, G_n^*) \le [c_2 + o(1)]\mathbb{E}e_n. \tag{15.7}$$

Moreover, if the maximal hitting time satisfies $t_* = o(\mathbb{E}e_n)$, then for all $\epsilon > 0$,

$$\left[\frac{1}{2} + o(1)\right]\mathbb{E}e_n \le \tau_1(\epsilon, G_n^*) \le [1 + o(1)]\mathbb{E}e_n. \tag{15.8}$$

Aldous [2] (Theorem 33 of chapter 6) showed that the condition $t_* = o(\mathbb{E}e_n)$ implies that the cover time has a sharp threshold, that is $e_n/\mathbb{E}e_n$ tends to 1 in probability. Theorem 15.6 thus says that in situations that give a sharp threshold for the cover time of G_n, there is also a threshold for the total variation mixing time on G_n^*, although the factor of 2 difference between the bounds means that we have not proved a sharp threshold.

Theorem 15.7 *Let $\{G_n\}$ be a sequence of regular graphs for which $|G_n| \to \infty$ and the maximal hitting time satisfies $t_* \le K|G_n|$ for some constant K. Then there are constants c_1, c_2 depending on ϵ and K such that*

$$c_1|G_n|(\tau_{rel}(G_n) + \log|G_n|) \le \tau_u(\epsilon, G_n^*) \le c_2|G_n|(\tau_1(G_n) + \log|G_n|). \tag{15.9}$$

Consider the 2-dimensional torus $\mathbb{Z}_n^2$.

$$t_*(\mathbb{Z}_n^2) \sim \frac{8}{\pi} n^2 \log n$$

$$\mathbb{E}e(\mathbb{Z}_n^2) \sim \frac{4}{\pi} n^2 \log^2 n$$

Results for this case are showed as the following theorem.

Theorem 15.8 *For the random walk $\{X_t\}$ on $(\mathbb{Z}_n^2)^*$, the relaxation time satisfies*

$$\frac{1}{\pi \log 2} \leq \frac{\tau_{rel}((\mathbb{Z}_n^2)^*)}{n^2 \log n} \leq \frac{16}{\pi \log 2} + o(1). \tag{15.10}$$

For any $\epsilon > 0$, the total variation mixing time satisfies

$$\lim_{n\to\infty} \frac{\tau_1(\epsilon, (\mathbb{Z}_n^2)^*)}{n^2 \log^2 n} = \frac{8}{\pi}, \tag{15.11}$$

and the uniform mixing time satisfies

$$C_2 \leq \frac{\tau_u(\epsilon, (\mathbb{Z}_n^2)^*)}{n^4} \leq C_2' \tag{15.12}$$

for some constants C_2 and C_2'.

To see the lower bound of equation (15.12), note that by the central limit theorem,

$$\mathbf{P}(\text{lamplighter stays in lower half for time } kn^2) \geq c^k$$

Let A be the event that all the lamps in upper half are off. If $kn^2 = \tau_u(\epsilon, (\mathbb{Z}_n^2)^*)$, then

$$1 + \epsilon \geq \frac{P^{kn^2}(A)}{\pi(A)} \geq \frac{c^k}{2^{-n^2/2}}$$

This lead to $k \geq C_2 n^2$ with some constant C_2. So $\tau_u(\epsilon, (\mathbb{Z}_n^2)^*) \geq C_2 n^4$.

The factor of 2 difference between the upper and lower bounds in (15.8) comes from the question of whether or not it suffices to cover all but the last $\sqrt{n}$ sites of the graph. For many graphs, the amount of time to cover all but the last $\sqrt{|G_n|}$ sites is $\mathbb{E}e_n/2$, which will be the lower bound of (15.8). When the unvisited sites are clumped together instead of being uniformly distributed, it will turn out to be necessary to visit all the sites, and the upper bound of (15.8) will be sharp. In $\mathbb{Z}_n^2$, at time $(1-\delta)\mathbb{E}e$, the set of unvisited points in high probability has holes of radius $> \ n^{\delta'}$.

