Mixing times of Markov chains

Perla Sousi*

January 26, 2020

Contents

*University of Cambridge

1 Mixing times

1.1 Background

These notes are largely based on [1].

Definition 1.1. A sequence of random variables $(X_n)_{n\geq 0}$ taking values in a space E is called a Markov chain if for all $x_0, \ldots, x_n \in E$ such that $\mathbb{P}(X_0 = x_0, \ldots, X_{n-1} = x_{n-1}) > 0$ we have

$$\mathbb{P}(X_n = x_n \mid X_0 = x_0, \ldots, X_{n-1} = x_{n-1}) = \mathbb{P}(X_n = x_n \mid X_{n-1} = x_{n-1}).$$

In other words, the future of the process is independent of the past given the present.

For an event A we write $\mathbb{P}_i(A)$ to denote $\mathbb{P}(A \mid X_0 = i)$.

A Markov chain is defined by its transition matrix P given by

$$P(i,j) = \mathbb{P}(X_1 = j \mid X_0 = i) \quad \forall\, i, j \in E.$$

We will also write $p_{i,j}(n)$ or $p_n(i,j)$ for $P^n(i,j)$.

Definition 1.2. A Markov chain is called irreducible if for all $x, y \in E$ there exists $n \geq 0$ such that $P^n(x,y) > 0$.

An irreducible Markov chain is called recurrent if for all i we have $\mathbb{P}_i(T_i < \infty) = 1$, where $T_i = \inf\{n \geq 1 : X_n = i\}$. Otherwise, it is called transient.

A Markov chain is called aperiodic, if for all x we have g.c.d.$\{n \geq 1 : P^n(x,x) > 0\} = 1$.

Let E be a countable (infinite or finite) state space and let π be a probability distribution on E. We call π an invariant distribution if $\pi P = P$. This means that if $X_0 \sim \pi$, then $X_n \sim \pi$ for all n.

Let π be the invariant distribution and suppose that $X_0 \sim \pi$. Fix N and consider the chain $Y_i = X_{N-i}$ for all $i \in \{0, \ldots, N\}$. Then Y is also a Markov chain with transition matrix given by

$$P^*(x,y) = \frac{\pi(y)}{\pi(x)} P(y,x) \quad \text{for all} \quad x, y.$$

We call P^* the reversal of P. It is easy to check that P^* is the adjoint operator, in the sense that for all $f, g : E \to \mathbb{R}$ we have

$$\langle Pf, g\rangle_\pi = \langle f, P^* g\rangle_\pi,$$

where $\langle f, g\rangle_\pi = \mathbb{E}_\pi[fg] = \sum_x \pi(x) f(x) g(x)$.

A Markov chain X is called reversible if for all N when $X_0 \sim \pi$, then $(X_0, \ldots, X_N)$ has the same distribution as $(X_N, \ldots, X_0)$. This is equivalent to the detailed balance equations, i.e. that

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

Let $G = (V, E)$ be a connected graph, which may be infinite or finite. A simple random walk on G is a Markov chain evolving on the vertices V with transition matrix given by

$$P(i,j) = \frac{1}{\deg(i)},$$

for i and j neighbours, i.e. joined by an edge, where $\deg(i)$ is equal to the total number of neighbours of i. If G is finite, then random walk on G is reversible and the invariant distribution is given by

$$\pi(x) = \frac{\deg(x)}{2|E|} \quad \text{for all} \quad x \in V.$$

1.2 Total variation distance and coupling

Recall the convergence to equilibrium theorem for Markov chains.

Theorem 1.3. *Suppose that X is an irreducible and aperiodic Markov chain on a finite state space with invariant distribution π. Then for all x, y we have*

$$P^t(x, y) \to \pi(y) \text{ as } t \to \infty.$$

The theorem above does not tell us anything about the rate of the convergence to equilibrium. Also we need to define a metric between probability measures in order to be able to measure distance between P^t and π. The most widely used metric is the total variation distance.

Definition 1.4. Let E be a finite space and let μ and ν be two probability distributions on E. We define

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

This is a probabilistic definition, as the distance between μ and ν is given in terms of the probabilities assigned to events A.

Proposition 1.5. *Let μ and ν be two probability distributions on Ω. Then*

$$\|\mu - \nu\|_{\mathrm{TV}} = \frac{1}{2} \sum_x |\mu(x) - \nu(x)| = \sum_{x:\mu(x) \geq \nu(x)} (\mu(x) - \nu(x)).$$

Proof. Let $B = \{x : \mu(x) \geq \nu(x)\}$ and let $A \subseteq \Omega$ be an arbitrary set. Then we have

$$\mu(A) - \nu(A) = (\mu(A \cap B) - \nu(A \cap B)) + (\mu(A \cap B^c) - \nu(A \cap B^c)) \leq \mu(A \cap B) - \nu(A \cap B),$$

because $\mu(A \cap B^c) - \nu(A \cap B^c) \leq 0$ by the definition of B. We further have

$$\mu(A \cap B) - \nu(A \cap B) = (\mu(B) - \nu(B)) - (\mu(A \cap B^c) - \nu(A \cap B^c)) \leq \mu(B) - \nu(B),$$

again using the definition of B. Similarly we can get

$$\nu(A) - \mu(A) \leq \nu(B^c) - \mu(B^c) = \mu(B) - \nu(B).$$

This proves that

$$\max_{A \subseteq \Omega} |\mu(A) - \nu(A)| \leq \mu(B) - \nu(B)$$

and taking $A = B$ we get

$$\max_{A \subseteq \Omega} |\mu(A) - \nu(A)| = \mu(B) - \nu(B) = \sum_{x:\mu(x) \geq \nu(x)} (\mu(x) - \nu(x)).$$

Using that $\nu(B^c) - \mu(B^c) = \mu(B) - \nu(B)$ finally gives

$$\|\mu - \nu\|_{\mathrm{TV}} = \frac{1}{2} \sum_X |\mu(x) - \nu(x)|$$

and completes the proof. □

Remark 1.6. The theorem above shows that the total variation distance satisfies the triangle inequality.

Definition 1.7. A coupling of two probability distributions μ and ν is a pair of random variables X and Y defined on the same probability space such that the marginal distribution of X is μ and that of Y is ν.

Example 1.8. *Suppose that $\mu = \nu$. Then one coupling of μ and ν is to take X and Y be independent random variables with distribution μ. Another coupling is to take $X = Y$.*

Proposition 1.9. *Let μ and ν be two probability distributions on Ω. Then*

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

and there is a coupling achieving the infimum above. We will call this coupling the optimal coupling of μ and ν.

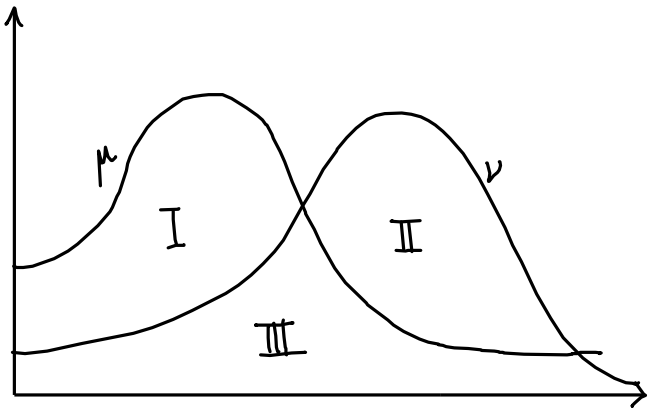
Proof. Let (X, Y) be a coupling of μ and ν. Then for any event A we have

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

This shows that

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

In order to prove the equality, we will construct a coupling for which $\mathbb{P}(X \neq Y)$ is exactly equal to the total variation distance. We want to construct their coupling in such a way so that they are equal as often as possible. We do it as follows: imagine we throw a point in the regions $I \cup II \cup III$. If the point lands in I, then we set $X = Y$. Otherwise, we throw X in the region $I \setminus III$ and Y in $II \setminus III$. In this way, they are equal only if the initial point lands in III.



More formally, let $p = \sum_x \mu(x) \wedge \nu(x)$, where we write $a \wedge b = \min(a, b)$. We toss a coin with probability of heads equal to p. If the coin comes up heads, then we sample Z according to the distribution

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

and we set $X = Y = Z$. If the coin comes up tails, then we sample X according to

$$\gamma_I(x) = \frac{\mu(x) - \nu(x)}{1-p} \mathbf{1}(\mu(x) > \nu(x))$$

and we sample Y according to

$$\gamma_{II}(x) = \frac{\nu(x) - \mu(x)}{1-p} \mathbf{1}(\nu(x) > \mu(x)).$$

First it is easy to check that X and Y have the correct distributions. Indeed,

$$\mathbb{P}(X = x) = p \cdot \frac{\mu(x) \wedge \nu(x)}{p} + (1-p) \cdot \frac{\mu(x) - \nu(x)}{1-p} \mathbf{1}(\mu(x) > \nu(x)) = \mu(x)$$
$$\mathbb{P}(Y = x) = p \cdot \frac{\mu(x) \wedge \nu(x)}{p} + (1-p) \cdot \frac{\nu(x) - \mu(x)}{1-p} \mathbf{1}(\nu(x) > \mu(x)) = \nu(x).$$

Also from the construction, since γ_I and γ_{II} are supported on disjoint sets, it follows that $X \neq Y$ only if the coin comes up tails. Thus we get

$$\mathbb{P}(X \neq Y) = 1 - p = 1 - \sum_x \mu(x) \wedge \nu(x) = \sum_{x:\mu(x) \geq \nu(x)} (\mu(x) - \nu(x)) = \|\mu - \nu\|_{\mathrm{TV}}$$

and this completes the proof. □

1.3 Distance to stationarity

Let X be a Markov chain with transition matrix P and invariant distribution π. We define the distance to stationarity

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

We also define

$$\overline{d}(t) = \max_{x,y} \left\| P^t(x, \cdot) - P^t(y, \cdot) \right\|_{\mathrm{TV}}.$$

Lemma 1.10. *For all t we have the following*

$$d(t) \leq \overline{d}(t) \leq 2d(t).$$

Proof. The second inequality is immediately from the triangle inequality for the total variation distance. To prove the first inequality, we use that for any set A, by stationarity $\pi(A) =$

$\sum_y \pi(y)P^t(y,A)$. Therefore, we get for all x

$$\left\|P^t(x,\cdot)-\pi\right\|_{\mathrm{TV}} = \max_A |P^t(x,A)-\pi(A)| = \max_A \left|\sum_y \pi(y)(P^t(x,A)-P^t(y,A))\right|$$
$$\leq \sum_y \pi(y) \max_A \left|P^t(x,A)-P^t(y,A)\right| \leq \overline{d}(t),$$

since $\sum_y \pi(y) = 1$. □

Theorem 1.11. *Let X be an irreducible and aperiodic Markov chain on a finite state space with invariant distribution π. Then there exist $\alpha \in (0,1)$ and a positive constant C so that for all t we have*

$$\max_x \left\|P^t(x,\cdot)-\pi\right\|_{\mathrm{TV}} \leq C\alpha^t.$$

Proof. By irreducibility and aperiodicity we get that there exists $r > 0$ so that the matrix P^r has strictly positive entries. Define $\alpha = \min_{x,y}(P^r(x,y)/\pi(y))$. Then $\alpha > 0$ and for all x, y we have

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

Therefore, we can write

$$P^r(x,y) = \alpha\pi(y) + (1-\alpha)Q(x,y).$$

From this it is easy to check that Q is a stochastic matrix. A probabilistic interpretation of the above equality is that with probability α we sample y according to π and with probability $1-\alpha$ we sample according to $Q(x,y)$. So in order to sample $P^{rk}(x,y)$, with probability $(1-\alpha)^k$ we never sampled from π and we sampled from $Q^k(x,\cdot)$ and with the complementary probability, $1-(1-\alpha)^k$ we did, in which case it is distributed according to π. This means that

$$P^{rk} = (1-\alpha)^k Q^k + (1-(1-\alpha)^k)\pi,$$

where we think of the vector π as a matrix whose rows are all equal to π. Using stationarity, we get for all j

$$P^{rk+j} = (1-\alpha)^k Q^k P^j + (1-(1-\alpha)^k)\pi.$$

This now gives

$$\left\|P^{rk+j}-\pi\right\|_{\mathrm{TV}} = (1-\alpha)^k \left\|Q^kP^j-\pi\right\|_{\mathrm{TV}} \leq (1-\alpha)^k$$

and this concludes the proof. □

The theorem above says that the Markov chain run long enough will converge to equilibrium, but it does not give information on the rate of convergence.

Exercise 1.12. *Check that $d(t)$ is a non-increasing function of t.*

We define the mixing time to be the first time the total variation distance from stationarity drops below ε, i.e.

$$t_{\mathrm{mix}}(\varepsilon) = \min\{t : d(t) \leq \varepsilon\}.$$

Lemma 1.13. *The function $\overline{d}$ is submultiplicative, i.e. $\overline{d}(s+t) \le \overline{d}(s)\overline{d}(t)$ for all s, t.*

Proof. Fix x and y and let (X, Y) be the optimal coupling of $P^s(x, \cdot)$ and $P^s(y, \cdot)$, i.e.

$$\|P^s(x,\cdot) - P^s(y,\cdot)\|_{\mathrm{TV}} = \mathbb{P}(X \neq Y)\,. \tag{1.1}$$

By the definition of total variation we have

$$\left\|P^{s+t}(x,\cdot) - P^{s+t}(y,\cdot)\right\|_{\mathrm{TV}} = \frac{1}{2}\sum_z \left|P^{s+t}(x,z) - P^{s+t}(y,z)\right|.$$

But by the Markov property, we also have

$$P^{s+t}(x,z) = \mathbb{E}\big[P^t(X,z)\big] \quad \text{and} \quad P^{s+t}(y,z) = \mathbb{E}\big[P^t(Y,z)\big]\,.$$

Substituting this above gives

$$\begin{aligned}
&\left\|P^{s+t}(x,\cdot) - P^{s+t}(y,\cdot)\right\|_{\mathrm{TV}} = \frac{1}{2}\sum_z \left|\mathbb{E}\big[P^t(X,z)\big] - \mathbb{E}\big[P^t(Y,z)\big]\right| \\
&\le \mathbb{E}\left[\frac{1}{2}\sum_z \left|P^t(X,z) - P^t(Y,z)\right|\right] = \mathbb{E}\left[\mathbf{1}(X \neq Y)\cdot\frac{1}{2}\sum_z \left|P^t(X,z) - P^t(Y,z)\right|\right] \\
&\le \mathbb{E}\big[\mathbf{1}(X \neq Y)\cdot \overline{d}(t)\big] = \mathbb{P}(X \neq Y)\,\overline{d}(t).
\end{aligned}$$

Maximising over x and y and using (1.1) completes the proof. □

Exercise 1.14. *By slightly modifying the proof above, show that*

$$d(s+t) \le d(s)\overline{d}(t).$$

In the definition of mixing time we usually take $\varepsilon = 1/4$ and in this case we write $t_{\mathrm{mix}} = t_{\mathrm{mix}}(1/4)$. The choice of $1/4$ is rather arbitrary – any number strictly smaller than $1/2$ would do. Indeed, by the submultiplicativity of $\overline{d}$ and Lemma 1.10 we have

$$d(\ell t_{\mathrm{mix}}(\varepsilon)) \le \overline{d}(\ell t_{\mathrm{mix}}(\varepsilon)) \le \overline{d}(t_{\mathrm{mix}}(\varepsilon))^\ell \le (2\varepsilon)^\ell.$$

So taking $\varepsilon = 1/4$ gives

$$d(\ell t_{\mathrm{mix}}) \le 2^{-\ell} \text{ and } t_{\mathrm{mix}}(\varepsilon) \le \left\lceil \log_2 \frac{1}{\varepsilon} \right\rceil t_{\mathrm{mix}}.$$

Remark 1.15. By Exercise 1.14 one can actually slightly improve and get

$$d(\ell t_{\mathrm{mix}}) \le \varepsilon(2\varepsilon)^{\ell-1}.$$

2 Markovian coupling and other metrics

2.1 Coupling

Definition 2.1. A coupling of Markov chains with transition matrix P is a process $(X_t, Y_t)_t$ so that both X and Y are Markov chains with transition matrix P and with possibly different starting distributions.

