Mixing times of Markov Chains

Lecture notes from Cambridge Part III course from Michaelmas 2024 and 2025

Contents

0 Prerequisites and Course Description

An ergodic Markov chain is known to converge to its equilibrium distribution as time goes to infinity. But how long should one wait until the distribution of the chain is close to the invariant one? How many times should one shuffle a deck of cards until the order becomes uniform? This question lies at the heart of the modern theory of mixing times for Markov chains. The classical theory of Markov chains studied fixed chains, focusing on large-time asymptotics of their distribution. Recently, the need to analyse large spaces has increased, shifting the focus to studying asymptotics of the *mixing time* (the first time distribution of the chain gets close enough to invariant one) as the size of the state space tends to infinity.

In this course, we will develop the basic theory and some of the main techniques and tools from probability and spectral theory used to estimate mixing times. We will apply them to study the mixing time of several chains of interest. We shall also discuss the *cutoff phenomenon*, first discovered by Diaconis in the context of card shuffling, which says that a Markov chain converges to equilibrium abruptly. This phenomenon seems to be widespread but it remains a challenging question to obtain criteria for cutoff for general classes of chains.

This course assumes almost no background, except for prior exposure to Markov chains at an elementary level. Parallel attendance of the first part of Advanced Probability, covering martingales, is encouraged but not necessary.

0.1 Background

Notation: In this course we will use $\mathbb{N} = \{1, 2, \ldots\}$ and $\mathbb{Z}^+ = \{0, 1, \ldots\}$.

We start by giving some definitions and results on Markov Chains which we do not prove. For proofs, more details and examples see any undergraduate Markov chains course notes, e.g. these.

Markov chains are processes where given the current step the next step of the process is independent from the past. Formally, we can define a Markov chain as a random process specified by a *state space, transition matrix* and *initial distribution* as follows:

Definition 0.1 (Markov Chain)**.** For a set S called the *state space*, a *transition matrix* P on S, (i.e. any $P = (P(x, y))_{x,y \in S}$ which is *stochastic*, which means that it satisfies $\forall x \in S$, $\sum_{y \in S} P(x, y) = 1$, and $\forall x, y \in S, P(x, y) \geq 0$) and an *initial distribution* ν on S, (i.e. $\nu = (\nu(x))_{x \in S}$ such that $\sum_{x \in S} \nu(x) = 1$ and $\forall x, \in S, \nu(x) \geq 0$) we can define a *Markov chain* with parameters (S, P, ν) to be a sequence of random variables $(X_t)_{t \geq 0}$ taking values in S such that

$$\mathbb{P}(X_0 = x_0, \ldots, X_{n-1} = x_{n-1}, X_t = x_t) = \nu(x_0) P(x_0, x_1) \ldots P(x_{t-1}, x_t)$$

for all $x_0, \ldots, x_t \in S$. //

Alternatively, we can think of a Markov chain as a process which satisfies the Markov property

Definition 0.2 (Markov Property)**.** *Markov chain* with transition matrix P, is any sequence of random variables $(X_t)_{t \geq 0}$ taking values in the set S, which satisfies the *Markov property*, which states that

$$\mathbb{P}(X_t = x_t \mid X_0 = x_0, \ldots, X_{t-1} = x_{t-1}) = \mathbb{P}(X_t = x_t \mid X_{t-1} = x_{t-1}) = P(x_{t-1}, x_t)$$

for all $x_0, \ldots, x_t \in S$ such that $\mathbb{P}(X_0 = x_0, \ldots, X_{t-1} = x_{t-1}) > 0$. //

Notation: For any event A and any $i \in S$ we will write $\mathbb{P}_i(A) = \mathbb{P}(A \mid X_0 = i)$ and $\mathbb{E}_i[A] = \mathbb{E}[A \mid X_0 = i]$ and for $i, j \in S$ and $t \in \mathbb{N}$ we write $p_t(i, j) = p_{i,j}(t) = P^t(i, j)$ and so we have that $\mathbb{P}_i(X_1 = j) = P(i, j)$ and

$$\begin{aligned} P^t(i, j) &= \sum_{x_1, \ldots x_{t-1} \in S} P(i, x_1) P(x_1, x_2) \ldots P(x_{t-1}, j) \\ &= \sum_{x_1, \ldots x_{t-1} \in S} \mathbb{P}_i(X_1 = x_1, X_2 = x_2 \ldots X_{t-1} = x_{t-1}, X_t = j) = \mathbb{P}_i(X_t = j)\,. \end{aligned}$$

We will denote by T_i the hitting time of state i, i.e. $T_i = \inf\{t \geq 0 : X_t = i\}$ and by T_i^+ the return time to state i, i.e. $T_i^+ = \inf\{t \geq 1 : X_t = i\}$. Similarly, if A is a set $T_A = \inf\{t \geq 0 : X_t \in A\}$ is the hitting time of set A and by $T_A^+ = \inf\{t \geq 1 : X_t \in A\}$ is the return time to set A.

For a probability distribution μ on S we write $X_t \sim \mu$ to mean that the law of X_t is μ. We also have that given $X_0 \sim \mu$ then $\mu P = (\sum_{j \in S} \mu(j) P(j, i))_{i \in S}$ is the law of X_1 (i.e. $X_1 \sim \mu P$). Similarly, if $X_0 \sim \mu$ then $X_t \sim \mu P^t$.

Theorem 0.3 (Stopping times and strong Markov property)**.** *A radnom variable T taking values in $\{0, 1, 2, 3, \ldots\} \cup \{\infty\}$ is called a stopping time if the event $\{T = t\}$ depends only on $X_0, X_1, \ldots, X_t$ for all $t \geq 0$. Strong Markov property states that for a Markov chain X on S with transition matrix P, for any stopping time T, state $i \in S$ we have that conditional on $T < \infty$ and $X_T = i$ the process $(X_{T+t})_{t=0}^{\infty}$ is a Markov chain starting from i with transition matrix P which is independent of $X_0, X_1, \ldots X_T$.*

Definition 0.4 (Irreducibility)**.** A Markov chain is called *irreducible* if for all $i, j \in S$ there is some $t \in \mathbb{N}$ such that $P^t(i,j) > 0$. We say that states $i, j \in S$ communicate if there are some $t, s \in \mathbb{N}$ such that $P^t(i,j) > 0$ and $P^s(j,i) > 0$. This defines an equivalence relation on S and classes of this relation are called communicating classes. So chain is irreducible if it consists of one communicating class only. //