Proof:[Proof of Theorem 15.5] For the lower bound, we will use the variational formula for the second eigenvalue:

$$1 - |\lambda_2| = \min_{\varphi:\mathbf{V}ar\varphi>0} \frac{\mathcal{E}(\varphi,\varphi)}{\mathbf{V}ar\varphi}, \tag{15.13}$$

For the lower bound of (15.6), we use (15.13) to show that the spectral gap for the transition kernel p^t is bounded away from 1 when $t = t_*/4$. Fix a vertex $o \in G$, and let $\varphi(f,x) = f(o)$. Then $\mathbf{V}ar\varphi = 1/4$ and by running for t steps,

$$\mathcal{E}(\varphi,\varphi) = \frac{1}{2}\mathbb{E}\left[\varphi(\xi_t) - \varphi(\xi_0)\right]^2 = \frac{1}{2}\sum_{x\in G} \nu(x) \frac{1}{2}\mathbf{P}_x[T_o < t],$$

where ν is the stationary measure on G, and $\{\xi_t\}$ is the stationary Markov chain on G^*. For any t,

$$\mathbb{E}_x T_o \leq t + t_*(1 - \mathbf{P}_x[T_o < t]).$$

For a vertex transitive graph, we have by Lemma 15 in Chapter 3 of [2], that

$$t_* \leq 2 \sum_{x \in G} \nu(x) \mathbb{E}_x T_o.$$

Let $\mathbb{E}_\nu = \sum_x \nu(x) \mathbb{E}_x$ and $\mathbf{P}_\nu = \sum_x \nu(x) \mathbf{P}(x)$. Then

$$t_* \leq 2\mathbb{E}_\nu T_o \leq 2t + 2t_*[1 - \mathbf{P}_\nu(T_o < t)].$$

Substituting $t = t_*/4$ yields

$$\mathbf{P}_\nu[T_0 < t_*/4] \leq \frac{3}{4}.$$

We thus have

$$1 - |\lambda_2|^{t_*/4} \leq \frac{3}{4},$$

and so

$$\log 4 \geq \frac{t_*}{4}(1 - |\lambda_2|),$$

which gives the claimed lower bound on $\tau_{rel}(G^*)$.

For the upper bound, we use a coupling argument from [6]. Suppose that φ is an eigenfunction for p with eigenvalue λ_2. To conclude that $\tau_{rel}(G^*) \leq \frac{(2+o(1))t_*}{\log 2}$, it suffices to show that $\lambda_2^{2t_*} \leq 1/2$. For a configuration h on G, let $|h|$ denote the Hamming length of h. Let

$$M = \sup_{f,g,x} \frac{|\varphi(f,x) - \varphi(g,x)|}{|f - g|}$$

be the maximal amount that φ can vary over two elements with the same lamplighter position. If $M = 0$, then $\varphi(f,x)$ depends only on x, and so $\psi(x) = \varphi(f,x)$ is an eigenfunction for the transition operator on G. Since $\tau_{rel}(G) \leq t_*$ (see [2], Chapter 4), this would imply that $|\lambda_2^{2t_*}| \leq e^{-4}$. We may thus assume that $M > 0$.

Consider two walks, one started at (f,x) and one at (g,x). Couple the lamplighter component of each walk and adjust the configurations to agree at each site visited by the lamplighter. Let (f',x') and (g',x') denote the position of the coupled walks after $2t_*$ steps. Let K denote the transition operator of this coupling. Because φ is an eigenfunction,

$$\begin{aligned}
\lambda_2^{2t_*} M &= \sup_{f,g,x} \frac{|p^{2t_*}\varphi(f,x) - p^{2t_*}\varphi(g,x)|}{|f-g|} \\
&\leq \sup_{f,g,x} \sum_{f',g',x'} K^{2t_*}[(f,g,x) \to (f',g',x')] \frac{|\varphi(f',x') - \varphi(g',x')}{|f'-g'|} \frac{|f'-g'|}{|f-g|} \\
&\leq M \sup_{f,g,x} \frac{\mathbb{E}|f'-g'|}{|f-g|}.
\end{aligned}$$

But at time $2t_*$, each lamp that contributes to $|f - g|$ has probability of at least $1/2$ of having been visited, and so $\mathbb{E}|f' - g'| \leq |f - g|/2$. Dividing by M gives the required bound of $\lambda_2^{2t_*} \leq 1/2$. ■

Lecture 16: Ising Model on Trees

16.1 Detecting λ_2

Given a Markov chain on finite state space with transition matrix P, suppose we have found an eigenfunction $\mathrm{P}f = \lambda f$. We'd like to know how to check whether $\lambda = \lambda_2$, the second largest eigenvalue.