A Markovian coupling of P is a coupling of Markov chains which is itself a Markov chain which also satisfies that for all x, x', y, y'

$$\mathbb{P}\big(X_1 = x' \mid X_0 = x, Y_0 = y\big) = P(x, x') \quad \text{and} \quad \mathbb{P}\big(Y_1 = y' \mid X_0 = x, Y_0 = y\big) = P(y, y').$$

A coupling is called coalescent, if whenever there exists s such that $X_s = Y_s$, then $X_t = Y_t$ for all $t \geq s$.

Remark 2.2. Any Markovian coupling can be modified so that it becomes coalescent. Simply run the Markov chains using their Markovian coupling until they meet for the first time and then continue them together.

All couplings used in this course will be Markovian.

Theorem 2.3. *Let (X, Y) be a Markovian coalescent coupling with $X_0 = x$ and $Y_0 = y$. Let*

$$\tau_{\text{couple}} = \inf\{t \geq 0 : X_t = Y_t\}.$$

Then

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

Moreover, if for each pair of states (x, y), there is a Markovian coalescent coupling with τ_{couple} the coupling time, then

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

Proof. Since the coupling is Markovian we have

$$P^t(x, x') = \mathbb{P}_{x,y}\big(X_t = x'\big) \quad \text{and} \quad P^t(y, y') = \mathbb{P}_{x,y}\big(Y_t = y'\big)\,.$$

So (X_t, Y_t) is a coupling of $P^t(x, \cdot)$ and $P^t(y, \cdot)$, and hence we get

$$\big\|P^t(x, \cdot) - P^t(y, \cdot)\big\|_{\text{TV}} \leq \mathbb{P}_{x,y}(X_t \neq Y_t) \leq \mathbb{P}_{x,y}(\tau_{\text{couple}} > t)\,.$$

The final assertion of the theorem follows now from Lemma 1.10. □

We now look at some examples in order to illustrate the use of coupling as a means of upper bounding mixing times.

Notation For functions f, g we will write $f(n) \lesssim g(n)$ if there exists a constant $c > 0$ such that $f(n) \leq cg(n)$ for all n. We write $f(n) \gtrsim g(n)$ if $g(n) \lesssim f(n)$. Finally, we write $f(n) \asymp g(n)$ if both $f(n) \lesssim g(n)$ and $f(n) \gtrsim g(n)$.

The lazy version of a Markov chain with transition matrix P is the Markov chain with transition matrix $(P+I)/2$, i.e. at every state it stays in place with probability $1/2$ and with probability $1/2$ it jumps according to P.

Random walk on $\mathbb{Z}_n$. Consider the integers mod n, i.e. $\mathbb{Z}_n = \{0, \dots, n-1\}$. A simple random walk on $\mathbb{Z}_n$ is a Markov chain with transition probabilities $P(i, (i \pm 1) \bmod n) = 1/2$. In order to get rid of the periodicity issue, we consider the lazy version of this chain, i.e. the Markov chain with transition matrix $(P+I)/2$.

Claim 2.1. *The mixing time of lazy simple random walk on $\mathbb{Z}_n$ satisfies $t_{\text{mix}} \asymp n^2$.*

Proof. We start with the upper bound. We are going to construct a coupling of two lazy walks X and Y starting from x and y respectively. At each step we toss a fair coin independently of previous tosses. If the coin comes up heads, then X makes a move (left or right with equal probability), otherwise Y moves. If at some point they are in the same location, then they continue moving together. Therefore, the coupling time is the time it takes for X and Y to meet under the dynamics we defined. The clockwise distance between the two walks evolves as a simple symmetric random walk on $\mathbb{Z}$ started from $x-y$ and run until it hits 0 or n for the first time (and then gets absorbed there). By gambler's ruin we get

$$\mathbb{E}_{x,y}[\tau_{\text{couple}}] = k(n-k),$$

where k is the clockwise distance between x and y. So using Theorem 2.3 we obtain

$$d(t) \le \max_{x,y \in \mathbb{Z}_n} \mathbb{P}_{x,y}(\tau_{\text{couple}} > t) \le \frac{\mathbb{E}_{x,y}[\tau_{\text{couple}}]}{t} \le \frac{n^2}{4t},$$

where we used Markov's inequality for the second inequality. Taking $t = n^2$ gives $d(n^2) \le 1/4$, and hence this proves the upper bound on t_{mix}.

For the lower bound, let S be a lazy simple random walk on $\mathbb{Z}$, i.e. with probability $1/4$ it jumps to the right, with $1/4$ to the left and with $1/2$ it stays in place. Then we can write $X_t = S_t \bmod n$ and by Chebyshev's inequality

$$\mathbb{P}_0(X_t \in \{\lceil n/4 \rceil + 1, \dots, \lceil 3n/4 \rceil\}) \le \mathbb{P}_0(|S_t| > n/4) \le 16 \frac{\text{Var}\,(S_t)}{n^2} = \frac{8t}{n^2}.$$

Taking now $t = n^2/32$ gives $\mathbb{P}_0(X_t \in A) \le 1/4$, where $A = \{\lceil n/4 \rceil + 1, \dots, \lceil 3n/4 \rceil\}$. But $\pi(A) > 1/2$, and hence we deduce

$$d(t) \ge \pi(A) - \mathbb{P}_0(X_t \in A) \ge 1/4,$$

which shows that $t_{\text{mix}} \ge n^2/32$ and completes the proof of the lower bound. □

2.1.1 Random walk on the binary tree

Consider a finite rooted tree on n vertices with the property that the root has degree two and every other vertex has degree 3. We are interested in the mixing time of a lazy simple random walk on this tree. In order to find an upper bound, we will construct a coupling of two chains X and Y started from two different vertices x and y respectively. Until the first time that the two walks are in the same level, at each step we toss a fair coin. If it comes up Heads, then we X jumps to a

neighbour chosen uniformly at random and Y stays in place. If Tails, then we do the corresponding thing for Y. The first time they reach the same level, we move them up or down together. Then if we wait for the first time they have visited the root after having visited the leaves, they must have coupled. By reducing to a biased random walk on the segment, if τ is the first time they couple, then $\mathbb{E}_{x,y}[\tau] \leq Cn$ for a positive constant C. Therefore, we obtain

$$\mathbb{P}_{x,y}(\tau > t) \leq \frac{\mathbb{E}_{x,y}[\tau]}{t} \leq \frac{Cn}{t},$$

and hence taking $t = 4Cn$ shows that $t_{\text{mix}} \leq 4Cn$. In order to obtain a lower bound, we use Exercise 7 and let A be the right half of the tree. Starting from a leaf on the left side of the tree, the expected time to hit A is of order n. Therefore, $t_{\text{mix}} \gtrsim n$, thus $t_{\text{mix}} \asymp n$.

2.2 Strong stationary times

We start with an example, the top to random shuffle. Consider a deck of n cards and suppose we shuffle it with the following method: at each time step we pick the top card and insert it in a random location. This is a Markov chain taking values in the space of permutations of n elements S_n.

Proposition 2.4. *Let X be the Markov chain corresponding to the order of the cards in the top to random shuffle and let τ_{top} be one step after the first time that the original bottom card arrives at the top of the deck. Then at this time the order of the cards is uniform in S_n and the time τ_{top} is independent of $X_{\tau_{\text{top}}}$.*

Proof. We first prove by induction on the number of steps that the set of cards under the original bottom card is in a uniform order. Indeed, at time $t = 0$, the claim trivially holds. Now suppose that it holds at time t. We show it also holds at time $t + 1$. There are two possibilities. Either a card is placed under the original bottom card or not. In the second case, the order remains uniform by the induction hypothesis. In the first case, the order is again uniform, since the new card was inserted in a random location.

The claim we just proved shows that at time τ_{top} the order of the cards under the original bottom card is uniform, and hence $X_{\tau_{\text{top}}}$ is in a uniform order and independent of τ_{top}. □

For the Markov chain X we have found a random time τ with the property that τ is independent of X_τ and X_τ has the desired distribution, uniform over S_n in this case. We now show how to use the expectation of such a time in order to bound the mixing time of a chain.

Definition 2.5. A stopping time is a random variable T with the property that $\{T \leq t\}$ is completely determined by $X_0, \ldots, X_t$ for all t and more generally by the filtration $\mathcal{F}_t$ to which X is adapted.

Let X be a Markov chain with stationary distribution π. A stopping time τ is called a **stationary time** (possibly depending on the starting point) if for all y we have $\mathbb{P}_x(X_\tau = y) = \pi(y)$.

A stationary time τ is called a **strong stationary time** (possibly depending on the starting point) if X_τ is independent of τ, i.e. it satisfies

$$\mathbb{P}_x(X_\tau = y, \tau = t) = \mathbb{P}_x(\tau = t)\,\pi(y) \quad \forall\ y.$$

Proposition 2.6. *If τ is a strong stationary time when $X_0 = x$, then for all t*

$$\left\|P^t(x,\cdot) - \pi\right\|_{\mathrm{TV}} \leq \mathbb{P}_x(\tau > t)\,.$$

Definition 2.7. We define the separation distance

$$s(t) = \max_{x,y}\left(1 - \frac{P^t(x,y)}{\pi(y)}\right).$$

Lemma 2.8. *For all x we have*

$$\left\|P^t(x,\cdot) - \pi\right\|_{\mathrm{TV}} \leq \max_{x}\left(1 - \frac{P^t(x,y)}{\pi(y)}\right) =: s_x(t),$$

and hence $d(t) \leq s(t)$.

Proof. Using the definition of total variation distance we have

$$\left\|P^t(x,\cdot) - \pi\right\|_{\mathrm{TV}} = \sum_{y:P^t(x,y)<\pi(y)} (\pi(y) - P^t(x,y)) = \sum_{y:P^t(x,y)<\pi(y)} \pi(y)\left(1 - \frac{P^t(x,y)}{\pi(y)}\right) \leq s_x(t)$$

and this concludes the proof. □

Proof of Proposition 2.6. Using Lemma 2.8 it suffices to show

$$s_x(t) \leq \mathbb{P}_x(\tau > t)\,.$$

For all x and y we have

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

We now show that $\mathbb{P}_x(X_t = y, \tau \leq t) = \mathbb{P}_x(\tau \leq t)\,\pi(y)$. Indeed, we have

$$\mathbb{P}_x(X_t = y, \tau \leq t) = \sum_{s\leq t}\sum_{z} \mathbb{P}_x(X_t = y, \tau = s, X_s = z) = \sum_{s\leq t}\sum_{z} \mathbb{P}_x(\tau = s, X_s = z)\,P^{t-s}(z,y),$$

where for the last equality we used the strong Markov property at the stopping time τ. Since τ is a strong stationary time, we now have

$$\mathbb{P}_x(X_t = y, \tau \leq t) = \sum_{s\leq t}\sum_{z} \pi(z)P^{t-s}(z,y)\mathbb{P}_x(\tau = s) = \pi(y)\mathbb{P}_x(\tau \leq t)\,,$$

where for the last equality we used the stationarity of π. This concludes the proof. □

Lemma 2.9. *For reversible chains we have*

$$s(2t) \leq 1 - (1 - \bar{d}(t))^2.$$

Proof. Note that by reversibility we have $P^t(x,y)/\pi(y) = P^t(y,x)/\pi(x)$. Therefore, we obtain

$$\frac{P^{2t}(x,y)}{\pi(y)} = \sum_z P^t(x,z)\cdot\frac{P^t(z,y)}{\pi(y)} = \sum_z P^t(x,z)\cdot\frac{P^t(y,z)}{\pi(z)} = \sum_z \frac{P^t(x,z)P^t(y,z)}{\pi(z)^2}\cdot\pi(z)$$
$$\geq \left(\sum_z \sqrt{P^t(x,z)P^t(y,z)}\right)^2 \geq \left(\sum_z P^t(x,z)\wedge P^t(y,z)\right)^2 = \left(1-\left\|P^t(x,\cdot)-P^t(y,\cdot)\right\|_{\mathrm{TV}}\right)^2,$$

where for the inequality we used Cauchy-Schwarz. Rearranging the above and taking the maximum over all x and y proves the lemma. □

2.3 Examples

Start with coupon collector, since it will be used for both examples.

Proposition 2.10. *A company issues n different types of coupons. A collector needs all n types to win a prize. We suppose that each coupon he acquires is equally likely each of the n types. Let τ be the number of coupons he acquires until he obtains a full set. Then $\mathbb{E}[\tau] = n\sum_{k=1}^n 1/k$ and for any $c>0$*

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

Proof. Let τ_i be the number of coupons he acquires in order to get $i+1$ distinct coupons when he starts with i distinct ones. Then τ_i has the geometric distribution with parameter $(n-i)/n$. We can then write

$$\tau = \tau_0 + \tau_1 + \ldots + \tau_{n-1},$$

and hence, taking expectations proves the desired equality. Regarding the second claim, we let A_i be the event that the i-th coupon does not appear in the first $\lceil n\log n + cn\rceil$ coupons drawn. Then

$$\mathbb{P}(\tau > \lceil n\log n + cn\rceil) = \mathbb{P}(\cup_{i=1}^n A_i) \leq \sum_{i=1}^n \mathbb{P}(A_i)\,.$$

Since the probability of not drawing coupon i in a given trial is $1-1/n$ and the trials are independent, we obtain

$$\mathbb{P}(A_i) = \left(1-\frac{1}{n}\right)^{\lceil n\log n+cn\rceil}.$$

This finally gives

$$\mathbb{P}(\tau > \lceil n\log n + cn\rceil) \leq n\left(1-\frac{1}{n}\right)^{\lceil n\log n+cn\rceil} \leq e^{-c}$$

and concludes the proof. □

Random walk on the hypercube

The n-dimensional hypercube is the graph whose vertex set is $\{0,1\}^n$ and two vertices are joined by an edge if they differ in exactly one coordinate. The lazy simple random walk on $\{0,1\}^n$ can be realised by choosing at every step a coordinate at random and refreshing its bit with a uniform one.