Definition 0.5 (Recurrence)**.** A state $i \in S$ is called recurrent if $\mathbb{P}_i\big(T_i^+ < \infty\big) = 1$ and it is called *positive recurrent* if $\mathbb{E}_i\big[T_i^+\big] < \infty$. Markov chain is called recurrent if all states are recurrent and it is called positive recurrent if all states are positive recurrent. A state (or chain) which is not recurrent is called transient. //

Remark 0.6. *In an irreducible chain if one state is recurrent then all states are recurrent and if one state is positive recurrent then all states are positive recurrent (i.e. recurrence and positive recurrence are properties of communicating classes).*

Definition 0.7 (Invariant distribution)**.** A measure π is called *invariant measure* if $\pi P = \pi$. If π is also a distribution it is called *invariant distribution* (also known as *stationary* or *equilibrium distribution*). //

Remark 0.8. *Invariant distribution exists for any positive recurrent chain and if the chain is irreducible then it is unique and has full support (i.e. for all $i \in S$, $\pi(i) > 0$). Moreover, for an irreducible chain this distribution is given by $\pi(i) = \frac{1}{\mathbb{E}_i\left[T_i^+\right]}$ for all $i \in S$. Furthermore, any irreducible chain on a finite state space is positive recurrent and has a unique and fully supported invariant distribution.*

In this course, we will only consider irreducible Markov chains on a finite state space. Therefore all chains we study will have unique invariant distribution. Unless otherwise states, we will denote a transition matrix of X by P, its invariant distribution by π and the state space by S.

Definition 0.9 (Time reversal)**.** Let π be the invariant distribution and suppose that $X_0 \sim \pi$. Fix $T \in \mathbb{N}$ and consider the chain $Y_t = X_{T-t}$ for all $t \in \{0, \dots, T\}$. Then $(Y)_{t=0}^T$ is also a Markov chain with a transition matrix given by

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

We call P^* the *reversal* of P. //

It is not hard to check by writing out the definition of conditional probability that

$$\mathbb{P}(X_0 = y \mid X_1 = x, X_0 \sim \pi) = P^*(x,y)$$

and so the above definition is consistent.

It is also easy to check that P^* is the adjoint operator, in the sense that for all $f, g : S \to \mathbb{R}$ we have

$$\langle Pf, g\rangle_\pi = \langle f, P^*g\rangle_\pi, \quad \text{where } \langle f, g\rangle_\pi = \mathbb{E}_\pi[fg] = \sum_x \pi(x)f(x)g(x).$$

Definition 0.10 (Reversible chain)**.** A Markov chain X is called (*time*) *reversible* if for all T when $X_0 \sim \pi$, then $(X_0, \dots, X_T)$ has the same distribution as $(X_T, \dots, X_0)$. This is equivalent to the detailed balance equations, i.e. equations saying that

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

//

We recall that if some distribution on S satisfies detailed balance equations then it is an invariant distribution. Therefore when trying to find an invariant distribution it is a good idea to try finding a distribution which satisfies detailed balance equations rather than solving $\pi P = \pi$ (as this is harder to solve). Here are a couple of examples of chains which satisfy detailed balance and are therefore time reversible.

Example 0.11 (Simple random walk)**.** *Let $G = (V, E)$ be a connected, undirected finite graph. Denote by* $\deg(v)$ *the degree of v which is defined to be the total number of neighbours of v (i.e. number of vertices u such that u and v are connected with an edge, which is also just the number of edges coming out of vertex v). A simple random walk on G is a Markov chain on state space V with a transition matrix given by*

$$P(v, u) = \frac{1}{\deg(v)} \mathbb{1}_{u \sim v},$$

where by $u \sim v$ we mean that u and v are neighbours, i.e. joined by an edge. Random walk on G is reversible and checking detailed balance equations gives that the invariant distribution is given by

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

Example 0.12 (Birth and death chains)**.** *A birth and death chain is a Markov chain on a state space which is an interval in $\mathbb{Z}$ (or all of $\mathbb{Z}$) and where the only possible transitions are by $+1$, -1 or staying in place. If such a chain is irreducible and positive recurrent it also satisfies detailed balance equations and therefore is reversible. This means that a biased random walk on a finite interval $\{0, \dots n\}$ (which is a Markov chain which for some $p \in (0, 1)$ moves by $+1$ with probability p and by -1 otherwise, except when at endpoints 0 or n, when it has some probability in $[0, 1)$ to stay in place and otherwise moves to 1 or $n - 1$, respectively) is reversible and we can find π by solving detailed balance (just relate $\pi(i)$ to $\pi(0)$ and normalise).*

Example 0.13 (Markov chain on the tree)**.** *We call a graph which doesn't have cycles a tree and we usually label it $T = (V, E)$. A general Markov chain transition matrix P on the tree T (not just simple random walk) is a transition for which $P(x, y) > 0$ for all $(x, y) \in E$ and if $P(x, y) > 0$ then either $(x, y) \in E$ or $x = y$. We label the invariant measure of P by π. It can be shown that P is reversible with respect to π. This is again because we can fix some vertex o and then express $\pi(x)$ for any vertex x in terms of $\pi(o)$ using detailed balance equations (along the unique path between x and o).*

Definition 0.14 (Aperiodic chain)**.** A Markov chain is called aperiodic if for all $i \in S$ we have that the period of the state i, which is defined to be $\gcd\{t \geq 1 : P^t(i, i) > 0\}$ is equal to 1. The chain is called periodic if it is not aperiodic. //

An example of a periodic chain is a simple symmetric random walk on $\mathbb{Z}$ which is a Markov chain on a state space $\mathbb{Z}$ which at every step moves by $+1$ or -1 with equal probability. This is a periodic chain since if the chain starts from 0 at even times it can only be in $2\mathbb{Z}$ while at odd times it can only be in states in $2\mathbb{Z} + 1$ and clearly the period of any state is 2.