Recall that the chain is a *monotone system* if there is a partial ordering $\leq$ on the state space, such that for any states $x \leq y$, there is a coupling (X, Y) of the probability measures $\delta_x\mathrm{P}$ and $\delta_y\mathrm{P}$ with the property that that $X \leq Y$. Note that if f is a function on the state space which is increasing with respect to this partial ordering, then monotonicity implies that for any $x \leq y$,

$$\mathrm{P}f(x) = \mathbb{E}f(X) \leq \mathbb{E}f(Y) = \mathrm{P}f(y)$$

hence $\mathrm{P}f$ is also increasing.

Lemma 16.1 *In the monotone and reversible case (e.g. the Ising model) the equation*

$$\mathrm{P}f = \lambda_2 f$$

has an increasing solution f.

Proof: Let $\{f_i\}_{i=1}^n$ be a basis of eigenfunctions for P. Since the strictly increasing functions of mean zero ($\int f d\pi = 0$) form an open subset of the set of all mean zero functions, we can find a strictly increasing function $h = \sum a_i f_i$ with $a_1 = 0, a_2 \neq 0$. Now consider the sequence of functions

$$h_m = \lambda_2^{-m}\mathrm{P}^m h = a_2 f_2 + \sum_{i=3}^{n} \left(\frac{\lambda_i}{\lambda_2}\right)^m f_i.$$

By monotonicity, h_m is increasing for each m. Since $\lambda_2 > |\lambda_i|$ for $i \geq 3$, the sequence h_m converges to $a_2 f_2$ as $m \to \infty$, and hence f_2 is increasing. ∎

The converse holds as well: if $\mathrm{P}f = \lambda f$ and f is increasing and nonconstant, then $\lambda = \lambda_2$. See Serban Nacu's paper in PTRF, "Glauber dynamics on the cycle is monotone." Proof sketch: If $\mathrm{P}f_2 = \lambda_2 f_2$ and $\mathrm{P}f = \lambda f$ with both f, f_2 increasing, and $\lambda \neq \lambda_2$, one can use the FKG inequality to show $\int f f_2 d\pi > 0$, contradicting orthogonality.

Open Question: Find other criteria that imply $\lambda = \lambda_2$ (in the absence of monotonicity).

16.2 Positive Correlations

A probability measure μ on partially ordered set Ω has *positive correlations* if for any increasing f, g we have $\int fg d\mu \geq \int f d\mu \int g d\mu$.

Lemma 16.2 (Chebyshev) *If Ω is totally ordered, then any probability measure μ has positive correlations.*

Proof: This was the first historical use of a coupling argument. Given increasing functions f, g on Ω, and random variables X, Y with distribution μ, the events $f(X) \le f(Y)$ and $g(X) \le g(Y)$ coincide, hence

$$(f(X) - f(Y))(g(X) - g(Y)) \ge 0.$$

Integrate $d\mu(x)d\mu(y)$ to get

$$\int f(x)g(x)d\mu(x) - \int f(x)d\mu(x) \int g(y)d\mu(y) \ge 0.$$

■

Lemma 16.3 (Harris Inequality) *Any product measure with independent components that are totally ordered has positive correlations, using the coordinatewise partial order on the product.*

Proof: It suffices to check that if μ_i has positive correlations on Ω_i for $i = 1, 2$, then $\mu_1 \times \mu_2$ has positive correlations on $\Omega_1 \times \Omega_2$. If f, g are increasing on the product space, we have

$$\begin{aligned} \int\int f(x,y)g(x,y)d\mu_1(x)d\mu_2(y) &\ge \int \left[\int f(x,y)d\mu_1(x)\right]\left[\int g(x,y)d\mu_1(x)\right] d\mu_2(y) \\ &\ge \int\int f d\mu_1 d\mu_2 \int\int g d\mu_1 d\mu_2. \end{aligned}$$

■

16.3 Glauber Dynamics

Consider the Ising model on a finite graph G. Denote by $\Omega = \{+, -\}^G$ the set of spin configurations. Starting from a fixed configuration σ, run the Glauber dynamics using a *systematic scan*: Fix an ordering of the vertices, and update in this order. So there is no randomness involved in how we choose the vertices to update. We update using uniform i.i.d. $U_j \in [0, 1]$ in order to get a monotone system.