Define τ_{refresh} to be the first time that all coordinates have been picked at least once. Then this is a strong stationary time. The time τ_{refresh} has the same distribution as the coupon collector time. Therefore, taking $t = n \log n + cn$ we obtain from Proposition 2.10 that

$$d(t) \leq \mathbb{P}(\tau_{\text{refresh}} > t) \leq e^{-c},$$

and hence by taking c large, we get that $t_{\text{mix}} \leq n \log n + cn$.

Top to random shuffle In Proposition 2.4 we showed that τ_{top} is a strong stationary time for the top to random shuffle. It is not hard to see that τ_{top} has the same distribution as the coupon collector time. Indeed, when there are k cards under the original bottom card, then at the next step the probability that there are $k+1$ cards under it is equal to $(k+1)/n$. Therefore, taking $t = n \log n + cn$ we get

$$d(t) \leq \mathbb{P}(\tau_{\text{top}} > t) \leq e^{-c},$$

and hence we obtain that $t_{\text{mix}}(\varepsilon) \leq n \log n + c(\varepsilon) n$ for all $\varepsilon \in (0,1)$.

We will now establish a lower bound on $t_{\text{mix}}(\varepsilon)$. To do so, let j be an index to be determined later. Suppose we start from the identity permutation. We define A to be the event that the original j bottom cards retain their original relative order. Then $\pi(A) = 1/j!$ and if τ_j is the first time the card original j-th from the bottom makes it to the top of the deck, then similarly to the coupon collector proof we obtain

$$\mathbb{E}[\tau_j] \geq n(\log n - \log j) \quad \text{and} \quad \text{Var}\,(\tau_j) \leq \frac{n^2}{j-1}.$$

It is clear that if $\tau_j \geq t$, then the event A holds. Taking $t = n \log n - cn$, we thus deduce

$$P^t(\text{id}, A) \geq \mathbb{P}(\tau_j \geq t) \geq 1 - \frac{1}{j-1}$$

for $c \geq \log j + 1$ using Chebyshev's inequality. So

$$d(t) \geq P^t(\text{id}, A) - \pi(A) \geq 1 - \frac{2}{j-1}.$$

Taking $j = \lfloor e^{c-1} \rfloor$ provided $n \geq j$ we get

$$d(t) \geq 1 - \frac{2}{e^{c-2} - 1},$$

and hence taking c sufficiently large gives that $t_{\text{mix}}(\varepsilon) \geq n \log n - c(\varepsilon) n$.

Definition 2.11. A sequence of Markov chains X^n is said to exhibit cutoff, if for all $\varepsilon \in (0,1)$

$$\lim_{n \to \infty} \frac{t^n_{\text{mix}}(\varepsilon)}{t^n_{\text{mix}}(1-\varepsilon)} = 1.$$

Equivalently, writing $d_n(t)$ for $d(t)$ defined with respect to X^n, there is a sequence t_n such that for all $\delta > 0$

$$d_n((1-\delta)t_n) \to 1 \quad \text{and} \quad d_n((1+\delta)t_n) \to 0 \text{ as } n \to \infty.$$

2.4 $\mathcal{L}^p$ distance

Instead of the total variation distance (which is equal to 1/2 the $\mathcal{L}^1$ norm) one can consider other distances. We start by defining the $\mathcal{L}^p$ norm for $p \in [1, \infty]$. Let π be a probability distribution and $f : E \to \mathbb{R}$ be a function. Then

$$\|f\|_p = \|f\|_{p,\pi} = \begin{cases} (\sum_x |f(x)|^p \pi(x))^{1/p} & \text{if } 1 \le p < \infty \\ \max_y |f(y)| & \text{if } p = \infty. \end{cases}$$

For functions f, g we define the scalar product $\langle f, g \rangle_\pi = \sum_x f(x)g(x)\pi(x)$. Finally we define $q_t(x, y) = P^t(x, y)/\pi(y)$. When the chain is reversible, then $q_t(x, y) = q_t(y, x)$. We define the $\mathcal{L}^p$ distance via

$$d_p(t) = \max_x \|q_t(x, \cdot) - 1\|_p .$$

Using Jensen it is easy to see that $2d(t) = d_1(t) \le d_2(t) \le d_\infty(t)$.

We define the $\mathcal{L}^p$ mixing time via

$$t^{(p)}_{\text{mix}}(\varepsilon) = \min\{t \ge 0 : d_p(t) \le \varepsilon\}.$$

When $p = \infty$, we call $t^{(\infty)}_{\text{mix}}(\varepsilon)$ the uniform mixing time.

Proposition 2.12. *For reversible Markov chains we have*

$$d_\infty(2t) = (d_2(t))^2 = \max_x \frac{P^{2t}(x, x)}{\pi(x)} - 1.$$

Proof. By reversibility we have

$$\frac{P^{2t}(x, y)}{\pi(y)} - 1 = \sum_z \left(\frac{P^t(x, z)}{\pi(z)} - 1 \right) \left(\frac{P^t(y, z)}{\pi(z)} - 1 \right) \pi(z).$$

Taking $x = y$ proves the second equality of the proposition. Applying Cauchy-Schwarz we now obtain

$$\begin{aligned} \left| \frac{P^{2t}(x, y)}{\pi(y)} - 1 \right| &\le \sqrt{\sum_z \left(\frac{P^t(x, z)}{\pi(z)} - 1 \right)^2 \pi(z) \cdot \sum_z \left(\frac{P^t(y, z)}{\pi(z)} - 1 \right)^2 \pi(z)} \\ &= \sqrt{\left(\frac{P^{2t}(x, x)}{\pi(x)} - 1 \right) \cdot \left(\frac{P^{2t}(y, y)}{\pi(y)} - 1 \right)}. \end{aligned}$$

Taking the maximum over all x and y shows that $d_\infty(2t) \le (d_2(t))^2$ and then taking $x = y$ proves the equality. □

3 Spectral techniques

3.1 Spectral decomposition and relaxation time

In this section we focus on reversible chains with transition matrix P and invariant distribution π. Recall the inner product $\langle \cdot, \cdot \rangle_\pi$ defined to be $\langle f, g \rangle_\pi = \sum_x f(x)g(x)\pi(x)$.

Theorem 3.1. *Let P be reversible with respect to π. The inner product space $(\mathbb{R}^E, \langle \cdot, \cdot \rangle_\pi)$ has an orthonormal basis of real-valued eigenfunctions $(f_j)_{j \le |E|}$ corresponding to real eigenvalues (λ_j) and the eigenfunction f_1 corresponding to $\lambda_1 = 1$ can be taken to be the constant vector $(1, \dots, 1)$. Moreover, the transition matrix P^t can be decomposed as*

$$\frac{P^t(x,y)}{\pi(y)} = 1 + \sum_{j=2}^{|E|} f_j(x) f_j(y) \lambda_j^t.$$

Proof. We consider the matrix $A(x,y) = \sqrt{\pi(x)}P(x,y)/\sqrt{\pi(y)}$ which using reversibility of P is easily seen to be symmetric. Therefore, we can apply the spectral theorem for symmetric matrices and get the existence of an orthonormal basis (g_j) corresponding to real eigenvalues. It is easy to check that $\sqrt{\pi}$ is an eigenfunction of A with eigenvalue 1. Let D be the diagonal matrix with elements $(\sqrt{\pi(x)})$. Then $A = DPD^{-1}$ and it is easy to check that $f_j = D^{-1}g_j$ are eigenfunctions of P and $\langle f_j, f_i \rangle_\pi = \mathbf{1}(i = j)$. So we have $P^t f_j = \lambda_j^t f_j$ and hence

$$P^t(x,y) = (P^t \mathbf{1}_y)(x) = \sum_{j=1}^{|E|} \lambda_j^t f_j(x) \langle f_j, \mathbf{1}_y \rangle_\pi = \sum_{j=1}^{|E|} \lambda_j^t f_j(x) f_j(y) \pi(y).$$

Using that $f_1 = 1$ and $\lambda_1 = 1$ gives the desired decomposition. □

Let P be a reversible matrix with respect to π. We order its eigenvalues

$$1 = \lambda_1 \ge \lambda_2 \ge \dots \lambda_{|E|} \ge -1.$$

We let $\lambda_* = \max\{|\lambda| : \lambda \text{ is an eigenvalue of } P, \lambda \ne 1\}$ and define $\gamma_* = 1 - \lambda_*$ to be the absolute spectral gap. The spectral gap is defined to be $\gamma = 1 - \lambda_2$.

Exercise 3.2. *Check that if the chain is lazy then $\gamma_* = \gamma$.*

Definition 3.3. The relaxation time for a reversible Markov chain is defined to be

$$t_{\text{rel}} = \frac{1}{\gamma_*}.$$

For a probability measure ν we write $\|\nu - \pi\|_2 = \|\nu/\pi - 1\|_{2,\pi}$.

Theorem 3.4 (Poincaré inequality)**.** *Let P be a reversible matrix with respect to the invariant distribution π. Then for all starting distributions ν we have*

$$\|\mathbb{P}_\nu(X_t = \cdot) - \pi\|_2 \le (1 - \gamma_*)^t \|\nu - \pi\|_2 \le e^{-t/t_{\text{rel}}} \|\nu - \pi\|_2 .$$

Proof. We have

$$\|\mathbb{P}_\nu(X_t=\cdot)-\pi\|_2^2=\sum_y\frac{\mathbb{P}_\nu(X_t=y)^2}{\pi(y)}-1=\sum_y\frac{1}{\pi(y)}\left(\sum_x\nu(x)\mathbb{P}_x(X_t=y)\right)^2-1.$$

Using the decomposition of $P^t(x,y)$ from Theorem 3.1 we obtain

$$\sum_y\frac{1}{\pi(y)}\left(\sum_x\nu(x)\mathbb{P}_x(X_t=y)\right)^2=\sum_{x,x'}\nu(x)\nu(x')\sum_{i,j=1}^n\lambda_i^t\lambda_j^t f_i(x)f_i(x')\sum_y f_i(y)f_j(y)\pi(y)$$

By the orthogonality of the eigenfunctions we obtain that this last sum is equal to

$$\begin{aligned}\sum_{x,x'}\nu(x)\nu(x')\sum_{i,j=1}^n\lambda_i^t\lambda_j^t f_i(x)f_i(x')\mathbf{1}(i=j)&=\sum_{x,x'}\nu(x)\nu(x')\sum_{i=1}^n\lambda_i^{2t}f_i(x)f_i(x')\\&=1+\sum_{x,x'}\nu(x)\nu(x')\sum_{i=2}^n\lambda_i^{2t}f_i(x)f_i(x')\le 1+\lambda_*^{2t}\sum_{x,x'}\nu(x)\nu(x')\sum_{i=2}^n f_i(x)f_i(x').\end{aligned}$$

Using that $P^0(\cdot,\cdot)=I$ and Theorem 3.1 we deduce

$$\sum_{i=2}^n f_i(x)f_i(x')=\frac{\mathbf{1}(x=x')}{\pi(x)}-1,$$

and hence this finally gives

$$\begin{aligned}\|\mathbb{P}_\nu(X_t=\cdot)-\pi\|_2^2&\le\lambda_*^{2t}\left(\sum_{x,x'}\nu(x)\nu(x')\frac{\mathbf{1}(x=x')}{\pi(x)}-1\right)=\lambda_*^{2t}\left(\sum_x\frac{\nu(x)^2}{\pi(x)}-1\right)\\&=(1-\gamma_*)^{2t}\|\nu-\pi\|_2^2\le e^{-2t\gamma_*}\|\nu-\pi\|_2^2=e^{-2t/t_{\mathrm{rel}}}\|\nu-\pi\|_2^2\end{aligned}$$

and this concludes the proof. □

If we do not upper bound the eigenvalues by the second one, then the proof above also gives the following lemma.

Lemma 3.5. *Let P be reversible with respect to π and let*

$$1=\lambda_1\ge\lambda_2\ge\ldots\ge\lambda_n\ge-1$$

be its eigenvalues and (f_j) the corresponding orthonormal eigenfunctions. Then for all x we have

$$4\left\|P^t(x,\cdot)-\pi\right\|_{\mathrm{TV}}^2\le\left\|P^t(x,\cdot)-\pi\right\|_2^2=\sum_{j=2}^n f_j(x)^2\lambda_j^{2t}.$$

Definition 3.6. A Markov chain with transition matrix P is **transitive** if for all x,y in the state space there is a bijection $\varphi=\varphi_{(x,y)}$ such that $\varphi(x)=y$ and $P(z,w)=P(\varphi(z),\varphi(w))$ for all z,w.

Lemma 3.7. *Let P be reversible and transitive. Then for all x we have*

$$\left\|P^t(x,\cdot)-\pi\right\|_2^2=\sum_{j=2}^n\lambda_j^{2t}.$$

Proof. First of all it is easy to check that the uniform distribution, i.e. $\pi(x)=1/n$ for all x, is invariant for P. Next recall that for all x we have

$$\left\|P^t(x,\cdot)-\pi\right\|_2^2=\frac{P^{2t}(x,x)}{\pi(x)}-1=nP^{2t}(x,x)-1.$$

By the definition of transitivity, it follows that the right hand side above is independent of x. Therefore, by Lemma 3.5 we get that $\sum_{j=2}^n f_j(x)^2\lambda_j^{2t}$ is independent of x. Taking the sum over all x and using that (f_j) constitutes an orthonormal basis implies that

$$\sum_{j=2}^n f_j(x)^2\lambda_j^{2t}=\sum_{j=2}^n\lambda_j^{2t}$$

and this concludes the proof. □

Theorem 3.8. *Let P be reversible with respect to the invariant distribution π and let $\pi_{\min}=\min_x\pi(x)$. Then for all $\varepsilon\in(0,1)$ we have*

$$t_{\mathrm{mix}}(\varepsilon)\le t_{mix}^{(\infty)}(\varepsilon)\le t_{\mathrm{rel}}\log\left(\frac{1}{\varepsilon\pi_{\min}}\right).$$

Proof. By the monotonicity of the $\mathcal{L}^p$ norms it suffices to prove the second inequality above. Also, using that $2t_{\mathrm{mix}}^{(\infty)}(\varepsilon)\le t_{\mathrm{mix}}^{(2)}(\sqrt{\varepsilon})$, it suffices to prove that

$$t_{\mathrm{mix}}^{(2)}(\sqrt{\varepsilon})\le\frac{1}{2}t_{\mathrm{rel}}\log\left(\frac{1}{\varepsilon\pi_{\min}}\right). \tag{3.1}$$

To this end, fix x in the state space and by the Poincaré inequality (Theorem 3.4) we have

$$\left\|P^t(x,\cdot)-\pi\right\|_2\le e^{-t/t_{\mathrm{rel}}}\left\|\delta_x-\pi\right\|_2=e^{-t/t_{\mathrm{rel}}}\left(\frac{1}{\pi(x)}-1\right)^{1/2}\le e^{-t/t_{\mathrm{rel}}}\frac{1}{\sqrt{\pi(x)}}\le e^{-t/t_{\mathrm{rel}}}\frac{1}{\sqrt{\pi_{\min}}}.$$