Definition 0.15 (Laziness)**.** For a transition matrix P we say that a *lazy* chain corresponding to P is a Markov chain $\widehat{X}$ with transition matrix

$$\widehat{P} = \frac{P + I}{2}$$

where I is the identity matrix on S. For $\alpha \in (0, 1)$ we say that the chain with transition $\alpha I + (1-\alpha)P$ is an *α-lazy chain* and the lazy chain is just α-lazy chain with parameter $\alpha = 1/2$. //

The reason lazy chains are useful is that they are a way to make an irreducible, periodic chain into an aperiodic chain. Any lazy chain is clearly aperiodic as it has a positive probability of staying in place. Any transition matrix with diagonal elements at least $1/2$ is a lazy version of some Markov chain. We can run a chain $\widehat{X}$ by tossing a fair coin repeatedly and if it comes up heads we make a transition according to P while otherwise we do nothing. We can notice that lazy chain $\widehat{X}$ is then just a time-changed version of X, where the chain $\widehat{X}$ makes a new step after geometric $1/2$ time. At this random time when the lazy chain makes the step we have that the distribution of the step is the same as the distribution of a single step of X. So we have a sequence of i.i.d. random variables T_i which are Geom$(1/2)$ such that the chain $\widehat{X}$ does not move between time $\sum_{i=1}^{m} T_i$ and $\sum_{i=1}^{m+1} T_i - 1$ and that the law of $(\widehat{X}_{T_n})_{n\geq 0}$ is the same as the law of $(X_n)_{n\geq 0}$. It is easy to check that the invariant distributions of a Markov chain and its lazy version are the same.

Definition 0.16 (Ergodicity)**.** A Markov chain is called *ergodic* if it is irreducible, positive recurrent and aperiodic. //

Notice that for an irreducible, positive recurrent chain the lazy version of it will be ergodic. As we will only work with chains on a finite state space which are also irreducible (so positive recurrent), we have that by considering the lazy version of a chain we get an ergodic chain. So we can think of laziness as being a tool for making all Markov chains we consider ergodic. [Note that another way of doing this would be looking at the continuous time chains (as they are always aperiodic), but in this course, we will only work with discrete time chains (however, many results would generalise to continuous time).]

The following result, which you probably covered in any undergraduate Markov chains course, is the reason we care about ergodic chains.

Theorem 0.17 (Ergodic theorem)**.** *Suppose that X is an ergodic Markov chain with invariant distribution π. Then for all $i, j \in S$ we have*

$$P^t(i,j) \to \pi(j) \quad \text{as} \quad t \to \infty.$$

We will see proofs of this theorem later in the course. The above theorem does not tell us anything about the speed or rate of convergence and in this course, we will want to study this rate. To do that we first have to define a metric on probability distributions. We do this in the next section.

Useful results for Mixing times of Markov chains

The purpose of this sheet is to recall some standard results from Markov Chain theory, which are useful to know. If you did any Markov chains course you have probably already seen most of these results. You are encouraged to try proving them again or read the proof. The final two questions state two important linear algebra results, which we will use a couple of times in this course. You are **not expected to know how to prove any of these results**, but you should know their statements. Proofs of standard Markov chains results can be found in any course notes, such as https://www.dpmms.cam.ac.uk/ rrw1/markov/M.pdf.

1. (*Strong Markov Property*). A radnom variable T taking values in $\{0,1,2,3,\ldots\}\cup\{\infty\}$ is called a *stopping time* if the event $\{T=t\}$ depends only on $X_0, X_1, \ldots, X_t$ for all $t \geq 0$. The Strong Markov property states that for a Markov chain X on S with transition matrix P, for any stopping time T and state $i \in S$ we have that conditional on $T < \infty$ and $X_T = i$ the process $(X_{T+t})_{t=0}^{\infty}$ is a Markov chain starting from i with transition matrix P which is independent of $X_0, X_1, \ldots X_T$. Prove the strong Markov property.

2. (*Alternative interpretation of aperiodicity*). Prove that for an irreducible, aperiodic transition matrix P on a finite state space S there exists $T \in \mathbb{N}$ such that $P^t(x,y) > 0$ for all $t \geq T$ and $x, y \in S$.

3. (*Recurrence is a class property*). Consider a chain with transition matrix P on a state space S. A state $x \in S$ is called recurrent if $\mathbb{P}_x(T_x^+ < \infty) = 1$ and otherwise it is called transient. Show that a state $x \in S$ is recurrent if and only if $\sum_{t=0}^{\infty} P^t(x,x) = \infty$. Moreover, prove that the recurrence of x is also equivalent to $\mathbb{P}_x(X_t = x \text{ for infinitely many } t) = 1$. Prove that for a transient state x, the total number of visits to the state x for the chain started from x is distributed as a geometric random variable. Show that for an irreducible chain we have that one state is recurrent if and only if all states are recurrent (i.e. recurrence is a class property).

4. (*Expected return time and π*). A state x is called positive recurrent if $\mathbb{E}_x[T_x^+] < \infty$. In this course we will need to know that the irreducible Markov chain on a finite state space has invariant distribution π, that it is positive recurrent and that $\mathbb{E}_x\left[\sum_{t=1}^{T_x^+} \mathbb{1}_{\{X_t=y\}}\right] = \frac{\pi(y)}{\pi(x)}$ and $\pi(x) = \frac{1}{\mathbb{E}_x[T_x^+]}$. This question leads you through the proof of these standard Markov chain results.

   (a) For an irreducible recurrent chain for all $x, y \in S$ let $\nu^x(y) = \mathbb{E}_x\left[\sum_{t=1}^{T_x^+} \mathbb{1}_{\{X_t=y\}}\right]$ be the expected number of visits to state y between two consecutive visits of x (notice that $\nu^x(x) = 1$). Show that $\nu = \nu P$ and that $0 < \nu^x(y) < \infty$.

   (b) Show for an irreducible chain that if a state $x \in S$ is positive recurrent then the chain has the invariant distribution, given by $\pi(y) = \frac{\nu^x(y)}{\mu(x)}$, where $\mu(x) = \mathbb{E}_x[T_i^+]$.

   (c) Show that for an irreducible, recurrent chain any invariant measure satisfying $\eta(x) = 1$ has to be equal to measure ν^x and conclude that if the chain has an invariant distribution then all states are positive recurrent and $\pi(x) = \frac{1}{\mu(x)}$. Conclude that $\nu^x(y) = \frac{\pi(y)}{\pi(x)}$.

   (d) Show that for an irreducible chain one state is positive recurrent if and only if all states are positive recurrent (i.e. positive recurrence is a class property).

   (e) Show (using linear algebra) that any irreducible chain on a finite state space has an invariant distribution and deduce that all states in a finite irreducible chain are positive recurrent.

The following are the only two linear algebra results we will need to recall in this course. The questions ask you to apply them to transition matrices and we will see some of these applications again in the course.

5. (*Spectral theorem applied to Markov chains*). Recall that the spectral theorem for symmetric matrices says that a finite symmetric real matrix has an orthonormal basis of eigenvectors and that all of its eigenvalues are real. By considering $A(x,y) = \sqrt{\frac{\pi(x)}{\pi(y)}}P(x,y)$ show that a transition matrix P on S which is reversible with respect to π has an orthonormal basis of eigenfunctions (i.e. eigenvectors f satisfying $\sum_{y\in S} P(x,y)f(y) = \lambda f(x)$ for some λ and all $x \in S$) corresponding to the inner product $\langle f, g\rangle_\pi = \sum_{x\in S} f(x)g(x)\pi(x)$.