Lemma 16.4 *No matter what the initial configuration σ, if μ_t is the distribution after t steps, then μ_t has positive correlations.*

Proof: Starting from σ, the new state γ can be written $\gamma = \Gamma(U_1, \ldots, U_t)$ with $\Gamma : [0, 1]^t \to \Omega$ increasing. Given increasing functions f, g on Ω, the compositions $f \circ \Gamma$ and $g \circ \Gamma$ are increasing on $[0, 1]^t$. By Lemma 16.3 we obtain

$$\begin{aligned} \int fg\ d\mu_t &= \int (f \circ \Gamma)(g \circ \Gamma) dU_1 \ldots dU_t \\ &\ge \int f \circ \Gamma\ dU_1 \ldots dU_t \int g \circ \Gamma\ dU_1 \ldots dU_t \\ &= \int f d\mu_t \int g d\mu_t. \end{aligned}$$

■

Now suppose our underlying graph G is a regular b-ary tree of depth r. We determine the outcome of an update at a vertex v by a Bernoulli variable

$$\sigma_v = \begin{cases} \sigma_w, & \text{with probability } 1-2\epsilon \\ -\sigma_w, & \text{with probability } 2\epsilon, \end{cases}$$

where w is the parent of v.

The parameter ϵ is related to the inverse temperature β by $\frac{\epsilon}{1-\epsilon} = e^{-2\beta}$.

The rest of this chapter is concerned with answering the following questions: Under what conditions on ϵ do we get order $n \log n$ mixing for the Glauber dynamics, and under what conditions do we get order $\frac{1}{n}$ spectral gap, where $n = 1 + b + \ldots + b^r$ is the number of vertices in the tree? The answer to both questions is: if and only if $\Theta := 1 - 2\epsilon < \frac{1}{\sqrt{b}}$.

Remark. A different transition occurs at $\Theta = \frac{1}{b}$. When do the spins at depth k affect the root? For path coupling, we need

$$(b+1)\Theta = (b+1)\tanh(\beta) < 1,$$

or $\Theta < \frac{1}{b+1}$. This can be improved to $\Theta < \frac{1}{b}$ using block dynamics (update small subtrees at random).

For the spectral gap, one direction is easy. If $\Theta > \frac{1}{\sqrt{b}}$, we get a gap $< n^{-1-\delta}$ using the variational principle with test function

$$S_k(\sigma) = \sum_{\text{level}(v)=k} \sigma_v.$$

This gives

$$\text{gap} \leq \frac{\mathcal{E}(S_k, S_k)}{\mathbf{V}ar(S_k)} \asymp \frac{1}{\mathbf{V}ar(S_k)}.$$

To estimte the variance, use the fact that

$$\mathbb{E}\sigma_v\sigma_w = \Theta^{\text{dist}(v,w)}.$$

See Berger-Kenyon-Mossel-Peres for the calculation.

In the case $\Theta < \frac{1}{\sqrt{b}}$, we will get bounds on the mixing time and spectral gap by proving a contraction in block dynamics. The trick is to use a weighted metric, with weight $\frac{1}{\sqrt{b^j}}$ at level j. So a spin defect at level j contributes $\frac{1}{\sqrt{b^j}}$ to the distance. Gives a contraction in block dynamics if $\Theta < \frac{1}{\sqrt{b}}$.

A correlation inequality is needed to prove the contraction. In *any* tree, given boundary conditions (fixed spins) η.

$$\mathbb{E}_\eta(\sigma_v|\sigma_w = 1) - \mathbb{E}_\eta(\sigma_v|\sigma_w = -1) \leq \mathbb{E}(\sigma_v|\sigma_w = 1) - \mathbb{E}(\sigma_v|\sigma_w = -1), \tag{16.1}$$

where $\mathbb{E}_\eta$ denotes expectation conditional on the boundary conditions. In words, the effect of a flip is maximized when there is no conditioning. The same is true if the boundary condtions are replaced by an external field. We know that all of this holds for trees, but the corresponding questions are open for arbitrary graphs!

To prove the contraction in block dynamics, start with a single defect $\sigma(u) = +1$, $\tau(u) = -1$ at level ℓ. In our weighted metric, $d(\sigma, \sigma') = (\frac{1}{\sqrt{b}})^\ell$. Choose a directed subtree (block) T of height h. If T contains u, we can remove the defect. Since there are h blocks containing u (one for each ancestor of u removed by h or fewer generations) the distance decreases by $(\frac{1}{\sqrt{b}})^\ell$ with probability $\frac{h+1}{n}$.