Taking $t=t_{\mathrm{rel}}\log(1/(\varepsilon\pi_{\min}))/2$ in the above inequality shows that $t_{\mathrm{mix}}^{(2)}(\sqrt{\varepsilon})\le t$ and thus proves (3.1) and completes the proof of the theorem. □

Theorem 3.9. *Let P be a reversible matrix with respect to π. Let λ be an eigenvalue with $\lambda\neq 1$. Then $2d(t)\ge|\lambda|^t$. Moreover, for all $\varepsilon\in(0,1)$ we have*

$$t_{\mathrm{mix}}(\varepsilon)\ge(t_{\mathrm{rel}}-1)\log\left(\frac{1}{2\varepsilon}\right).$$

Proof. Let φ be the eigenfunction corresponding to the eigenvalue λ. Then, by the orthogonality of the eigenfunctions, f is orthogonal to $f_1=(1,\dots,1)$ corresponding to $\lambda_1=1$. Therefore,

$\mathbb{E}_\pi[f] = \langle f, 1\rangle_\pi = 0$. Using that $P^t f = \lambda^t f$ for all $t \geq 0$ gives

$$|\lambda^t f(x)| = |P^t f(x)| = \left| \sum_y \big(P^t(x,y)f(y) - \pi(y)f(y)\big) \right| \leq \max_y |f(y)| \cdot 2d(t).$$

Taking now x such that $|f(x)| = \max_y |f(y)|$ shows that $|\lambda|^t \leq 2d(t)$, and hence $|\lambda|^{t_{\text{mix}}(\varepsilon)} \leq 2\varepsilon$, which implies that

$$t_{\text{mix}}(\varepsilon) \geq \log\left(\frac{1}{2\varepsilon}\right) \frac{1}{\log(1/|\lambda|)}.$$

Maximising over all eigenvalues $\lambda \neq 1$ and using that $\log x \leq x - 1$ for all $x > 0$ shows that

$$t_{\text{mix}}(\varepsilon) \geq \log\left(\frac{1}{2\varepsilon}\right) \cdot \frac{1}{\log(1/|\lambda_*|)} \geq \log\left(\frac{1}{2\varepsilon}\right) \cdot \frac{1}{\frac{1}{|\lambda_*|} - 1} = \log\left(\frac{1}{2\varepsilon}\right) \cdot \frac{|\lambda_*|}{1 - |\lambda_*|} = \log\left(\frac{1}{2\varepsilon}\right) \cdot (t_{\text{rel}} - 1)$$

and this completes the proof. □

Corollary 3.10. *Let P be reversible with respect to π. Then we have*

$$d(t)^{1/t} \to \lambda_* \quad \text{as} \quad t \to \infty.$$

Proof. Theorem 3.9 gives one direction. For the other one we use again the monotonicity of $\mathcal{L}^p$ norms in p to get

$$d(t) \leq d_2(t) \leq (1 - \gamma_*)^t \cdot \frac{1}{\sqrt{\pi_{\min}}},$$

where the last inequality follows from the Poincaré inequality, Theorem 3.4. □

Recall that a sequence of chains exhibits cutoff if for all $\varepsilon \in (0,1)$

$$\lim_{n\to\infty} \frac{t^n_{\text{mix}}(\varepsilon)}{t^n_{\text{mix}}(1-\varepsilon)} = 1.$$

A sequence of chains satisfies a weaker condition called **pre-cutoff** if

$$\sup_{0<\varepsilon<1/2} \limsup_{n\to\infty} \frac{t^n_{\text{mix}}(\varepsilon)}{t^n_{\text{mix}}(1-\varepsilon)} < \infty.$$

Proposition 3.11. *Let $P^{(n)}$ be a sequence of reversible Markov chains with mixing times $t^{(n)}_{\text{mix}}$ and relaxation times $t^{(n)}_{\text{rel}}$. If $t^{(n)}_{\text{mix}}/t^{(n)}_{\text{rel}}$ is bounded from above, then there is no pre-cutoff.*

Proof. Dividing both sides of the statement of Theorem 3.9 by $t^{(n)}_{\text{mix}}$ we get

$$\frac{t^{(n)}_{\text{mix}}(\varepsilon)}{t^{(n)}_{\text{mix}}} \geq \frac{t^{(n)}_{\text{rel}} - 1}{t^{(n)}_{\text{mix}}} \log\left(\frac{1}{2\varepsilon}\right).$$

Using now that $t^{(n)}_{\text{mix}}/t^{(n)}_{\text{rel}}$ is bounded from above, we obtain that the right hand side above is lower bounded by $c_1 \log(1/(2\varepsilon))$ for a positive constant c_1. Letting $\varepsilon \to 0$ proves the proposition. □

3.2 Examples

Lazy random walk on the cycle $\mathbb{Z}_n$

Consider the lazy simple random walk on $\mathbb{Z}_n$. We want to find the eigenfunctions and the corresponding eigenvalues. Let f be an eigenfunction with eigenvalue λ. Then it must satisfy

$$\frac{f(x)}{2} + \frac{f(x+1)}{4} + \frac{f(x-1)}{4} = \lambda f(x), \tag{3.2}$$

for all $x \in \mathbb{Z}_n$ where addition and subtraction above are taken mod n. Thinking of the points on the cycle are the roots of unity, we set for $k = 0, \ldots, n-1$ and all $x \in \mathbb{Z}_n$

$$f_k(x) = \exp\left(\frac{2\pi k i x}{n}\right).$$

Then it is straightforward to check that for each k, the function f_k satisfies (3.2) with

$$\lambda_{k+1} = \frac{1 + \cos(2\pi k/n)}{2}.$$

Since for each k the function f_k is an eigenfunction of a real matrix corresponding to a real eigenvalue, it follows that both its real and imaginary parts are also eigenfunctions. So let

$$\varphi_k(x) = \cos\left(\frac{2\pi k x}{n}\right).$$

Taking $k = 0$ gives (as expected) $\lambda_1 = 1$ and taking $k = 1$ gives the second eigenvalue, which by laziness also corresponds to the second maximum. So we get

$$\lambda_* = \lambda_2 = \frac{1 + \cos(2\pi/n)}{2} = 1 - \frac{\pi^2}{n^2} + O(n^{-4}).$$

Therefore, this implies that $t_{\text{rel}} \sim n^2/\pi^2$ as $n \to \infty$.

Since $t_{\text{rel}} \asymp t_{\text{mix}}$, it follows that the lazy random walk on the cycle does not exhibit pre-cutoff.

Lazy random walk on the hypercube $\{0,1\}^n$

Recall that the lazy random walk on the hypercube can be realised by every time picking a coordinate at random and refreshing its bit with a uniform $\{0,1\}$ bit. The walk on the hypercube can be thought of as the product of n matrices each corresponding to the Markov chain on $\{0,1\}$ with transition matrix $P(x,y) = 1/2$ for all x, y. The product means that every time we pick one coordinate at random and we use the corresponding matrix to move it to the next value. So we start by considering the case $n = 1$ and the walk on $\{0,1\}$ with transition matrix $P(x,y) = 1/2$ for all x, y. It is straightforward to check that the eigenfunctions of this chain are $g_1(x) = 1$ for all x corresponding to $\lambda = 1$ and $g_2(x) = 1 - 2x$ corresponding to $\lambda = 0$.

For $x \in \{0,1\}^n$ we now set

$$f(x_1, \ldots, x_n) = \prod_{i=1}^{n} f_i(x_i),$$

where f_i is either g_1 or g_2 for each. i. It is straightforward to check that indeed f is an eigenfunction for the lazy random walk on $\{0,1\}^n$. Now for every subset $I \subseteq \{1,\ldots,n\}$ we take

$$f_I(x_1,\ldots,x_n) = \prod_{i\in I} g_2(x_i).$$

Then this is an eigenfunction corresponding to the eigenvalue

$$\lambda_I = \frac{n-|I|}{n}.$$

It is easy to check that if $I \neq J$, then f_I and f_J are orthogonal. Since there are in total 2^n subsets I, we get an orthonormal basis of eigenfunctions. When $I = \varnothing$, this gives the eigenvalue $\lambda_\varnothing = 1$ and for $|I| = 1$ we get $\lambda_* = \lambda_2 = 1 - 1/n$, which implies that $t_{\text{rel}} = n$. Note that $\pi_{\min} = 2^{-n}$, and hence applying Theorem 3.8 gives

$$t_{\text{mix}}(\varepsilon) \leq n\left(\log(1/\varepsilon) + \log(2^n)\right) \lesssim n^2.$$

This is not a good bound, since we have already obtained a better upper bound of order $n\log n$ using strong stationary times. However, using the full spectrum and not just the second eigenvalue we will see now how we can get the correct order as well as the correct constant.

It is clear by symmetry that the lazy random walk on the hypercube is a transitive chain. Therefore, we can use Lemma 3.7 to get for all x

$$\begin{aligned} 4\left\|P^t(x,\cdot)-\pi\right\|_{\text{TV}} \leq \left\|P^t(x,\cdot)-\pi\right\|_2^2 &= \sum_{\varnothing\neq I\subseteq\{1,\ldots,n\}} \lambda_I^{2t} = \sum_{k=1}^n \binom{n}{k}\cdot\left(\frac{n-k}{n}\right)^{2t} \\ &\leq \sum_{k=1}^n \binom{n}{k} e^{-2kt/n} = \left(1+e^{-2t/n}\right)^n - 1. \end{aligned}$$

Taking now $t = n\log n/2 + cn$ gives

$$4\left\|P^t(x,\cdot)-\pi\right\|_{\text{TV}}^2 \leq e^{e^{-2c}} - 1,$$

and hence taking c sufficiently large (independent of n) shows that the right hand side above can be made arbitrarily small, thus showing that for all $\varepsilon \in (0,1)$ we have

$$t_{\text{mix}}(\varepsilon) \leq \frac{1}{2} n\log n + c(\varepsilon)n.$$

We now prove a matching (to the leading order and constant) lower bound.

Suppose we start with $X_0 = (0,\ldots,0)$. For $x = (x_1,\ldots,x_n)$ we define

$$\Phi(x_1,\ldots,x_n) = \sum_{i=1}^n (1-2x_i).$$

Then $\Phi(X_t)$ satisfies

$$\mathbb{E}[\Phi(X_{t+1}) \mid X_0, \dots, X_t] = \left(1 - \frac{1}{n}\right) \Phi(X_t), \tag{3.3}$$

and hence using that $\Phi(X_0) = n$ this immediately gives

$$\mathbb{E}[\Phi(X_t)] = n \left(1 - \frac{1}{n}\right)^t.$$

Now letting $t \to \infty$ shows that $\mathbb{E}_\pi[\Phi(X)] = 0$, which also follows from the fact that each coordinate is equally likely to be either 0 or 1. Since changing each coordinate changes the value of Φ by ± 2, this gives

$$\mathbb{E}\big[(\Phi(X_{t+1}) - \Phi(X_t))^2 \;\big|\; X_0, \dots, X_t\big] = 2.$$

Therefore, combining this with (3.3) and using again that $\Phi(X_0) = n$ imply

$$\mathbb{E}\big[\Phi(X_t)^2\big] = n + n(n-1) \left(1 - \frac{2}{n}\right)^t.$$

Therefore, this gives that

$$\mathrm{Var}\left(\Phi(X_t)\right) = n + n(n-1) \left(1 - \frac{2}{n}\right)^t - n^2 \left(1 - \frac{1}{n}\right)^{2t} \leq n,$$

since $1 - 2/n \leq (1 - 1/n)^2$. Notice that when $X \sim \pi$, then $\mathrm{Var}_\pi\left(\Phi(X)\right) = n$. So we now get

$$d(t) \geq \mathbb{P}_1\left(\Phi(X_t) \geq \frac{1}{2} n \left(1 - \frac{1}{n}\right)^t\right) - \mathbb{P}_\pi\left(\Phi(X) \geq \frac{1}{2} n \left(1 - \frac{1}{n}\right)^t\right) \geq 1 - \frac{8}{n\left(1 - \frac{1}{n}\right)^{2t}},$$

where for the last inequality we used Chebyshev. Taking now $t = 1/2n \log n - cn$ for a suitable constant c shows that the right hand side above can be made arbitrarily close to 1, hence showing that for all $\varepsilon \in (0,1)$ we have

$$t_{\mathrm{mix}}(\varepsilon) \geq \frac{1}{2} n \log n - c(\varepsilon) n$$

and thus concluding the proof of the lower bound.

The previous technique generalises to any eigenfunction and gives lower bounds on mixing.

Theorem 3.12 (Wilson's method)**.** *Let X be an irreducible and aperiodic Markov chain and let Φ be an eigenfunction corresponding to eigenvalue λ with $1/2 < \lambda < 1$. Suppose there exists $R > 0$ such that*

$$\mathbb{E}_x\big[(\Phi(X_1) - \Phi(X_0))^2\big] \leq R \quad \forall\ x.$$

Then for all $\varepsilon \in (0,1)$ and all x we have

$$t_{\mathrm{mix}}(\varepsilon) \geq \frac{1}{2\log(1/\lambda)} \left(\log\left(\frac{(1-\lambda)\Phi(x)^2}{2R}\right) + \log\left(\frac{1-\varepsilon}{\varepsilon}\right)\right).$$

3.3 Hitting time bound

For a Markov chain X and a state x we let

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

be the first hitting time of x. We also define $t_{\text{hit}} = \max_{x,y} \mathbb{E}_x[\tau_y]$.

Theorem 3.13. *Let P be a lazy reversible Markov chain with invariant distribution π. Then*

$$t_{\text{mix}} \leq 4t_{\text{hit}}.$$

First proof. Recall the definition of the separation distance

$$s(t) = \max_{x,y} \left(1 - \frac{P^t(x,y)}{\pi(y)}\right)$$

and the separation mixing time is defined to be $t_{\text{sep}} = \min\{t \geq 0 : s(t) \leq 1/4\}$. We showed in Lemma 2.8 that $d(t) \leq s(t)$ for all t. So it suffices to prove the bound on the separation mixing. We now have

$$\frac{P^t(x,y)}{\pi(y)} \geq \mathbb{P}_x(\tau_y \leq t) \min_s \frac{P^s(y,y)}{\pi(y)}. \tag{3.4}$$

Since the chain was assumed to be lazy, it follows that for all times t and all states x we have $P^t(x,x) \geq \pi(x)$. Therefore, this shows that the right hand side of (3.4) is larger than $\mathbb{P}_x(\tau_y \leq t)$. So this shows that

$$s(t) \leq \max_{x,y} \mathbb{P}_x(\tau_y > t)\,.$$

Taking now $t = 4t_{\text{hit}}$ gives that $s(t) \leq 1/4$ and finishes the proof. □

Second proof. Recall from the example sheet that if P is aperiodic and irreducible, then

$$\pi(x)\mathbb{E}_\pi[\tau_x] = \sum_{t=0}^{\infty} (P^t(x,x) - \pi(x)).$$

Since the chain is lazy and reversible, by the spectral theorem it is easy to see that $P^t(x,x)$ is decreasing in t and converges to $\pi(x)$ as $t \to \infty$. Therefore, we can lower bound the sum above

$$\pi(x)\mathbb{E}_\pi[\tau_x] \geq \sum_{t=0}^{T} (P^t(x,x) - \pi(x)) \geq T(P^T(x,x) - \pi(x)),$$

where in the last inequality we used again the decreasing property. Dividing through by $T\pi(x)$ gives

$$\frac{\mathbb{E}_\pi[\tau_x]}{T} \geq \frac{P^T(x,x)}{\pi(x)} - 1.$$

Recall from Proposition 2.12 that

$$d_\infty(2t) = \max_x \frac{P^{2t}(x,x)}{\pi(x)} - 1.$$

So we obtain

$$d_\infty(2T) = \max_x \left(\frac{P^{2T}(x,x)}{\pi(x)} - 1 \right) \leq \frac{\max_x \mathbb{E}_\pi[\tau_x]}{2T}.$$

Taking now $T = 2\max_x \mathbb{E}_\pi[\tau_x]$, shows that $t_{\text{mix}}^{(1/4)}(\infty) \leq 4\max_x \mathbb{E}_\pi[\tau_x]$ and this concludes the second proof. □

Remark 3.14. Note that the reversibility assumption in Theorem 3.13 is essential. Consider a biased random walk on $\mathbb{Z}_n$, for which $t_{\text{mix}} \asymp n^2$ while $t_{\text{hit}} \asymp n$.