6. (*Perron-Frobenius and application to Markov chains*). The Perron-Frobenius theorem states the following. Suppose A is a square matrix, which is non-negative and irreducible, i.e. satisfies that for all i, j, we have $(A^t)_{ij} > 0$ for some $t \in \mathbb{N}$. Then there exists a positive real number, λ, such that:

   - λ is an eigenvalue of A with multiplicity 1,
   - both the left and right-hand eigenvectors corresponding to λ are strictly positive,
   - no other eigenvector of A is strictly positive,
   - all eigenvalues μ of A satisfy $|\mu| \leq |\lambda|$,
   - if, moreover, A is strictly positive then all other eigenvalues μ of A satisfy $|\mu| < |\lambda|$
   - and $\min_i \sum_j a_{ij} \leq \lambda \leq \max_i \sum_j a_{ij}$, or in other words, λ lies between the minimum and maximum of row sums of A.

   Show that in the case that $A = P$ is the transition matrix of an irreducible Markov chain, then $\lambda = 1$ is an eigenvalue of multiplicity 1 and $(1, 1, \ldots, 1)$ is the corresponding unique positive eigenfunction. Moreover, use the above to show that if P is aperiodic all other eigenvalues have a modulus strictly less than 1. Check directly without using the above theorem that any transition matrix has all eigenvalues of modulus at most 1. Justify using the above why an invariant distribution π exists for all finite chains. Show also that π is the uniform distribution on the state space if and only if the transpose of P is stochastic (or in other words if the sum of incoming probability to each state is 1, i.e. $\sum_i P(i,j) = 1$ for all j).

1 Introduction

Lecture 1

1.1 Brief motivation and the definition of mixing time

We start by giving a very brief example of a question the area of mixing times attempts to answer.

Example 1.1 (Shuffling cards)**.** *Suppose you have a deck of $n = 52$ cards which are ordered from 1 to n which you would like to shuffle well. Assume you are only allowed to do one of the following two shuffling methods:*

- *Riffle shuffle : we repeat the following two-step procedure:*

  (i) cut the deck into two (possibly unequal) parts;

  (ii) interleave cards from the two parts to produce a new deck.

  Formally, identify each card with a unique label $i \in [n] = \{1, \dots, n\}$ and represent a deck of cards by a permutation $\sigma \in S_n$, where $\sigma(i)$ indicates the label of the i-th top card in the deck. Then, the above procedure transforms σ into a new permutation of the form $\sigma' = \sigma \circ \gamma_I$, where the set $I \subseteq [n]$ indicates the positions to which the "top part" gets relocated, and where γ_I is the permutation that takes values $1, 2, \dots, |I|$ (in this order) on I and $|I| + 1, \dots, n$ (in this order) on $[n] \setminus I$. We choose the subset $I \subseteq [n]$ at random according to some prescribed distribution (e.g., uniform) and repeat this procedure independently at each step. The most common definition of riffle shuffle will pick I to be a uniform subset, i.e. we can choose elements of I by adding each $i \in [n]$ to I with probability $1/2$, independently of the other elements.

- *Random-To-Top shuffle: at each step, take a randomly chosen card from the deck and insert it at the top of the deck (we are allowed to choose a top card as a random card and place it on top, and in this case, the order of cards just stays the same). The operation of taking the card in position i of the deck and moving it to the top of the deck can be described mathematically as multiplying on the right the current permutation σ by the inverse of the cycle $(1\,2\,\dots\,i)$.*

Which method would you use? How many shuffles do you need to shuffle the deck well for each of these methods?

Intuitively, the riffle shuffle should be shuffling cards better, simply as it simultaneously changes the relative order of many cards. But to formally answer which method shuffles the cards better, we first need to define what it means to shuffle them well. The definition we want to use might change depending on what we want to use the cards for. For example, if we are going to play a game where after one shuffle you ask the opponent to guess the top card and you win if they do not guess it, you would want to use the random-to-top shuffle as the top card after one such shuffle is a uniform card from the deck, while in the riffle shuffle, there is a probability $1/2$ that the top card is card 1. However, if you want to play any reasonable cards game, it is unlikely that the only important feature of the deck is going to be the value of the top card. For example, you can notice that after one random-to-top shuffle, the probability that card $i > 1$ is either in position i or $i+1$ is going to be $\frac{n-1}{n}$ while for the riffle shuffle, this probability is going to be a lot smaller. In a uniform deck for a fixed i this probability is $\frac{2}{n}$, so if want the deck to look like a uniform permutation of cards, maybe it would be more efficient to use the riffle shuffle.

From the Ergodic theorem (Theorem 0.17) it is not hard to see that for both shuffling methods, as the number of shuffles goes to infinity, the distribution of the deck will converge to the uniform permutation (as it is not hard to check that the invariant distribution for both Markov chains in uniform). To figure out how many shuffles are needed before we shuffle the deck well, i.e. before the deck is distributed close to a uniform distribution, we need to work out how the distribution of cards looks after some fixed number of shuffles, but more importantly, we need to define what we mean when we say that the two distributions are close to each other. For this, we need to define a distance on probability distributions. In the context of mixing times, the most widely used distance is total variation distance, which we define formally soon. Intuitively, the idea behind it is the following. We want to say that the two distributions are close if it is hard to distinguish between them. To distinguish between the distributions, we can come up with some test questions, such as "What is the probability that card i is before j in the deck?" and we would say that the two distributions are close if the answers to all possible test questions for these two distributions do not have a large difference. If after t shuffles all possible test questions have roughly the same probabilities as in the uniform arrangement of cards, it would be natural to say that the deck is shuffled well, i.e. close to uniform, as it would not be easy to distinguish our deck from a uniform deck.

Formally, we define total variation distance as follows.