The distance increases if T is rooted at a child of u. There are b such blocks. In this case, we use the correlation inequality (16.1) to remove all boundary conditions other than u. Then the expected increase in distance is at most

$$\sum_{j=1}^{h} b^j \Theta^j (\frac{1}{\sqrt{b}})^{\ell+j} \le \frac{b^{-\ell/2}}{1-\Theta\sqrt{b}}$$

The distance also increases if T is rooted at the ancestor exactly $h+1$ generations above u. A similar calculation applies in this case.

Putting things together, the total expected change in distance is

$$\mathbb{E}(d(\sigma,\tau) - d(\sigma',\tau')) \le \frac{1}{n}\left(\frac{b^{1-\ell/2}}{1-\Theta\sqrt{b}} - (h+1)b^{-\ell/2}\right).$$

Taking the block height h sufficiently large, we obtain

$$\mathbb{E}(d(\sigma,\tau) - d(\sigma',\tau')) \le \frac{-cb^{-\ell/2}}{n} = \frac{-c}{n} d(\sigma,\tau),$$

for some positive constant c. By path coupling, we conclude that the block dynamics have mixing time $O(n \log n)$ and spectral gap $O(\frac{1}{n})$.

To get a result for the single-site dynamics, use horizontal censoring lines spaced by h. Shift the censoring lines after running for a while, to get rid of boundary effects. The censored single-site dynamics, run for a long time, closely approximate the block dynamics, which contract. So the censored single-site dynamics also contract.

16.4 Censoring Inequality

To get a contraction for the uncensored single-site dynamics, we will use a "censoring inequality" of Peres and Winkler.

Write $\mu \preceq \nu$ if ν stochastically dominates μ (i.e. $\int f d\mu \le \int f d\nu$ for all increasing functions f)

Theorem 16.5 (Peres-Winkler) *For the Ising model and other monotone systems, starting from the maximal state (all +), let μ be the distribution resulting from updates at sites $v_1,\ldots,v_m$, and let ν be the distribution resulting from updates at a subsequence $v_{i_1},\ldots,v_{i_k}$. Then $\mu \preceq \nu$, and*

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

In words, censoring updates always brings the distribution further away from stationarity.

The proof relies on monotonicity. The analogous question is open for nonmonotone systems like the Potts model.

By induction, we can assume μ was updated at $v_1,\ldots,v_{j-1},v_{j+1},\ldots,v_m$. To prove the censoring inequality, we will establish a stronger fact by induction: $\frac{\mu}{\nu}, \frac{\mu}{\pi}, and \frac{\nu}{\pi}$ are all increasing.

Given a spin configuration σ, a vertex v and a spin s, denote by σ_v^s the configuration obtained from σ by changing the spin at v to s. Write $\sigma_v^\bullet = \{\sigma_v^s\}_{s\in S}$ for the set of spin configurations that are identical to σ except possibly at v. Given a distribution μ, denote by μ_v the distribution resulting from an update at v. Then

$$\mu_v(\sigma) = \frac{\pi(\sigma)}{\pi(\sigma_v^\bullet)}\mu(\sigma_v^\bullet). \tag{16.2}$$

Lemma 16.6 *For any distribution μ, if $\frac{\mu}{\pi}$ is increasing, then $\frac{\mu_v}{\pi}$ is also increasing for any site v.*

Proof: Define $f : S^V \to \mathbf{R}$ by

$$f(\sigma) := \max\Big\{\frac{\mu(\omega)}{\pi(\omega)} \ : \ \omega \in \Omega, \ \omega \leq \sigma\Big\} \tag{16.3}$$

with the convention that $f(\sigma) = 0$ if there is no $\omega \in \Omega$ satisfying $\omega \leq \sigma$. Then f is increasing on S^V, and f agrees with μ/π on Ω.