4 Dirichlet form and the bottleneck ratio

Recall the definition of the inner product: for $f, g : E \to \mathbb{R}$ be two functions we define

$$\langle f, g \rangle_\pi = \sum_x f(x)g(x)\pi(x).$$

Definition 4.1. Let P be a transition matrix with invariant distribution π. The Dirichlet form associated to P and π is defined for all $f, g : E \to \mathbb{R}$

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

Expanding in the definition of $\mathcal{E}$ we get

$$\mathcal{E}(f,g) = \sum_x (I-P)f(x)g(x)\pi(x) = \sum_{x,y} (f(x) - f(y))g(x)P(x,y)\pi(x).$$

When P is reversible with respect to π, then the right hand side above is also equal to

$$\mathcal{E}(f,g) = \sum_{x,y} (f(y) - f(x))g(y)P(x,y)\pi(x).$$

Therefore, in the reversible case we get

$$\mathcal{E}(f,g) = \frac{1}{2}\sum_{x,y} (f(x) - f(y))(g(x) - g(y))\pi(x)P(x,y).$$

When $f = g$ we simply write $\mathcal{E}(f) = \mathcal{E}(f,f)$.

Corollary 4.2. *Let P be a reversible matrix with respect to π. Then for all $f : E \to \mathbb{R}$ we have*

$$\mathcal{E}(f) = \frac{1}{2}\sum_{x,y} (f(x) - f(y))^2 \pi(x)P(x,y).$$

Theorem 4.3. *Let P be a reversible matrix with respect to π. Then the spectral gap $\gamma = 1 - \lambda_2$ satisfies*

$$\gamma = \min_{f:\ \|f\|_2=1, \mathbb{E}_\pi[f]=0} \mathcal{E}(f) = \min_{\substack{f:\ f\not\equiv 0 \\ \mathbb{E}_\pi[f]=0}} \frac{\mathcal{E}(f)}{\|f\|_2^2} = \min_{f:\ \mathrm{Var}_\pi(f)\neq 0} \frac{\mathcal{E}(f)}{\mathrm{Var}_\pi(f)}.$$

Proof. Using that $\mathcal{E}(f+c) = \mathcal{E}(f)$ for any constant $c \in \mathbb{R}$ and $\|f - \mathbb{E}_\pi[f]\|_2^2 = \mathrm{Var}_\pi(f)$ gives the third equality. Also taking $\widetilde{f} = f(x)/\|f\|_2$ gives the second one. So we now prove the first equality.

Let (f_j) be an orthonormal basis for the space $(\mathbb{R}^E, \langle\cdot,\cdot\rangle_\pi)$. Then any function f with $\mathbb{E}_\pi[f] = 0$ can be expressed as

$$f = \sum_{j=2}^{n} \langle f, f_j\rangle_\pi f_j,$$

and hence the Dirichlet form is equal to

$$\mathcal{E}(f) = \langle (I-P)f, f\rangle_\pi = \sum_{j=2}^{n} (1-\lambda_j)\langle f, f_j\rangle_\pi^2 \geq (1-\lambda_2)\sum_{j=2}^{n}\langle f, f_j\rangle_\pi^2.$$

Taking f with $\|f\|_2 = 1$, gives that the last sum appearing above is equal to 1, and hence proves that

$$\min_{f:\ \|f\|_2=1, \mathbb{E}_\pi[f]=0} \mathcal{E}(f) \geq 1 - \lambda_2.$$

Finally, taking $f = f_2$ we get $\mathcal{E}(f_2) = 1 - \lambda_2$ and this concludes the proof. □

Lemma 4.4. *Let P and $\widetilde{P}$ be two transition matrices reversible with respect to π and $\widetilde{\pi}$ respectively. Suppose that there exists a positive A such that $\widetilde{\mathcal{E}}(f) \leq A\mathcal{E}(f)$ for all functions $f : E \to \mathbb{R}$. Let γ and $\widetilde{\gamma}$ be the spectral gaps of P and $\widetilde{P}$ respectively. Then they satisfy*

$$\widetilde{\gamma} \leq \left(\max_x \frac{\pi(x)}{\widetilde{\pi}(x)}\right) A\gamma.$$

Proof. From Theorem 4.3 and the assumption we have

$$\widetilde{\gamma} = \min_{f \text{ not constant}} \frac{\widetilde{\mathcal{E}}(f)}{\mathrm{Var}_{\widetilde{\pi}}(f)} \leq A \cdot \min_{f \text{ not constant}} \frac{\mathcal{E}(f)}{\mathrm{Var}_{\widetilde{\pi}}(f)}. \tag{4.1}$$

Since the variance of a random variable X is the minimum of $E(X-a)^2$ over all $a \in \mathbb{R}$, it follows that

$$\begin{aligned}\mathrm{Var}_\pi(f) &= \mathbb{E}_\pi\left[(f - \mathbb{E}_\pi[f])^2\right] \leq \mathbb{E}_\pi\left[(f - \mathbb{E}_{\widetilde{\pi}}[f])^2\right] = \sum_x \pi(x)(f(x) - \mathbb{E}_{\widetilde{\pi}}[f])^2 \\ &= \sum_x \frac{\pi(x)}{\widetilde{\pi}(x)}\widetilde{\pi}(x)(f(x) - \mathbb{E}_{\widetilde{\pi}}[f])^2 \leq \left(\max_x \frac{\pi(x)}{\widetilde{\pi}(x)}\right)\cdot \mathrm{Var}_{\widetilde{\pi}}(f).\end{aligned}$$

Substituting this into (4.1) finishes the proof. □

4.1 Canonical paths

Suppose that for each x and y in the state space we choose a "path" $\Gamma_{xy} = (x_0, x_1, \ldots, x_k)$ with $x_0 = x$ and $x_k = y$ with the property that $P(x_i, x_{i+1}) > 0$ for all $i \leq k-1$. We write $|\Gamma_{xy}| = k$ for the length of the path, i.e. the number of edges used. We call $e = (x, y)$ an edge if $P(x, y) > 0$ and we write $Q(e) = \pi(x)P(x, y)$. We also let $E = \{(x, y) : P(x, y) > 0\}$.

Theorem 4.5. *Let P be a reversible transition matrix with invariant distribution π. Define the congestion ratio*

$$B = \max_{e \in E} \left(\frac{1}{Q(e)} \sum_{x,y: e \in \Gamma_{xy}} |\Gamma_{xy}| \pi(x)\pi(y) \right),$$

where $e \in \Gamma_{xy}$ means there exists i such that $e = (x_i, x_{i+1})$ with x_i, x_{i+1} consecutive vertices on Γ_{xy}. Then the spectral gap γ satisfies $\gamma \geq 1/B$.

Proof. For an edge $e = (x, y)$ we write $\nabla f(e) = f(x) - f(y)$. Let X and Y be independent and both distributed according to π. Then

$$\begin{aligned}
\mathrm{Var}_\pi(f) &= \frac{\mathbb{E}\left[(f(X) - f(Y))^2\right]}{2} = \frac{1}{2} \sum_{x,y} (f(x) - f(y))^2 \pi(x)\pi(y) \\
&= \frac{1}{2} \sum_{x,y} \left(\sum_{e \in \Gamma_{xy}} \nabla f(e) \right)^2 \pi(x)\pi(y) \leq \frac{1}{2} \sum_{x,y} |\Gamma_{xy}| \sum_{e \in \Gamma_{xy}} (\nabla f(e))^2 \pi(x)\pi(y) \\
&= \frac{1}{2} \sum_e Q(e)(\nabla f(e))^2 \cdot \frac{1}{Q(e)} \sum_{x,y: e \in \Gamma_{xy}} |\Gamma_{xy}| \pi(x)\pi(y) \leq \mathcal{E}(f) B,
\end{aligned}$$

where for the inequality we used Cauchy-Schwartz. Using Theorem 4.3 completes the proof. □

Claim 4.1. *Let X be a lazy simple random walk on the box $[1, n]^d \cap \mathbb{Z}^d$ with reflection at the boundary. There exists $c > 0$ such that $t_{\mathrm{rel}} \leq c(dn)^2$.*

Proof. We describe the choice of path Γ_{xy} in two dimensions. For each x and y we take Γ_{xy} to be the path that goes first horizontally and then vertically. Then for a given edge e the number of x, y with the property that $e \in \Gamma_{xy}$ is at most n^{d+1}. Also the invariant distribution satisfies $\pi(x) \leq c/n^d$ and $Q(e) \asymp (dn^d)^{-1}$. We bound the quantity B from Theorem 4.5 by

$$B \lesssim n^{1-d} d^2 n^{d+1} = d^2 n^2$$

and this concludes the proof. □

4.2 Comparison technique

The following is taken from Berestycki's notes.

Theorem 4.6. *Let P be a reversible matrix with respect to the invariant distribution π and let λ_j be its eigenvalues with $1 = \lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_n$. The for all $j \in \{1, \ldots, n\}$ we have*

$$1 - \lambda_j = \max_{\varphi_1, \ldots, \varphi_{j-1}} \min\{\mathcal{E}(f) : \|f\|_2 = 1, f \perp \varphi_1, \ldots, \varphi_{j-1}\}.$$

Proof. Let (f_j) be the eigenfunctions corresponding to the eigenvalues (λ_j). Let $\varphi_1, \ldots, \varphi_{j-1}$ be arbitrary functions. Consider $W = \text{span}(\varphi_1, \ldots, \varphi_{\text{j}-1})^\perp$. Then $\dim(W) \geq n - j + 1$, and hence $W \cap \text{span}(f_1, \ldots, f_j) \neq \varnothing$. So there exists g in the intersection. By normalising we can assume that $\|g\|_2 = 1$. Let $g = \sum_{i=1}^{j} a_j f_j$. Then $\sum_j a_j^2 = 1$ and we have

$$\mathcal{E}(g) = \langle (I - P)g, g \rangle_\pi = \left\langle \sum_{i=1}^{j} a_i(1 - \lambda_i) f_i, \sum_{i=1}^{j} a_i f_i \right\rangle_\pi = \sum_{i=1}^{j} a_i^2 (1 - \lambda_i) \leq 1 - \lambda_j.$$

Finally taking $\varphi_i = f_i$ for all $i \leq j - 1$ gives the equality. □

Corollary 4.7. *Let P and $\widetilde{P}$ be two transition matrices reversible with respect to the same invariant distribution π. Let $\mathcal{E}$ and $\widetilde{\mathcal{E}}$ be their Dirichlet forms and $(\lambda_i)_i$ and $(\widetilde{\lambda})_i$ their respective eigenvalues. If there exists a positive constant A such that for all $f : E \to \mathbb{R}$ we have $\widetilde{\mathcal{E}}(f) \leq A\mathcal{E}(f)$, then $1 - \widetilde{\lambda}_j \leq A(1 - \lambda_j)$ for all j.*

Theorem 4.8. *Let P and $\widetilde{P}$ be two transition matrices reversible with respect to the invariant distributions π and $\widetilde{\pi}$ respectively. Suppose that for each $(x, y) \in \widetilde{E}$ we pick a path Γ_{xy} in E and we set*

$$B = \max_{e \in E} \left(\frac{1}{Q(e)} \sum_{x,y : e \in \Gamma_{xy}} \widetilde{Q}(x, y) |\Gamma_{xy}| \right).$$

Then for all f we have $\widetilde{\mathcal{E}}(f) \leq B\mathcal{E}(f)$.

Proof. This proof is very similar to the proof of Theorem 4.5. We have

$$\begin{aligned} 2\mathcal{E}(f) &= \sum_{(x,y)} \widetilde{Q}(x, y)(f(x) - f(y))^2 = \sum_{(x,y)} \widetilde{Q}(x, y) \left(\sum_{e \in \Gamma_{xy}} \nabla f(e) \right)^2 \\ &\leq \sum_{x,y} \widetilde{Q}(x, y) |\Gamma_{xy}| \sum_{e \in \Gamma_{xy}} (\nabla f(e))^2 = \sum_e Q(e)(\nabla f(e))^2 \cdot \frac{1}{Q(e)} \sum_{x,y : e \in \Gamma_{xy}} \widetilde{Q}(x, y) |\Gamma_{xy}| \\ &\leq 2\mathcal{E}(f) B, \end{aligned}$$

where for the first inequality we used Cauchy-Schwarz. □

4.3 Bottleneck ratio

As before, we write $Q(x, y) = \pi(x)P(x, y)$ for any two states x, y and we define

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

Definition 4.9. The bottleneck ratio is defined to be

$$\Phi_* = \min_{S:\pi(S)\leq 1/2} \frac{Q(S, S^c)}{\pi(S)}.$$

Theorem 4.10. *For any irreducible transition matrix P we have*

$$t_{\text{mix}} = t_{\text{mix}}(1/4) \geq \frac{1}{4\Phi_*}.$$

Proof. Let A be such that $\pi(A) \leq 1/2$ and $Q(A, A^c)/\pi(A) = \Phi_*$. Then we have

$$\mathbb{P}_\pi(X_0 \in A, X_t \in A^c) \leq \sum_{i=1}^{t} \mathbb{P}_\pi(X_i \in A, X_{i+1} \in A^c) = t\mathbb{P}_\pi(X_0 \in A, X_1 \in A^c) = tQ(A, A^c).$$

Dividing through by $\pi(A)$ we obtain

$$\mathbb{P}_{\pi_A}(X_t \in A^c) \leq t\frac{Q(A, A^c)}{\pi(A)} = t\Phi_*.$$

Taking now $t = (4\Phi_*)^{-1}$ gives $\mathbb{P}_{\pi_A}(X_t \in A^c) \leq 1/4$, and therefore,

$$d(t) \geq \pi(A^c) - \mathbb{P}_{\pi|_A}(X_t \in A^c) \geq \frac{1}{2} - \frac{1}{4} = \frac{1}{4}$$

which completes the proof. □

Theorem 4.11 (Jerrum and Sinclair, Lawler and Sokal)**.** *Let P be a reversible transition matrix with respect to the invariant distribution π. Let γ be the spectral gap. Then we have*

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

Proof. We start with the easy direction that is the upper bound. Using the variational characterisation of γ from Theorem 4.3 we get

$$\gamma = \min_{f\neq 0, \mathbb{E}_\pi[f]=0} \frac{\mathcal{E}(f, f)}{\|f\|_2^2} = \min_{f\neq 0, \mathbb{E}_\pi[f]=0} \frac{\sum_{x,y}(f(x) - f(y))^2 Q(x, y)}{\sum_{x,y} \pi(x)\pi(y)(f(x) - f(y))^2}.$$

Let S be a set with $\pi(S) \leq 1/2$. Taking now $f(x) = -\pi(S^c)$ for $x \in S$ and $f(x) = \pi(S)$ for $x \in S$, we obtain

$$\gamma \leq \frac{2Q(S, S^c)}{2\pi(S)\pi(S^c)} \leq \frac{2Q(S, S^c)}{\pi(S)}.$$

Taking the minimum over all sets S with $\pi(S) \leq 1/2$ proves the upper bound.