Definition 1.2 (Total variation distance)**.** Let S be a finite space and let μ and ν be two probability distributions on S. We define the total variation distance or the TV distance of μ and ν to be

$$\|\mu - \nu\|_{\text{TV}} = \max_{A \subseteq S} |\mu(A) - \nu(A)| = \max_{A \subseteq S} (\mu(A) - \nu(A)),$$

where we see that the two maximums are equal as $\mu(A) - \nu(A) = (1 - \mu(A^c)) - (1 - \nu(A^c)) = \nu(A^c) - \mu(A^c)$. Total variation distance is also sometimes denoted by $d_{\text{TV}}(\mu, \nu) = \|\mu - \nu\|_{\text{TV}}$ //

We can use the example with card shufflings to get a better idea of what a total variation distance actually means. Let's look at two decks of cards, one is disstributed according to the uniform permutation (one can notice that invariant distribution for both riffle and random-to-top shuffling is a uniform distribution, so as time goes to infinity the distribution of the deck will converge to uniform) and one is distributed according to the distribution after k shuffles. The total variation distance is by definition the maximum over all subsets of permutations of the difference of probability that uniform permutation is in that subset and that the permutation after k shuffles is in that set. We mentioned "test questions", such as "What is the probability that i is before j?" and that we would want these to have similar answers for the uniform deck and deck after k shuffles. We can notice that this "test question" is just taking the subset of permutations for which i is before j and asking about the probabilities of uniform permutation vs the permutation after k shuffles being in this set. In fact any "test question" you could think of is exactly asking what the probability of permutation after k shuffles being in the set of all permutations with some property is and how that compares with the answer for the uniform permutation. So by the definition, the total variation distance of uniform permutation and the one after k shuffles is just upper bounding the difference of the answers to test questions for these two decks of cards. Moreover, we know that there is a test question for which the TV distance is exactly the difference of these probabilities (this test question is simply "What is the probability that the permutation is in set A?" where A is the set achieving the maximum in the TV distance definition). So it would make sense to call the deck well shuffled if the TV distance of the distribution of the uniform permutation and the permutation after k shuffles differ by at most some small constant we pick. This should motivate the following definition of the distance to equilibrium and mixing time and convince us that the total variation distance is a reasonable metric between probability distributions.

Formally, let X be a Markov chain with transition matrix P and invariant distribution π. We define the *distance to equilibrium* (also known as *distance to stationarity*)

$$d_P(t) = \max_{x \in S} \|P^t(x, \cdot) - \pi\|_{\mathrm{TV}}.$$

We further define the *ε-mixing time* of X to be the first time t such that the total variation distance at that time drops below ε i.e.

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

The *mixing time* is then defined to be the ε-mixing time for $\varepsilon = \frac{1}{4}$, i.e.

$$t_{\mathrm{mix}} = t_{\mathrm{mix}}\left(\frac{1}{4}\right).$$

For card shufflings, one could compute that the mixing time of the random-to-top shuffle requires hundreds of shuffles before being mixed, while for the riffle shuffle, the mixing time is just 11. So if you want to get a deck of cards which is distributed close to a uniform deck, you can just take any deck of cards and apply 11 riffle shuffles to it. Similarly, mixing times can be used to sample from complicated distributions on large state spaces. A typical example for this is the so-called Ising model, where each vertex in a graph gets a spin of $+1$ or -1, randomly, and the probability to get some specific configuration of spins is given by some function which depends on the number of neighbours which have different spins. If one wants to get a sample from the Ising model on a say 100×100 grid, they would need to evaluate probabilities of all 2^{10000} configurations, which would be impossible to do using any supercomputer. However, alternatively, one can start with an arbitrary assignment of $+1$s and -1s and run a Markov chain which has an invariant distribution corresponding to this Ising model. At the mixing time of this chain, the configuration of spins will be distributed close to the distribution of an Ising model. Luckily, it turns out that one can construct such a Markov chain (for which one step of the chain is not difficult to sample) and bound its mixing time by a function which has significantly smaller order than the number of states of that chain. A similar approach can be used to get a "close to uniform" proper vertex colouring of a graph (proper colouring just means that no two neighbours have the same colour) and even to get an algorithm which, with large probability, estimates the total number of proper colourings for any graph, in a reasonable time.

When studying mixing times of Markov chains, it is usually not particularly interesting to look at the mixing time of one fixed chain (say, shuffling of 52 cards or Ising model on a 100×100 grid). Usually, one wants to consider a sequence of Markov chains indexed in some n, and see how the mixing times behave roughly as a function of n. The sequence usually has a state space which increases with n, such as shuffling of n cards, which has state space of size $n!$ or a random walk on a cycle of length n, which has state space of size n. For example, we will show that for random-to-top shuffling, the mixing time is roughly $n \log n$, while one can show that for riffle shuffle it is $\frac{3}{2} \log n$.

1.2 Total Variation and Coupling

This definition of a total variation distance is a probabilistic definition, as the distance between μ and ν is given in terms of the probabilities assigned to events A. We can also notice that the total variation distance is always at most 1 and that it is equal to 1 if and only if μ and ν have disjoint support. It is clear that the total variation distance is symmetric and satisfies that

$\|\mu(x) - \nu(x)\|_{\text{TV}} \geq 0$ with equality iff $\mu = \nu$ (as if they differ on a state i by considering $A = \{i\}$ we get that TV distance is > 0) and therefore to show it is a metric we only need to check the triangle inequality. We get the triangle inequality easily from the following Proposition which gives alternative expressions for the total variation distance. These expressions are particularly useful to remember as they will be used in this course. The proof of this proposition is just a sequence of algebraic manipulations and the main idea is considering the set $B = \{x \in S : \mu(x) \geq \nu(x)\}$.

Proposition 1.3 (Alternative expressions for TV distance and TV distance is a metric)**.** *Let μ and ν be two probability distributions on S. Then*

$$\begin{aligned}\|\mu - \nu\|_{\text{TV}} &= \sum_{x\in S}(\mu(x) - \nu(x))^+ = 1 - \sum_{x\in S}\mu(x) \wedge \nu(x) \\ &= \frac{1}{2}\sum_{x\in S}|\mu(x) - \nu(x)| = \frac{1}{2}\sup_{f:S\to[-1,1]}|\mu f - \nu f|\end{aligned}$$

where $a\wedge b = \min\{a, b\}$ denotes the minimum and $(a-b)^+ = \max\{a-b, 0\}$ and $\mu f = \sum_{x\in S}\mu(x)f(x)$. Moreover, from the penultimate expression, it is easy to see that the total variation distance satisfies triangle inequality and is a metric.

The main idea of the proof is to consider a set $B = \{x \in S : \mu(x) \geq \nu(x)\}$, show that $\|\mu - \nu\|_{\text{TV}} = \mu(B) - \nu(B)$ and then do some simple algebraic manipulations.