Let $\sigma < \tau$ be two configurations in Ω; we wish to show that

$$\frac{\mu_v}{\pi}(\sigma) \leq \frac{\mu_v}{\pi}(\tau). \tag{16.4}$$

Note first that for any $s \in S$,

$$f(\sigma_v^s) \leq f(\tau_v^s)\,,$$

since f is increasing. Furthermore, $f(\tau_v^s)$ is an increasing function of s. Thus, by (16.2),

$$\begin{aligned} \frac{\mu_v}{\pi}(\sigma) &= \frac{\mu(\sigma_v^\bullet)}{\pi(\sigma_v^\bullet)} = \sum_{s\in S} f(\sigma_v^s)\frac{\pi(\sigma_v^s)}{\pi(\sigma_v^\bullet)} \\ &\leq \sum_{s\in S} f(\tau_v^s)\frac{\pi(\sigma_v^s)}{\pi(\sigma_v^\bullet)} \leq \sum_{s\in S} f(\tau_v^s)\frac{\pi(\tau_v^s)}{\pi(\tau_v^\bullet)} = \frac{\mu_v}{\pi}(\tau)\,, \end{aligned}$$

where the last inequality follows from the stochastic domination guaranteed by monotonicity of the system. ■

Lemma 16.7 *For any μ, ν such that $\frac{\nu}{\pi}$ is increasing, and $\nu \preceq \mu$, we have*

$$||\nu - \pi|| \leq ||\mu - \pi||.$$

Proof: Let $A = \{\sigma \ : \ \nu(\sigma) > \pi(\sigma)\}$. Then 1_A is increasing, so

$$||\nu - \pi|| = \sum_{\sigma\in A}(\nu(\sigma) - \pi(\sigma)) = \nu(A) - \pi(A) \leq \mu(A) - \pi(A) \leq ||\mu - \pi||.$$

■

Lemma 16.8 *If the set of spins S is totally ordered, and α and β are probability distributions on S such that $\frac{\alpha}{\beta}$ is increasing, and $\beta > 0$ on S, then $\alpha \succeq \beta$.*

Proof: If g is an increasing function on S, then by Chebyshev's result on positive correlations (Lemma 16.2) we have

$$\begin{aligned} \sum g(s)\alpha(s) &= \sum g(s)\frac{\alpha(s)}{\beta(s)}\alpha(s) \\ &\geq \sum g(s)\beta(s)\sum \frac{\alpha(s)}{\beta(s)}\beta(s) \\ &= \sum g(s)\beta(s). \end{aligned}$$

■

Lemma 16.9 *If $\frac{\mu}{\pi}$ is increasing, then $\mu \succeq \mu_v$ for all sites v.*

Proof: This is immediate from Lemma 16.8. ■

Proof:[Proof of Theorem 16.5] Let μ^0 be the distribution concentrated at the top configuration, and $\mu^i = (\mu^{i-1})_{u_i}$ for $i \geq 1$. Applying Lemma 16.6 inductively, we have that each μ^i/π is increasing, for $0 \leq i \leq k$. In particular, we see from Lemma 16.9 that $\mu^{j-1} \succeq (\mu_{j-1})_{u_j} = \mu^j$.

If we define ν^i in the same manner as μ^i, except that $\nu^j = \nu^{j-1}$, then because stochastic dominance persists under updates, we have $\nu^i \succeq \mu^i$ for all i; when $i = k$, we get $\mu \preceq \nu$ as desired.

For the second statement of the theorem, we apply Lemma 16.7, noting that ν^k/π is increasing by the same inductive argument used for μ. ■

References

[1] Aldous, D. and Diaconis, P. Shuffling cards and stopping times, *American Mathematical Monthly*, vol. 93 **5** (1986), 333-348.

[2] D. Aldous and J. Fill, Reversible Markov Chains and Random Walks on Graphs, (draft version avaible online at http://www.stat.berkeley.edu/ aldous/RWG/book.html)

[3] Alon, N. (1986). Eigenvalues and expanders. *Combinatorica* **6**, 83–96.

[4] Alon, N. and Milman, V. D. (1985). λ_1, Isoperimetric inequalities for graphs and superconcentrators, *J. Combinatorial Theory Ser. B* **38**, 73–88.

[5] Benjamini, I. and Mossel, E. (2003). On the mixing time of a simple random walk on the super critical percolation cluster. *Probab. Th. Rel. Fields* **125**, 408–420.

[6] Chen, M. *Trilogy of couplings and general formulas for lower bound of spectral gap*, Probability towards 2000 (New York, 1995), Lecture Notes in Statist., vol. 128, Springer, New York, 1998, 123–136.

[7] Chung, F. R. K. (1996) Laplacians of graphs and Cheeger's inequalities. In *Combinatorics, Paul ErdHos is eighty*, Vol. **2** , 157–172, J. Bolyai Soc. Math. Stud., Budapest.