We next turn to the lower bound. The proof will consist of three steps.

1st step Let f_2 be the eigenfunction corresponding to λ_2. Suppose without loss of generality that $\pi(f_2 > 0) \leq 1/2$, otherwise consider the function $-f_2$. Let $f = \max(f_2, 0)$. We claim that

$$(I - P)\psi(x) \leq \gamma\psi(x). \tag{4.2}$$

Indeed, if $f(x) = 0$, then this is obvious, since $f \geq 0$. If $f(x) > 0$, then $f(x) = f_2(x)$ and

$$(I - P)f(x) = f_2(x) - Pf(x) \leq f_2(x) - Pf_2(x) = \gamma f_2(x) = \gamma f(x),$$

where the inequality follows since $f \geq f_2$. Using that $f \geq 0$ and (4.2) give

$$\gamma \geq \frac{\mathcal{E}(f,f)}{\|f\|_2^2}. \tag{4.3}$$

2nd step For any non-negative function ψ define $S(\psi) = \{x : \psi(x) > 0\}$ and

$$h(\psi) = \inf\left\{\frac{Q(S, S^c)}{\pi(S)} : \varnothing \neq S \subseteq S(\psi)\right\}.$$

In this step we show that for any function $\psi \geq 0$

$$\mathcal{E}(\psi,\psi) \geq \frac{h(\psi)}{2}\|\psi\|_2^2. \tag{4.4}$$

Fix $\psi \geq 0$ and for every $t > 0$ define $S_t = \{x : \psi(x) > t\}$. Then by the definition of $h(\psi)$ for all t such that $S_t \neq \varnothing$ we have

$$\pi(S_t)h(\psi) \leq Q(S_t, S_t^c) = Q(S_t^c, S_t) \sum_{\psi(x) \leq t < \psi(y)} Q(x,y).$$

Multiplying both sides by $2t$ and integrating from 0 to ∞ we obtain

$$h(\psi)\int_0^\infty 2t\pi(\{x : \psi(x) > t\})\,dt \leq \sum_{\psi(x)<\psi(y)} \int_{\psi(x)}^{\psi(y)} 2tQ(x,y)\,dt = \sum_{\psi(x)<\psi(y)} (\psi(y)^2 - \psi(x)^2)Q(x,y)$$

$$= \frac{1}{2}\sum_{x,y}|\psi(x)^2 - \psi(y)^2|Q(x,y) \leq \frac{1}{2}\left(\sum_{x,y}(\psi(x) - \psi(y))^2 Q(x,y)\right)^{\frac{1}{2}} \cdot \left(\sum_{x,y}(\psi(x) + \psi(y))^2 Q(x,y)\right)^{\frac{1}{2}}$$

$$\leq \frac{1}{2}\sqrt{2\mathcal{E}(\psi)} \cdot 2\|\psi\|_2 = \sqrt{2\mathcal{E}(\psi)}\,\|\psi\|_2,$$

where for the inequality on the second line we used Cauchy Schwarz and for the final inequality we used $(a+b)^2 \leq 2(a^2+b^2)$. The left hand side on the first line above is equal to $h(\psi)\|\psi\|_2^2$. Thus rearranging proves (4.4) and completes the proof the second step.

3rd step Since we assumed $\pi(f_2 > 0) \leq 1/2$, using the definition of the bottleneck ratio and the definition of $h(f)$ we deduce

$$h(f) \geq \Phi_*.$$

This now combined with (4.3) and (4.4) completes the proof of the theorem. □

Remark 4.12. The bounds in Theorem 4.11 are the best one could hope for. Indeed, both bounds are achieved.

The lower bound is achieved for the lazy simple random walk on the cycle $\mathbb{Z}_n$. Indeed, in Section 3.2 it was proved that $t_{\text{rel}} \sim n^2/\pi^2$ as $n \to \infty$. It is easy to check that in this case $\Phi_* \asymp 1/n$.

The upper bound is achieved by the lazy simple random walk on the hypercube. Indeed, if we consider the set

$$A = \{x = (x_1, \ldots, x_n) \in \{0,1\}^n : x_1 = 0\},$$

then $\pi(A) = 1/2$ and $\Phi(A) = 1/(2n)$, which shows that $\Phi_* \leq 1/(2n)$. We proved in Section 3.2 that $\gamma = 1/n$, and hence $\Phi_* \geq 1/(2n)$, showing that $\Phi_* = 1/(2n)$, and hence the upper bound is sharp in this case.

4.4 Expander graphs

Definition 4.13. A sequence of graphs $G_n = (V_n, E_n)$ is called a (d, α)-expander family if $\lim_{n\to\infty} |V_n| = \infty$, for each n the graph G_n is d-regular and the bottleneck ratios $\Phi_*(G_n) \geq \alpha$ for all n.

Proposition 4.14. *Let G_n be a (d, α)-expander family. Then the mixing time of lazy simple random walk on G_n satisfies $t_{\text{mix}} = O(\log |V(G_n)|)$.*

Proof. Applying Theorem 4.11 we get that $\gamma \geq \alpha^2/2$, and hence the upper bound follows from Theorem 3.8. □

Claim 4.2. *Expander graphs of bounded degree have the fastest mixing time (up to constants) among all regular graphs.*

Proof. It is immediate that the diameter of an expander graph of bounded degree is at least (up to constants) $\log |V(G_n)|$. The claim then follows in view of the diameter lower bound on the mixing time. □

The definition of expanders does not make it clear why such graphs actually exist! Here we will show that a random 3-regular graph is an expander with high probability. Thus an expander exists!

Theorem 4.15. *There exists a random graph with $\mathbb{P}(\Phi_* > c) = 1 - o(1)$ as $n \to \infty$, where c is a positive constant.*

Proof. Here we describe Pinsker's method. Let $A = \{a_1, \ldots, a_n\}$ and $B = \{b_1, \ldots, b_n\}$. We will construct a bipartite graph between A and B. Let σ_1 and σ_2 be two independent permutations of $\{1, \ldots, n\}$. For every i we place three edges

$$E_i = \left\{(a_i, b_i), (a_i, b_{\sigma_1(i)}), (a_i, b_{\sigma_2(i)})\right\}.$$

We now claim that the resulting graph is an expander with high probability. We first prove that for any subset S of A with $|S| = k \leq n/2$ the number of neighbours $N(S)$ of S satisfies for δ sufficiently small

$$\mathbb{P}(N(S) \leq (1+\delta)k) = o(1) \text{ as } n \to \infty.$$

Indeed, $|N(S)|$ is at least k. We now want to find the probability that there are less than δk surplus vertices. To do this, we use a union bound over all possible sets of δk vertices and multiply it by

the probability that σ_1 and σ_2 fall within the specified set. So we have

$$\mathbb{P}(N(S) \leq (1+\delta)k) \leq \binom{n}{\delta k} \frac{\binom{(1+\delta)k}{\delta k}^2}{\binom{n}{k}^2}.$$

Using a union bound again we obtain

$$\mathbb{P}(\exists\, S \subseteq Q : |S| \leq n/2 \text{ and } N(S) \leq (1+\delta)|S|) \leq \sum_{k=1}^{n/2} \binom{n}{k}\binom{n}{\delta k} \frac{\binom{(1+\delta)k}{\delta k}^2}{\binom{n}{k}^2}.$$

It is now a calculus problem to show that for δ sufficiently small the quantity above tends to 0 as $n \to \infty$.

So we showed that all subsets $S \subseteq A$ of size $k \leq n/2$ have at least $(1+\delta)k$ neighbours with high probability. Similarly, the same result holds for all subsets of B. Using this, we will now show that $\Phi_* > \delta/2$.

So we now work on the high probability event above. Let $S \subseteq A \cup B$ of size $|S| \leq n$. We write $A' = S \cap A$ and $B' = S \cap B$. Wlog suppose that $|A'| \geq |B'|$. Then this implies that $|A'| \geq |S|/2$.

If $|A'| \leq n/2$, then by the assumption on all subsets of A of size $\leq n/2$ we get that A' will have at least $(1+\delta)|A'|$ neighbours in B. Since $|B'| \leq |S|/2 \leq |A'|$, it follows that A' will have al least $\delta|S|/2$ neighbours in $B \setminus B'$. These must be edges connecting S to S^c.

If $|A'| > n/2$, then take a subset A'' such that $|A''| = \lceil n/2 \rceil$. Then again, since $|B'| \leq n/2$, there will be at least $\delta|S|/2$ neighbours in $B \setminus B'$ and the corresponding edges will connect S to S^c.

Therefore, in either case the bottleneck ratio is lower bounded by $\delta/2$.

The graph produced by Pinsker's method is a multigraph. The expected number of triple edges is $1/n$ and the expected number of double edges is at most 2. Therefore, with probability bounded away from 0, the graph G_n will have at most 3 double edges and no triple edges. We can now subdivide every double edge and create a 3 regular graph. It is easy to check that the bottleneck ratio will be decreased by a multiplicative constant. □

5 Path coupling

Suppose we have a Markov chain with values in a space E which is endowed with a metric ρ satisfying $\rho(x,y) \geq \mathbf{1}(x \neq y)$.

Suppose that for all states x and y, there exists a coupling of $P(x,\cdot)$ and $P(y,\cdot)$ that contracts the distance in the sense

$$\mathbb{E}_{x,y}[\rho(X_1, Y_1)] \leq e^{-\alpha}\rho(x,y).$$

Define the diameter of E to be $\mathrm{diam}(E) = \max_{x,y} \rho(x,y)$. Iterating the above inequality we get that for all t

$$\mathbb{E}_{x,y}[\rho(X_t, Y_t)] \leq e^{-\alpha t}\mathrm{diam}(E).$$

This now gives a bound on the total variation mixing, since

$$\left\|P^t(x,\cdot) - P^t(y,\cdot)\right\|_{\mathrm{TV}} \leq \mathbb{P}_{x,y}(X_t \neq Y_t) \leq \mathbb{E}_{x,y}[\rho(X_t, Y_t)] \leq e^{-\alpha t}\mathrm{diam}(E).$$

Therefore,

$$t_{\mathrm{mix}}(\varepsilon) \leq \frac{1}{\alpha}\left(\log(\mathrm{diam}(E)) + \log(1/\varepsilon)\right).$$

In this section, we are going to see that when ρ is a path metric, to be defined below, then it suffices to check the contraction property only for neighbouring pairs x and y.

5.1 Transportation metric

The first step is to lift the metric on E to a metric between probability distributions on E. We define the transportation metric between the probability measures μ and ν

$$\rho_K(\mu,\nu) = \inf\{\mathbb{E}[\rho(X,Y)] : (X,Y) \text{ is a coupling of } \mu \text{ and } \nu\}. \tag{5.1}$$

We note that if $\mu = \delta_x$ and $\nu = \delta_y$, then $\rho_K(\mu,\nu) = \rho(x,y)$ and if $\rho(x,y) = \mathbf{1}(x \neq y)$, then $\rho_K(\mu,\nu) = \|\mu - \nu\|_{\mathrm{TV}}$.

Lemma 5.1. *There exists a coupling q_* of μ and ν such that*

$$\rho_K(\mu,\nu) = \sum_{(x,y)\in E\times E} q_*(x,y)\rho(x,y).$$

The coupling q_* is called the optimal ρ-coupling of μ and ν.

Proof. We can think of all the couplings q of μ and ν as vectors $(q(x,y))_{(x,y)}$ with $\sum_y q(x,y) = \mu(x)$ and $\sum_x q(x,y) = \nu(y)$. Therefore, the space of all couplings is a compact subset of the $|E|^2 - 1$ dimensional simplex. The function

$$q \mapsto \sum_{(x,y)} \rho(x,y)q(x,y)$$

is continuous on this set, and hence there exists q_* where the minimum is attained and we have

$$\sum_{(x,y)} \rho(x,y)q_*(x,y) = \rho_K(\mu,\nu).$$

This concludes the proof. □

Lemma 5.2. *The function ρ_K defines a metric on the space of probability distributions on E.*

Proof. It is obvious that $\rho_K(\mu,\nu) \geq 0$ and symmetric. Also, if $\rho_K(\mu,\nu) = 0$, then this means that

$$\sum_{(x,y)} \rho(x,y)q_*(x,y) = 0,$$

which implies that for all (x, y) for which $\rho(x, y) > 0$, then $q_*(x, y) = 0$. So q_* is supported on the diagonal $\{(x, x) : x \in E\}$. This immediately gives now that $\mu = \nu$.