Proof. Let $B = \{x \in S : \mu(x) \geq \nu(x)\}$ and let $A \subseteq S$ 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

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

again using the definition of B. This proves that

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

and taking $A = B$ we get

$$\max_{A\subseteq S}\mu(A) - \nu(A) = \mu(B) - \nu(B) = \sum_{x\in S:\mu(x)\geq\nu(x)}(\mu(x) - \nu(x)) = \sum_{x\in S}(\mu(x) - \nu(x))^+.$$

Now as $(a - b)^+ = a - a \wedge b$ we have

$$\sum_{x\in S}(\mu(x) - \nu(x))^+ = \sum_{x\in S}\mu(x) - \sum_{x\in S}\mu(x) \wedge \nu(x) = 1 - \sum_{x\in S}\mu(x) \wedge \nu(x)$$

as μ is a probability distribution on S.

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

$$\begin{aligned}\|\mu - \nu\|_{\text{TV}} &= \frac{1}{2}(\mu(B) - \nu(B) + \nu(B^c) - \mu(B^c)) \\ &= \frac{1}{2}\left(\sum_{x\in B}|\mu(x) - \nu(x)| + \sum_{x\in B^c}|\nu(x) - \mu(x)|\right) = \frac{1}{2}\sum_{x\in S}|\mu(x) - \nu(x)|\end{aligned}$$

$$|\mu f - \nu f| = |\sum_{x\in S}(\mu(x)-\nu(x))f(x)| \le \sum_{x\in S}|\mu(x)-\nu(x)||f(x)| \le \sum_{x\in S}|\mu(x)-\nu(x)|$$

and that the above inequalities are equalities for a function $f(x) = \text{sign}(\mu(x) - \nu(x))$ completes the proof. □

Proposition 1.4 (Convexity and contraction)**.** *The function $(\mu, \nu) \mapsto \|\mu - \nu\|_{\text{TV}}$ is convex on the set of all pairs of distributions on S. Also, any transition matrix P on S contracts TV distance, i.e. for all distributions on S, μ and ν we have*

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

For convexity check the definition, for the inequality use the alternative expression of TV distance with functions.

Proof. The first claim follows from the observation that $(\mu, \nu) \mapsto \mu(A) - \nu(A)$ is trivially convex for each $A \subseteq S$, and that any maximum of convex functions is convex. Alternatively, we can just check for any distributions on S $\mu_1, \nu_1, \mu_2, \nu_2$ and $p \in [0, 1]$ that

$$\begin{aligned}\|(p\mu_1 + (1-p)\mu_2) - (p\nu_1 + (1-p)\nu_2)\|_{\text{TV}} &= \sup_{A\subset S}(p(\mu_1(A)-\nu_1(A)) + (1-p)(\mu_2(A)-\nu_2(A)))\\ &\le p\sup_{A\subset S}\mu_1(A)-\nu_1(A) + (1-p)\sup_{A\subset S}\mu_2(A)-\nu_2(A)\\ &= p\|\mu_1-\nu_1\|_{\text{TV}} + (1-p)\|\mu_2-\nu_2\|_{\text{TV}}\end{aligned}$$

The contraction follows from the last expression in Proposition 1.3, as

$$\|\mu P - \nu P\|_{\text{TV}} = \frac{1}{2}\sup_{f:S\to[-1,1]}|\mu Pf - \nu Pf| = \frac{1}{2}|\mu g - \nu g| \le \|\mu - \nu\|_{\text{TV}}$$

where $g = Pf^*$ and f^* is the function attaining the supremum above, and the last inequality holds as $g : S \to [-1, 1]$ as P is stochastic and $f^* : S \to [-1, 1]$ [i.e. we can check $|g(x)| = |\sum_{y\in S} P(x, y)f^*(y)| \le \sum_{y\in S} P(x, y)|f^*(y)| \le \sum_{y\in S} P(x, y) = 1$]. □

Example 1.5 (Total variation distance for Bernoulli)**.** *It is easy to see that a total variation distance of μ and ν which are the laws of the two Bernoulli random variables with parameters p and q respectively is $|p - q|$. The two distributions are supported on two states, we can call them states 1 and 0. If A is $\{0, 1\}$ or an empty set clearly $\mu(A) - \nu(A) = 0$ and $|\mu(1) - \nu(1)| = |p - q|$ while $|\mu(0) - \nu(0)| = |(1 - p) - (1 - q)| = |p - q| \ge 0$.*

We now give one more way of bounding the total variation distance and that is through *coupling*. Coupling is a general technique in probability where one tries to compare two distributions by considering some random variables distributed according to them. Formally we define it as follows:

Definition 1.6 (Coupling)**.** A *coupling* of two probability measures μ, ν is a pair (X, Y) of random variables defined on the same probability space, such that $X \sim \mu$ and $Y \sim \nu$. //

An example of coupling would be to take X and Y to be independent, or if $\mu = \nu$ to take them to be the same random variable. Of course, this is not a very useful coupling and we will usually want to couple μ and ν which are not the same in some way that helps us reveal more about the pair

(μ, ν). The above definition lets us pick a way in which X and Y depend on each other however we want as long as their marginals are correct.

For example, a coupling of two Bernoulli distributions with parameters p (i.e. distribution μ such that $\mu(1) = p$ and $\mu(0) = 1 - p$) and q, with $p > q$ could be one where we pick the joint law of X and Y such that

- $X = Y = 1$ with probability q,
- $X = 1, Y = 0$ with probability $p - q$
- $X = Y = 0$, otherwise.

We can see that the marginals of X and Y are μ and ν, so this is a valid coupling. This coupling is somewhat interesting because we coupled (μ, ν) in such a way that $X \geq Y$ always. A good way to think of this coupling would be using a uniform random variable in $[0, 1]$, U and setting $X = \mathbb{1}_{\{U \leq p\}}$ and $Y = \mathbb{1}_{\{U \leq q\}}$.

We can notice that this coupling of Bernoulli random variables has $\mathbb{P}(X \neq Y) = p - q$ and recall that we have also checked that this is the total variation distance of laws of these Bernoulli distributions. It is not hard to see that for any coupling (X, Y) of any distributions their total variation distance is upper bounded by $\mathbb{P}(X \neq Y)$. It turns out that, in general, for any two distributions, there is a coupling (X, Y), known as an optimal coupling, for which $\mathbb{P}(X \neq Y)$ is exactly the total variation distance. We now prove this fact.

Theorem 1.7 (TV distance is given by an optimal coupling)**.** *Let μ and ν be two probability distributions on S. 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 an optimal coupling of μ and ν.

Just try to set X and Y to be equal as often as possible. This can be done with probability $p = \sum_{x \in S} \mu(x) \wedge \nu(x)$ and it is illustrated in the picture from the proof. If we pick a uniform point under the curve of μ we can notice that if the point lands in the region III one can set $X = Y$ and if it lands in the region I we pick X from region I and Y from II. The result then follows from one of the alternative expressions for TV distance.