[8] Chung, F. R. K. and Yau, S. T. (1995) Eigenvalues of graphs and Sobolev inequalities, *Combinatorics, Probability and Computing* **4**, 11–26.

[9] Diaconis, P. and Fill, J. A. (1990) Strong stationary times via a new form of duality. *Ann. Probab.* **18**, 1483-1522.

[10] Coulhon, T. (1996). Ultracontractivity and Nash type inequalities. *J. Funct. Anal.* **141**, 510–539.

[11] Coulhon, T., Grigorýan, A. and Pittet, C. (2001). A geometric approach to on-diagonal heat kernel lower bounds on groups. *Ann. Inst. Fourier (Grenoble)* **51**, 1763–1827.

[12] P. Diaconis and A. Ram, Analysis of systematic scan Metropolis Algorithms Using Iwahori-Hecke Algebra Techniques, *Michigan Math. J.* **48** (2000), 157–190.

[13] Diaconis, P. and Saloff-Coste, L. (1996). Nash inequalities for finite Markov chains. *J. Theoret. Probab.* **9**, 459–510.

[14] Fill, J. A. (1991). Eigenvalue bounds on convergence to stationarity for nonreversible Markov chains, with an application to the exclusion process. *Ann. Appl. Probab.* **1**, 62–87.

[15] Häggström, O. and Jonasson, J. (1997). Rates of convergence for lamplighter processes. *Stochastic Process. Appl.* **67**, 227–249.

[16] Houdré, C. and Tetali, P. (2004). Isoperimetric Invariants for Product Markov Chains and Graph Products. *Combinatorica* **24**, 359–388.

[17] Jerrum, M. R. and Sinclair, A. J. (1989). Approximating the permanent. *SIAM Journal on Computing* **18**, 1149–1178.

[18] Kannan, R. (2002). Rapid Mixing in Markov Chains *Proceedings of International Congress of Math.* 2002, Vol. III, 673–683.

[19] Lawler, G. and Sokal, A. (1988). Bounds on the L^2 spectrum for Markov chains and Markov processes: a generalization of Cheeger's inequality. *Trans. Amer. Math. Soc.* **309**, 557–580.

[20] Lovász, L. and R. Kannan, R. (1999). Faster mixing via average conductance *Proceedings of the 27th Annual ACM Symposium on theory of computing.*

[21] Lovász, L. and Winkler, P. (1993). A note on the last new vertex visited by a random walk. J. Graph Theory 17 , no. 5, 593–596.

[22] Mathieu, P. and Remy, E. (2004). Isoperimetry and heat kernel decay on percolation clusters. *Ann. Probab.* **32**, 100–128.

[23] Mihail, M. (1989). Conductance and convergence of Markov chains - A combinatorial treatment of expanders. *Proceedings of the 30th Annual Conference on Foundations of Computer Science*, 526–531.

[24] Montenegro, R. and Son, J.-B. (2001) Edge Isoperimetry and Rapid Mixing on Matroids and Geometric Markov Chains, *Proceedings of the 33rd Annual ACM Symposium on theory of computing.*

[25] Morris, B. (2002). A new, probabilistic approach to heat kernel bounds. Lecture at Sectional AMS meeting, Atlanta, GA, March 2002.

[26] Morris, B. and Peres, Y. (2005). Evolving sets, mixing and heat kernel bounds. To appear PTRF, available at `http://front.math.ucdavis.edu/math.PR/0305349`

[27] Pittet, C. and Saloff-Coste, L. (2002) A survey on the relationships between volume growth, isoperimetry, and the behavior of simple random walk on Cayley graphs, with examples. Unpublished manuscript, available at `http://www.math.cornell.edu/~lsc/lau.html`

[28] Quastel, Jeremy (1992). Diffusion of color in the simple exclusion process. Comm. Pure Appl. Math. 45, no. 6, 623–679.

[29] Saloff-Coste, L. (1997). *Lectures on finite Markov chains.* Lecture Notes in Math. **1665**, Springer, Berlin, 301–413.

[30] Sinclair, A. (1993). *Algorithms for Random Generation and Counting: A Markov Chain Approach*, Birkhäuser, Boston.

[31] Varopoulos, N. Th. (1985) Isoperimetric inequalities and Markov chains. *J. Funct. Anal.* **63**, 215–239.

[32] Woess, W. (2000). *Random walks on infinite graphs and groups.* Cambridge Tracts in Mathematics 138, Cambridge University Press.