We now prove the triangle inequality. Let μ, ν, η be three probability distributions. We will show

$$\rho_K(\mu, \eta) \leq \rho_K(\mu, \nu) + \rho_K(\nu, \eta). \tag{5.2}$$

Let $p(x, y)$ be the optimal ρ-coupling of μ and ν and let $q(y, z)$ be the optimal ρ-coupling of ν and η. Define now

$$r(x, y, z) = \frac{p(x, y)q(y, z)}{\nu(y)}.$$

Then this is a coupling of μ, ν, η and if $(X, Y, Z) \sim r$, then (X, Z) is a coupling of μ and η. By the triangle inequality for the metric ρ we get

$$\mathbb{E}[\rho(X, Z)] \leq \mathbb{E}[\rho(X, Y)] + \mathbb{E}[\rho(Y, Z)] = \rho_K(\mu, \nu) + \rho_K(\nu, \eta).$$

Since (X, Z) is a coupling of μ and η this proves (5.2). □

5.2 Path metric

Suppose that the Markov chain takes values in the vertex set of a graph G which is endowed with a length function defined on the edges of the graph. Note, however, that the transition matrix of the Markov chain does not have to obey the graph structure. The length function ℓ satisfies $\ell(x, y) \geq 1$ for all edges (x, y). If $x_0, \ldots, x_r$ is a path, then its length is defined to be $\sum_{i=0}^{r-1} \ell(x_i, x_{i+1})$. We now define the path metric

$$\rho(x, y) = \min\{\text{length of } \xi : \xi \text{ is a path from } x \text{ to } y\}.$$

By the assumption on ℓ, we get $\rho(x, y) \geq \mathbf{1}(x \neq y)$, so this gives

$$\mathbb{P}(X \neq Y) \leq \mathbb{E}[\rho(X, Y)]\,,$$

and hence taking the minimum over all couplings (X, Y) of μ and ν we obtain

$$\|\mu - \nu\|_{\mathrm{TV}} \leq \rho_K(\mu, \nu).$$

Theorem 5.3. *[Bubley and Dyer (1997)] Suppose that X takes values in the vertex set V of a graph G with length function ℓ and let ρ be the corresponding path metric. Suppose that for all edges (x, y) there exists a coupling (X_1, Y_1) of $P(x, \cdot)$ and $P(y, \cdot)$ such that*

$$\mathbb{E}_{x,y}[\rho(X, Y)] \leq e^{-\alpha}\rho(x, y).$$

Then for any probability measures μ and ν we have

$$\rho_K(\mu P, \nu P) \leq e^{-\alpha}\rho_K(\mu, \nu).$$

In particular,

$$d(t) \leq e^{-\alpha t}\mathrm{diam}(V),$$

where $\operatorname{diam}(V) = \max_{x,y} \rho(x,y)$.

Proof. We first establish the inequality $\rho_K(\mu P, \nu P) \le e^{-\alpha}\rho_K(\mu,\nu)$ for $\mu = \delta_x$ and $\nu = \delta_y$. Let $x = x_0, x_1, \dots, x_k = y$ be the path from x to y of the shortest length. Then by the triangle inequality for the transportation metric we get

$$\begin{aligned}\rho_K(P(x,\cdot),P(y,\cdot)) &\le \sum_{i=0}^{k-1} \rho_K(P(x_i,\cdot),P(x_{i+1},\cdot)) \le e^{-\alpha}\sum_{i=0}^{k-1}\rho(x_i,x_{i+1}) \\ &\le e^{-\alpha}\sum_{i=0}^{k-1}\ell(x_i,x_{i+1}) = e^{-\alpha}\rho(x,y),\end{aligned}$$

which proves it in this case.

For general measures μ and ν, let (X,Y) be an optimal ρ-coupling of μ and ν. Given $(X,Y) = (x,y)$, generate (X',Y') using an optimal ρ-coupling of $P(x,\cdot)$ and $P(y,\cdot)$. Then (X',Y') is a coupling of μP and νP. We then have

$$\begin{aligned}\rho_K(\mu P,\nu P) &\le \mathbb{E}\big[\rho(X',Y')\big] = \sum_{(x,y)} \mathbb{P}((X,Y)=(x,y))\,\mathbb{E}\big[\rho(X',Y') \,\big|\, (X,Y)=(x,y)\big] \\ &= \sum_{(x,y)} \mathbb{P}((X,Y)=(x,y))\cdot\rho_K(P(x,\cdot),P(y,\cdot)) \le e^{-\alpha}\sum_{(x,y)}\mathbb{P}((X,Y)=(x,y))\,\rho(x,y) \\ &= e^{-\alpha}\mathbb{E}[\rho(X,Y)] = e^{-\alpha}\rho_K(\mu,\nu),\end{aligned}$$

where the last equality follows from the fact that (X,Y) is an optimal ρ-coupling of μ and ν.

Iterating this inequality we obtain that for all t

$$\rho_K(\mu P^t,\nu P^t) \le e^{-\alpha t}\rho_K(\mu,\nu) \le e^{-\alpha t}\operatorname{diam}(V).$$

Taking now $\mu = \delta_x$ and $\nu = \pi$, gives the second inequality of the theorem. □

5.3 Applications

Colourings Let $G = (V,E)$ be a graph. We consider the set of all proper colourings of G, i.e. the set

$$\mathcal{X} = \{x \in \{1,\dots,q\}^V : x(v) \ne x(w) \quad \forall\ (v,w) \in E\}.$$

We want to sample a proper colouring uniformly at random from $\mathcal{X}$. We use Glauber dynamics for sampling which work as follows: choose a vertex w of V uniformly at random and update the colour at w by choosing a colour at random from the set of colours not taken by any of its neighbours. It is easy to check that this Markov chain is reversible with respect to the uniform distribution on the set $\mathcal{X}$.

Theorem 5.4. *Let G be a graph on n vertices with maximal degree Δ and let $q > 2\Delta$. Then the Glauber dynamics chain has mixing time*

$$t_{\mathrm{mix}}(\varepsilon) \le \left\lceil \left(\frac{q-\Delta}{q-2\Delta}\right) n(\log n - \log\varepsilon)\right\rceil.$$

Proof. Let x and y be two proper colourings of the graph. Then we define their distance to be $\rho(x,y) = \sum_v \mathbf{1}(x(v) \neq y(v))$, i.e. it is given by the number of vertices where they differ. For a vertex v in V we write $x(v)$ for the colour of v in the configuration $x \in \mathcal{X}$. We also write $A_v(x)$ for the set of allowed colours for v, i.e. the set of colours not present in any of the neighbours of v.

We define a coupling for (X_1, Y_1) when $X_0 = x$ and $Y_0 = y$ with $\rho(x,y) = 1$. Let v be the vertex where x and y differ. We pick the same uniform vertex w in both configurations. If w is not a neighbour of v, then we update both configurations in the same way, since all such w's have the same allowed colours in both configurations. If $w \sim v$, then suppose without loss of generality that $|A_w(x)| \leq |A_w(y)|$. We then pick a colour U uniformly at random from $A_w(y)$. If $U \neq x(v)$, then we update $x(w)$ to U. If, however $U = x(v)$, then we consider two cases: if $|A_w(x)| = |A_w(y)|$, then we set $x(w) = y(v)$. If $|A_w(x)| < |A_w(y)|$ (in which case we have $|A_w(y)| = |A_w(x)| + 1$), then we update $x(w)$ to a uniform colour from $A_w(x)$. It is straightforward to then check that $x(w)$ is updated to a uniform colour from $A_w(x)$.

We now need to calculate $\mathbb{E}_{x,y}[\rho(X_1, Y_1)]$. We notice that the distance increases to 2 when we pick a neighbour of v and it is updated differently in both configurations. If we pick v, then the distance goes down to 0, while picking any other vertex other than v or a neighbour of v, results in the distance remaining equal to 1. Putting all things together we obtain

$$\mathbb{E}_{x,y}[\rho(X_1, Y_1)] = 2 \cdot \frac{\deg v}{n} \cdot \frac{1}{|A_w(y)|} + 1 - \frac{1}{n} - \frac{\deg v}{n} \cdot \frac{1}{|A_w(y)|} = 1 - \frac{1}{n} + \frac{\deg v}{n} \cdot \frac{1}{|A_w(y)|}.$$

Using the bound $|A_w(y)| \geq q - \Delta$ and $\deg v \leq \Delta$, we immediately get

$$\mathbb{E}_{x,y}[\rho(X_1, Y_1)] \leq 1 - \frac{1}{n} + \frac{\Delta}{n} \cdot \frac{1}{q - \Delta} = 1 - \frac{1}{n}\left(1 - \frac{\Delta}{q - \Delta}\right) \leq \exp\left(-\alpha(q,\Delta)/n\right),$$

where $\alpha = \Delta/(q - \Delta)$, which is in $(0,1)$ by the assumption $q > 2\Delta$. Theorem 5.3 together with the fact that $\operatorname{diam}(\mathcal{X}) = n$ completes the proof. □

Approximate counting colourings

Theorem 5.5 (Jerrum and Sinclair)**.** *Let G be a graph on n vertices with maximal degree Δ satisfying $q > 2\Delta$. Then there exists a random variable W which can be simulated by running*

$$n \left\lceil \frac{n \log n + n \log(6eqn/\varepsilon)}{1 - \frac{\Delta}{q-\Delta}} \right\rceil \left\lceil \frac{27qn}{\eta\varepsilon^2} \right\rceil$$

Glauber updates and it satisfies

$$\mathbb{P}\big((1-\varepsilon)|\mathcal{X}|^{-1} \leq W \leq (1+\varepsilon)|\mathcal{X}|^{-1}\big) \geq 1 - \eta.$$

The idea of the proof is to define a sequence of sets of proper colourings, $\mathcal{X}_k$, run Glauber dynamics on them, and approximate $|\mathcal{X}_{k-1}|/|\mathcal{X}_k|$. Then take the product.

Fix an ordering of the vertices of $G = \{v_1, \ldots, v_n\}$ and fix a proper colouring x_0. Define

$$\mathcal{X}_k = \{x \in \mathcal{X} : x(v_j) = x_0(v_j) \quad \forall\ j > k\}.$$

In the proof of Theorem 5.5 we will need to use that $|\mathcal{X}_{k-1}|/|\mathcal{X}_k|$ is not too small.

Lemma 5.6. *Let $q > 2\Delta$. Then for all k we have*

$$\frac{|\mathcal{X}_{k-1}|}{|\mathcal{X}_k|} \geq \frac{1}{eq}.$$

Proof. Suppose that v_k has r neighbours in the set $\{v_1, \ldots, v_{k-1}\}$. Start with the uniform distribution on $\mathcal{X}_k$ and update in the order given by the ordering of the graph the colours at the r neighbours of v_k and last the colour at v_k as follows: for each vertex to be updated choose a colour at random from the set of allowed colours. Then this clearly preserves the uniform distribution on $\mathcal{X}_k$. Let Y be the configuration of colours at the end of this process. Then Y is uniform on $\mathcal{X}_k$. Let A be the event that at the end of this process the colour at each of the r neighbours is different to $x_0(v_k)$ and the colour of v_k is updated to $x_0(v_k)$. Then $Y \in \mathcal{X}_{k-1}$ if and only if the event A occurs. So we have

$$\begin{aligned}\frac{|\mathcal{X}_{k-1}|}{|\mathcal{X}_k|} = \mathbb{P}(Y \in \mathcal{X}_{k-1}) = \mathbb{P}(A) &\geq \left(1 - \frac{1}{q-\Delta}\right)^r \frac{1}{q} \geq \left(1 - \frac{1}{q-\Delta}\right)^\Delta \frac{1}{q} \\ &\geq \left(\frac{\Delta}{\Delta+1}\right)^\Delta \frac{1}{q} \geq \frac{1}{eq},\end{aligned}$$

where for the first inequality we used that the set of allowed colours for every vertex is at least $q - \Delta$ and for the penultimate inequality we used the assumption that $q \geq 2\Delta + 1$. □

Proof of Theorem 5.5. First notice that $|\mathcal{X}_0| = 1$ and $|\mathcal{X}_n| = |\mathcal{X}|$. So we have

$$\prod_{i=0}^{n-1} \frac{|\mathcal{X}_i|}{|\mathcal{X}_{i+1}|} = \frac{1}{|\mathcal{X}|}.$$

The strategy of the proof is to define a random variable W_i which will be close to $\frac{|\mathcal{X}_{i-1}|}{|\mathcal{X}_i|}$ with high probability. Then we will define $W = \prod_{i=1}^n W_i$ and get that it will be close to $1/|\mathcal{X}|$.

Running Glauber dynamics on $\mathcal{X}_k$ with frozen boundary conditions at the vertices $v_{k+1}, \ldots, v_n$ will generate a uniform element of $\mathcal{X}_k$. The same proof as in Theorem 5.4 gives that if

$$t = \left\lceil \frac{n \log n + n \log(6eqn/\varepsilon)}{1 - \frac{\Delta}{q-\Delta}} \right\rceil,$$

then the distribution of Glauber dynamics on $\mathcal{X}_k$ at time t is within $\varepsilon/(6eqn)$ in total variation from the uniform distribution on $\mathcal{X}_k$.

We now take $a_n = \lceil 27qn/(\eta\varepsilon^2) \rceil$ independent copies of Glauber dynamics on $\mathcal{X}_k$ each run for t steps independently for different k's. For $i = 1, \ldots, a_n$ we let

$$Z_{k,i} = \mathbf{1}(i\text{-th sample is in } \mathcal{X}_{k-1}) \quad \text{and} \quad W_k = \frac{1}{a_n} \sum_{i=1}^{a_n} Z_{k,i}.$$

Using the mixing property at time t we get that

$$\left|\mathbb{E}[Z_{k,i}] - \frac{|\mathcal{X}_{k-1}|}{|\mathcal{X}_k|}\right| \leq \frac{\varepsilon}{6eqn} \quad \text{and} \quad \left|\mathbb{E}[W_k] - \frac{|\mathcal{X}_{k-1}|}{|\mathcal{X}_k|}\right| \leq \frac{\varepsilon}{6eqn}. \tag{5.3}$$

The second inequality together with Lemma 5.6 now give

$$1 - \frac{\varepsilon}{6n} \leq \frac{|\mathcal{X}_k|}{|\mathcal{X}_{k-1}|}\mathbb{E}[W_k] \leq 1 + \frac{\varepsilon}{6n}. \tag{5.4}$$

We now define $W = \prod_{i=1}^n W_i$. We will shortly show that each W_k is concentrated around its expectation, which is close to $|\mathcal{X}_{k-1}|/|\mathcal{X}_k|$. So by taking the product of W_i's in the definition of W we will get that W is close to the product of $|\mathcal{X}_{i-1}|/|\mathcal{X}_i|$ for $i = 1, \ldots, n$, which is equal to $1/|\mathcal{X}|$, since $|\mathcal{X}_0| = 1$ and $|\mathcal{X}_n| = |\mathcal{X}|$. Using the independence of W_k's we get

$$\frac{\mathrm{Var}\,(W)}{(\mathbb{E}[W])^2} = \frac{\mathbb{E}\big[W^2\big] - (\mathbb{E}[W])^2}{(\mathbb{E}[W])^2} = \frac{\prod_{i=1}^n \mathbb{E}\big[W_i^2\big]}{\prod_{i=1}^n (\mathbb{E}[W_i])^2} - 1 = \prod_{i=1}^n \left(1 + \frac{\mathrm{Var}\,(W_i)}{(\mathbb{E}[W_i])^2}\right) - 1. \tag{5.5}$$