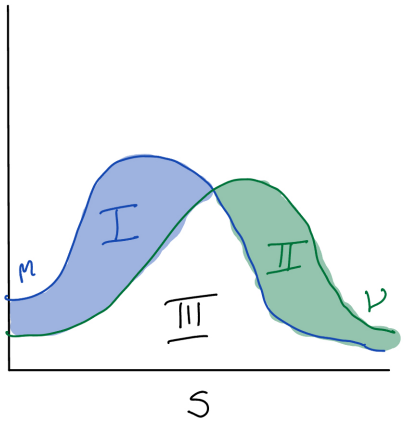
Proof. If (X, Y) is a coupling of μ and ν, then for any set $A \subseteq S$ we can write

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

By taking a maximum over $A \subseteq S$ we upper bound the TV distance by $\mathbb{P}(X \neq Y)$.

Conversely, let us construct an optimal coupling (X, Y). To minimise $\mathbb{P}(X \neq Y)$ we clearly want X and Y to be equal as often as possible. We essentially do it as follows: we consider the following sketch of probability distributions μ and ν (x axis corresponds to the elements in S and y axis to the probabilities, so the area under the curves μ and ν should be 1 and they should rather look like step functions) and imagine we throw a point in the regions $I \cup III$. If the point lands in the region

III, then we set $X = Y$. Otherwise, we throw X in the region I and Y in II. In this way, they are equal only if the initial point lands in III.



More formally, let $p = \sum_{x \in S} \mu(x) \wedge \nu(x)$. From the second expression in Proposition 1.3 we know that if $\mathbb{P}(X = Y) = p$ then $\mathbb{P}(X \neq Y) = 1 - p = \|\mu - \nu\|_{\text{TV}}$ so we would be done. We toss a coin with a probability of heads equal to p. If the coin comes up heads, then we sample a random variable 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

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

and we sample Y according to

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

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

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

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

Also from the construction, since γ_I and γ_{II} are supported on disjoint sets, it follows that $X = Y$ iff the coin comes up heads which has probability p and we are done. □

We remark that optimal couplings are not necessarily unique, if we have two distributions with disjoint supports such that support of each distribution has at least 2 elements (this is needed so that we know that there is more than one coupling), then any coupling is optimal as the total variation distance is 1.

The above representation of total variation distance is particularly useful as it involves a minimum and so any coupling we would find would give an upper bound.

1.3 Distance to stationarity and mixing time

We recall the definition of the mixing time first.

Definition 1.8 (Distance to equilibrium/stationarity)**.** Let X be a Markov chain with transition matrix P and invariant distribution π. We define the *distance to equilibrium* (also known as *distance to stationarity*)

$$d_P(t) = \max_{x \in S} \|P^t(x, \cdot) - \pi\|_{\mathrm{TV}}.$$

We also define $\bar{d}(t)$, i.e. *"distance from worst starting points"* as

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

When it is clear which chain we are referring to we will drop P from the notation and only write $d(t)$ and $\bar{d}(t)$. //

From the Proposition 1.4 we get directly using the contraction property that $\bar{d}_P(t)$ *is decreasing as a function of* t and also using that $\pi P = \pi$ that $d_P(t)$ *is also decreasing in* t and the convexity implies that for any starting distribution ν on S we have $d_P(t) \geq \|\nu P^t - \pi\|_{\mathrm{TV}}$.

Definition 1.9 (Mixing time)**.** For a Markov chain X with a transition matrix P we define the *ε-mixing time* of X to be the first time t such that the total variation distance at that time drops below ε i.e.

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

The *mixing time* is then defined to be the ε-mixing time for $\varepsilon = \frac{1}{4}$, i.e.

$$t_{\mathrm{mix}} = t_{\mathrm{mix}}\left(\frac{1}{4}\right)$$

//

We start by giving a proof of the Ergodic Theorem (Theorem 0.17) in the case of finite Markov chains. In fact for finite Markov chains, we get the following stronger results.

Theorem 1.10 (Geometric decay of distance to equilibrium)**.** *Let X be an irreducible and aperiodic Markov chain on a finite state space with invariant distribution π. Then there exist $a \in (0, 1)$ and a positive constant C so that for all t we have*

$$\max_{x \in S} \|P^t(x, \cdot) - \pi\|_{\mathrm{TV}} \leq Ca^t.$$

The main idea is that aperiodicity implies $\exists r, \alpha > 0$ so that $P^r(x, \cdot) = \alpha\pi + (1 - \alpha)Q(x, \cdot)$ for some transition matrix Q. This implies that $P^{kr+\ell}$ can be generated by taking a sample from π with probability $1 - (1 - \alpha)^k$ and otherwise a sample from $Q^k P^\ell$ and this implies wanted bound.

Proof. By irreducibility and aperiodicity from "Useful results for Mixing times of Markov chains", Question 2 there exists $r > 0$ such that the matrix P^r has strictly positive entries. We now set $\alpha := \min_{x,y \in S} \frac{P^r(x,y)}{\pi(y)}$. As $\frac{P^r(x,y)}{\pi(y)} > 0$ for all x, y and the state space is finite this minimum is attained at some x, y and so $\alpha > 0$. Moreover as $\sum_{y \in S} P^t(x, y) = 1 = \sum_{y \in S} \pi(y)$ for all x, we also have that $\alpha \leq 1$. If $\alpha = 1$, then $d(r) = 0$ as $P^r(x, \cdot) = \pi$ for all $x \in S$ and we are done (by taking any $a < 1$ and $C = \frac{1}{a^r}$), so assume $\alpha < 1$. This means that

$$P^r(x, y) \geq \alpha\pi(y) \text{ and so we can write } P^r(x, y) = \alpha\pi(y) + (1 - \alpha)Q(x, y)$$

where it is easy to check that $Q(x,y) := \frac{P^r(x,y)-\alpha\pi(y)}{1-\alpha}$ is a stochastic matrix (this is why we needed to take α to be the infimum over x, y). We can interpret this probabilistically as sampling the value of X_r when starting from x, by tossing a coin with probability α of heads (H) and if H falls we sample it according to the distribution π while for tails we sample according to Q. As $\pi P = \pi$ this means that to sample $P^{kr}(x,\cdot)$ we toss independent coins k times, and as soon as the first H falls, every next step will be distributed according to π, so if any H appears we just sample π, while if no H fell we sample it according to $Q^k(x,\cdot)$. As the probability of seeing only tails is $(1-\alpha)^k$ we have

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

where Π is a matrix with rows equal to π i.e. $\Pi(x,y) = \pi(y)$. Using again that π is invariant we have that $\Pi P = \Pi$ and so for all ℓ we have

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

We see now that for any $x \in S$

$$\|P^{rk+\ell}(x,\cdot) - \pi\|_{\mathrm{TV}} = (1-\alpha)^k \|Q^k P^\ell(x,\cdot) - \pi\|_{\mathrm{TV}} \leq (1-\alpha)^k$$

which concludes the proof by taking $a = (1-\alpha)^{\frac{1}{r}}$ and $C = \frac{1}{1-\alpha}$.