Using the independence of $Z_{k,i}$ for different i's we obtain for all k

$$\mathrm{Var}\,(W_k) = \frac{1}{a_n^2}\sum_{i=1}^{a_n} \mathbb{E}[Z_{k,i}]\,(1 - \mathbb{E}[Z_{k,i}]) \leq \frac{1}{a_n}\mathbb{E}[W_k]\,,$$

which means that

$$\frac{\mathrm{Var}\,(W_k)}{(\mathbb{E}[W_k])^2} \leq \frac{1}{a_n \mathbb{E}[W_k]} \leq \frac{3q}{a_n} \leq \frac{\eta\varepsilon^2}{9n}, \tag{5.6}$$

where for the second inequality we used that

$$\mathbb{E}[W_k] \geq \frac{1}{eq} - \frac{\varepsilon}{6eqn} \geq \frac{1}{3q},$$

which follows from (5.3) and Lemma 5.6. Plugging the bound of (5.6) into (5.5) we obtain

$$\frac{\mathrm{Var}\,(W)}{(\mathbb{E}[W])^2} \leq \prod_{i=1}^n \left(1 + \frac{\eta\varepsilon^2}{9n}\right) - 1 \leq e^{\eta\varepsilon^2/9} - 1 \leq 2\eta\varepsilon^2/9,$$

using that $e^x \leq 1 + 2x$ for $x \in [0, 1]$. Therefore, by Chebyshev's inequality we get

$$\mathbb{P}\left(|W - \mathbb{E}[W]\,| \geq \frac{\varepsilon\mathbb{E}[W]}{2}\right) \leq \eta.$$

By (5.4) we deduce

$$1 - \frac{\varepsilon}{6} \leq \left(1 - \frac{\varepsilon}{6n}\right)^n \leq |\mathcal{X}| \cdot \mathbb{E}[W] \leq \left(1 + \frac{\varepsilon}{6n}\right)^n \leq e^{\varepsilon/6} \leq 1 + \frac{\varepsilon}{3}.$$

Therefore,

$$\left|\mathbb{E}[W] - \frac{1}{|\mathcal{X}|}\right| \leq \frac{\varepsilon}{3|\mathcal{X}|}.$$

Using this we now see that on the event $\{|W - \mathbb{E}[W]| < \varepsilon\mathbb{E}[W]/2\}$ we have that

$$\left|W - \frac{1}{|\mathcal{X}|}\right| \leq \frac{\varepsilon}{3|\mathcal{X}|} + \frac{\varepsilon\mathbb{E}[W]}{2} \leq \frac{\varepsilon}{3|\mathcal{X}|} + \frac{\varepsilon}{2}\left(\frac{1}{|\mathcal{X}|} + \frac{\varepsilon}{3|\mathcal{X}|}\right) \leq \frac{\varepsilon}{|\mathcal{X}|}.$$

So in order to simulate the random variable W we need to run at most a_n copies of t steps of Glauber dynamics on each $\mathcal{X}_k$ for $k = 1, \ldots, n$ and this concludes the proof. □

5.4 Ising model

Definition 5.7. Let V and S be two finite sets. Let $\mathcal{X}$ be a subset of V^S and π a distribution on $\mathcal{X}$. The *Glauber dynamics* on $\mathcal{X}$ is the Markov chain that evolves as follows: when at state x, we pick a vertex of V uniformly at random and a new state is chosen with probability equal to π conditioned on the set of states equal to x at all vertices except for v. Formally, for each $x \in \mathcal{X}$ and $v \in V$ we define $A(x, v) = \{y \in \mathcal{X} : y(w) = x(w),\ \forall\ w \neq v\}$ and $\pi^{x,v}(y) = \mathbf{1}(y \in A(x, v))\pi(y)/\pi(A(x, v))$. So when at state $x \in \mathcal{X}$, the Glauber dynamics are defined by picking v uniformly at random from V and then choosing a new state according to $\pi^{x,v}$.

Remark 5.8. It is straightforward to check that the Glauber dynamics is a reversible Markov chain with respect to the distribution π.

Ising model. Let $G = (V, E)$ be a finite connected graph. The Ising model on G is the probability distribution on $\{-1, 1\}^V$ given by

$$\pi(\sigma) = \frac{1}{Z(\beta)} \cdot \exp\left(\beta \sum_{(i,j)\in E} \sigma(i)\sigma(j)\right),$$

where $\sigma \in \{-1, 1\}^V$ is a spin configuration. The parameter $\beta > 0$ is called the inverse temperature and the partition function $Z(\beta)$ is the normalising constant in order for π to be a probability distribution. When $\beta = 0$, then all spin configurations are equally likely, which means that π is uniform on $\{-1, 1\}^V$. When $\beta > 0$, the distribution π favours spin configurations where the spins of neighbouring vertices agree.

The Glauber dynamics for the Ising model evolve as follows: when at state $\sigma \in \{-1, 1\}^V$, a vertex v is picked uniformly at random and the new state $\sigma' \in \{-1, 1\}^V$ with $\sigma'(w) = \sigma(w)$ for all $w \neq v$ is chosen with probability

$$\frac{\pi(\sigma')}{\pi(A(\sigma, v))} = \frac{\pi(\sigma')}{\pi(\{z : z(w) = \sigma(w),\ \forall\ w \neq v\})} = \frac{e^{\beta\sigma'(v)S_v(\sigma)}}{e^{\beta S_v(\sigma)} + e^{-\beta S_v(\sigma)}},$$

where $S_v(\sigma) = \sum_{w\sim v} \sigma(w)$ with $i \sim j$ meaning that (i, j) is an edge of G.

Definition 5.9. Suppose that X is a Markov chain taking values in a partially ordered set $(S, \preceq)$. A coupling of two chains $(X_t, Y_t)_t$ is called monotone, if whenever $X_0 \preceq Y_0$, then $X_t \preceq Y_t$ for all t. The Markov chain X is called monotone, if for every two ordered initial states, there exists a monotone coupling.

Glauber dynamics for the Ising model is a monotone chain. We define the ordering $\sigma \preceq \sigma'$ if for all v we have $\sigma(v) \leq \sigma'(v)$. Indeed, suppose the current state is σ and the vertex chosen to be updated is v. Then one way to sample the new state is to take a uniform random variable U in $[0,1]$ and set the spin at v to be $+1$ if

$$U \leq \frac{1+\tanh(\beta S_v(\sigma))}{2}$$

and -1 otherwise. Since $\frac{1+\tanh(\beta S_v(\sigma))}{2}$ is non-decreasing in σ, it follows that the coupling is monotone.

6 Coupling from the past

6.1 Algorithm

Coupling from the past is an ingenious algorithm invented by Propp and Wilson in 1996 to exactly sample from the invariant distribution π. In order to describe it we start with the random function representation of a Markov chain with transition matrix P.

Lemma 6.1. *Let X be a Markov chain on S with transition matrix P. There exists a function $f : S \times [0,1] \to S$ such that if (U_i) is an i.i.d. sequence of random variables uniform on $[0,1]$, then*

$$X_{n+1} = f(X_n, U_n).$$

We can think of the function f as a grand coupling of X, in the sense that we couple all transitions from all starting points using the same randomness coming from the uniform random variables.

Let $(U_i)_{i\in\mathbb{Z}}$ be i.i.d. distributed as $\mathcal{U}[0,1]$. For every $t \in \mathbb{Z}$ we let $f_t : S \to S$ be given by

$$f_t(x) = f(x, U_t).$$

So for a Markov chain X, the functions f_t will define the evolution of X, i.e. $X_{t+1} = f_t(X_t)$.

For $s < t$ we now define

$$F_s^t(x) = (f_{t-1} \circ \ldots \circ f_s)(x) = f_{t-1}(f_{t-2}(\ldots f_s(x)\ldots)).$$

The function F_s^t gives the evolution of the Markov chain from time s to t and so for all x and y we have

$$\mathbb{P}\big(F_s^t(x) = y\big) = P^{t-s}(x,y).$$

We call the maps F_0^t the forward maps for the chain and the maps F_{-t}^0 the backward maps for $t > 0$. Now notice that if for some $t > 0$ we have F_{-t}^0 is a constant function, then for all $s > t$ we also have that F_{-s}^0 is a constant function equal to F_{-t}^0, since

$$F_{-s}^0 = F_{-t}^0 \circ (f_{-t-1} \circ \ldots f_{-s}).$$

The idea of coupling from the past is that under ergodicity assumptions, there will exist a random time T at which F_{-T}^0 will be a constant function and F_{-T}^0 will be distributed according to π.

Theorem 6.2 (Coupling from the past (Propp and Wilson))**.** *Let X be a Markov chain that is irreducible and aperiodic on the finite state space S with invariant distribution π. Then there exists an almost surely finite random time T satisfying F^0_{-T} is a constant function and the unique value F^0_{-T} (i.e. $F^0_{-T}(S)$) is distributed according to π.*

Remark 6.3. We note that it is crucial that in the algorithm above we consider time backwards. Indeed, if not, then if F^t_0 is constant for $t > 0$, then of course F^{t+1}_0 would also be a constant function, but not necessarily equal to F^t_0.

Proof of Theorem 6.2. By the ergodicity assumptions on X it follows that there exists L sufficiently large and $\varepsilon > 0$ so that

$$\mathbb{P}\big(F^0_{-L} \text{ is a constant function}\big) \geq \varepsilon.$$

By the i.i.d. property of the maps f_i we get that each of the maps $F^0_{-L}, F^{-L}_{-2L}, \ldots$ has probability at least ε of being constant and these events are independent. Therefore, with probability 1 one of these events will eventually happen, which means that $T < \infty$ a.s.

We now turn to prove that $F^0_{-T}(S) \sim \pi$.

It is clear by the i.i.d. property of the maps f_i that for all $t > 0$ we have

$$\mathbb{P}\big(F^0_{-t}(x) = y\big) = \mathbb{P}\big(F^t_0(x) = y\big)\,. \tag{6.1}$$

By the convergence to equilibrium theorem, we get that for all x and y

$$\lim_{t\to\infty} \mathbb{P}\big(F^t_0(x) = y\big) = \lim_{t\to\infty} P^t(x, y) = \pi(y).$$

This together with (6.1) give

$$\lim_{t\to\infty} \mathbb{P}\big(F^0_{-t}(x) = y\big) = \pi(y).$$

Since after time T, i.e. for $t > T$ the functions F^0_{-t} are all equal to F^0_{-T}, the above implies that $F^0_{-T}(S)$ is distributed according to π and this concludes the proof. □

6.2 Monotone chains

Let as above T be the first time that F^0_{-T} is a constant function and let C be the coalescence time, i.e. the first time t that F^t_0 is constant.

Claim 6.1. *The times T and C have the same distribution.*

Proof. Since for all t we have that F^t_0 and F^0_{-t} have the same distribution, it follows that for all k

$$\mathbb{P}(T > k) = \mathbb{P}\big(F^0_{-k} \text{ is not constant}\big) = \mathbb{P}\Big(F^k_0 \text{ is not constant}\Big) = \mathbb{P}(C > k)\,,$$

which concludes the proof. □

We now restrict our attention to monotone chains. Let $(S, \preceq)$ be a partially ordered set and suppose that it contains two elements $\widehat{0}$ and $\widehat{1}$ such that $\widehat{0} \leq x \leq \widehat{1}$ for all $x \in S$. Suppose that the monotonicity is preserved under the map f. Using monotonicity, we see that if there exists a time t such that $F^0_{-t}(\widehat{0}) = F^0_{-t}(\widehat{1})$, then since all the other states are sandwiched between $\widehat{0}$ and $\widehat{1}$, it follows that F^0_{-t} is a constant function. So for monotone chains, we do not need to run Markov chains starting from all initial states, but only from $\widehat{0}$ and $\widehat{1}$, making the algorithm computationally more efficient.

Theorem 6.4. *Let ℓ be the size of the largest totally ordered subset of S (or in other words the length of the longest chain in S). Then*

$$\mathbb{E}[T] \leq 2t_{\mathrm{mix}}(1 + \log_2 \ell).$$

Proof. Using Claim 6.1 it suffices to prove the bound of the statement for $\mathbb{E}[C]$. We start by proving that for all k

$$\frac{\mathbb{P}(C > k)}{\ell} \leq \overline{d}(k) \leq \mathbb{P}(C > k)\,. \tag{6.2}$$

For every x we write $h(x)$ for the length of the longest chain having x as the top element. Let (X_t) and (Y_t) be two Markov chains started from $\widehat{0}$ and $\widehat{1}$ respectively at time 0 coupled using the same maps f. Then we see that if $X_t \prec Y_t$, then $h(X_t) + 1 \leq h(Y_t)$. So we have

$$\mathbb{P}(C > k) = \mathbb{P}(X_k \neq Y_k) \leq \mathbb{E}[h(Y_k) - h(X_k)]\,,$$

where we used Markov's inequality. Let $(\widetilde{X}, \widetilde{Y})$ be the optimal coupling of the distributions of X_k and Y_k. Then we get

$$\begin{aligned}\mathbb{E}[h(Y_k) - h(X_k)] &= \mathbb{E}\Big[\Big(h(\widetilde{Y}) - h(\widetilde{X})\Big)\,\mathbf{1}(\widetilde{Y} \neq \widetilde{X})\Big] \leq \mathbb{P}\Big(\widetilde{Y} \neq \widetilde{X}\Big) \\ &\leq \Big(\max_x h(x) - \min_x h(x)\Big)\,\mathbb{P}\Big(\widetilde{X} \neq \widetilde{Y}\Big) \leq \ell\,\|\mathcal{L}(Y_k) - \mathcal{L}(X_k)\|_{\mathrm{TV}} \leq \ell\overline{d}(k),\end{aligned}$$

which proves the first inequality. The second one follows immediately from the coupling upper bound on total variation distance.

We next claim that for all k_1, k_2 we have $\mathbb{P}(C > k_1 + k_2) \leq \mathbb{P}(C > k_1)\,\mathbb{P}(C > k_2)$. Indeed, this follows since if $F_0^{k_1+k_2}$ is not a constant, then this means that both $F_0^{k_1}$ and $F_{k_1}^{k_1+k_2}$ are not constant. Using that these two last events are independent proves the sub-multiplicativity.

Using the sub-multiplicativity, we get that $\mathbb{P}(C > ik) \leq \mathbb{P}(C > k)^i$ for all k and i. Therefore,

$$\mathbb{E}[C] = \sum_{i=0}^{\infty} \mathbb{P}(C > i) \leq \sum_{i=0}^{\infty} k\mathbb{P}(C > ik) \leq \sum_{i=0}^{\infty} k\mathbb{P}(C > k)^i = \frac{k}{\mathbb{P}(C \leq k)}. \tag{6.3}$$

Let now $k = t_{\mathrm{mix}} \log_2(2\ell)$. Then using sub-multiplicativity of $\overline{d}$ we have

$$\overline{d}(k) \leq (\overline{d}(t_{\mathrm{mix}}))^{\log_2(2\ell)} \leq \frac{1}{2^{\log_2(2\ell)}} = \frac{1}{2\ell}.$$

From (6.2) we obtain

$$\mathbb{P}(C > k) \leq \ell \overline{d}(k) \leq \frac{1}{2},$$

and hence from (6.3) we deduce

$$\mathbb{E}[C] \leq 2t_{\text{mix}} \log_2(2\ell) = 2t_{\text{mix}}(1 + \log_2 \ell)$$

and this completes the proof. □

Remark 6.5. From the theorem above we see that the expected running time of the coupling from the past algorithm is governed by the mixing time of the chain. So if a chain is rapidly mixing, then it is also rapidly coupling.

References

[1] David A. Levin, Yuval Peres, and Elizabeth L. Wilmer. *Markov chains and mixing times*. American Mathematical Society, Providence, RI, 2009. With a chapter by James G. Propp and David B. Wilson.