[Notice that it would have been enough to prove that $d(kr) \leq a^{kr}$ for integers k and then use that as $d(t)$ is non-increasing and so for $t = kr+\ell$, and $\ell \leq r$ we would have $d(kr+\ell) \leq a^{kr} \leq a^t \frac{1}{a^r}$.] □

Remark 1.11. *We showed that for all finite aperiodic Markov chains the distance to stationarity decays eventually at least as fast as geometric decay, but we have no control over the constant C in the above inequality and over the value r from the proof. This does not tell us in fact how the distance to stationarity decays initially, i.e. for smaller values of t and does not tell us anything about the value of the $t_{\mathrm{mix}}(\varepsilon)$ for any given chain and value of ε. It mostly tells us that the definition of mixing times makes sense for finite ergodic chains as it indeed happens that eventually, the distance to stationarity goes to* 0.

The main purpose of the following sequence of lemmas is to help us establish useful inequalities relating mixing times for different values of ε.

Lemma 1.12 (Comparison of d and $\bar{d}$)**.** *For all t we have that $d(t) \leq \bar{d}(t) \leq 2d(t)$.*

As $\pi P^t = \pi$ averaging over y with weights $\pi(y)$ expression $|P^t(x,A) - P^t(y,A)|$ and maximising over A gives an upper bound $d(t) \leq \bar{d}(t)$ and triangle inequality gives directly $\bar{d}(t) \leq 2d(t)$.

Proof. The second inequality is immediately from the triangle inequality for the TV distance which we know from Proposition 1.3 (consider $\|P^t(x,\cdot) - P^t(y,\cdot)\|_{\mathrm{TV}} \leq \|P^t(x,\cdot) - \pi\|_{\mathrm{TV}} + \|\pi - P^t(y,\cdot)\|_{\mathrm{TV}}$, maximise LHS over x, y and then bound the RHS by $2d(t)$). To prove the first inequality, we use that for any set A, by stationarity $\pi(A) = \sum_{y\in S} \pi(y) P^t(y,A)$. Therefore, for all x

$$\begin{aligned}\|P^t(x,\cdot) - \pi\|_{\mathrm{TV}} &= \max_{A\subset S} \left|P^t(x,A) - \pi(A)\right| = \max_{A\subset S} \left|\sum_{y\in S} \pi(y)(P^t(x,A) - P^t(y,A))\right| \\ &\leq \sum_{y\in S} \pi(y) \max_{A\in S} |P^t(x,A) - P^t(y,A)| \leq \bar{d}(t),\end{aligned}$$

where we have used that $\sum_{y\in S} \pi(y) = 1$. □

In the above proof we just spotted that starting from π means "averaging" with weights according to π over the starting points y, so the "worst starting y" has to be "further away" from x than the "average starting y". We are actually just exploiting the convexity of the TV distance and the fact that $\sum_{y\in S}\pi(y)P^t(y,\cdot)=\pi(\cdot)$ (so we can weight $P^t(y,\cdot)$ with $\pi(y)$ for each y).

The following lemma says that if we ran two chains from the "worst starting points" for time $t+s$ their distance (the TV distance of their laws at time $t+s$) would be smaller than the product of the TV distances in case we would run them from the "worst starting points" for time s and then restart them from the "worst starting points" and run them for time t.

Lemma 1.13 ("Sub-multiplicativity" of $\bar{d}$)**.** *For all $s,t\geq 0$ we have that $\bar{d}(s+t)\leq\bar{d}(s)\bar{d}(t)$. We also have that $d(s+t)\leq d(s)\bar{d}(t)$.*

Idea is to use the existence of an optimal coupling for $P^s(x,\cdot)$ and $P^s(y,\cdot)$, which tells us that there is a way to sample laws of chains starting from x and y such that they are not the same at time s with probability $\|P^s(x,\cdot)-P^s(y,\cdot)\|_{\mathrm{TV}}$. Starting from there if the chains have already collided at time s the TV distance of the laws after another t steps will be 0 while if they have not collided this distance will be at most $\bar{d}(t)$. We then maximise over $x,y\in S$.

Proof. Fix x and y and let (X,Y) be an optimal coupling (which exists by Theorem 1.7) of $P^s(x,\cdot)$ and $P^s(y,\cdot)$, i.e. a coupling for which

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

By the definition of total variation we have

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

But by the Markov property, we also have

$$P^{s+t}(x,z)=\sum_{u\in S}P^s(x,u)P^t(u,z)=\sum_{u\in S}\mathbb{P}(X=u)\,P^t(u,z)=\mathbb{E}\big[P^t(X,z)\big]$$

and similarly that

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

Substituting this above gives

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

Maximising over x and y completes the proof of sub-multiplicativity of $\bar{d}$. To show $d(s+t)\leq d(s)\bar{d}(t)$ we just amend the above proof by setting (X,Y) to be an optimal coupling of $P^s(x,\cdot)$ and $\pi(\cdot)$ and use that $P^t(Y,\cdot)$ is then distributed according to π. Figuring out the details is left as an exercise. □

Corollary 1.14 (Sub-multiplicativity of $2d$)**.** *For all $s, t \geq 0$ we have that $d(s+t) \leq 2d(s)d(t)$. *[1] Moreover, this means that the $(2d(t))^{\frac{1}{t}}$ has a limit as $t \to \infty$ and this limit is also strictly less than 1 when the chain is ergodic, as in that case we eventually have $d(t) < 1/2$.*

Proof. The proof of the sub-multiplicativity follows directly from Lemma 1.12 and 1.13. *The existence of the limit follows by considering $\log(2d(t))/t$ and directly applying Fekete's lemma, which states that for any subadditive sequence a_n the sequence a_n/n converges to its infimum. □

The above Corollary tells us that we essentially know that $d(t)$ will decay exponentially starting from some arbitrarily large time t, but this decay will usually be noticeable only for times which are a lot larger than the mixing time and therefore this doesn't necessarily help us that much with determining the ε-mixing times. However, we have the following useful consequence of the above results.

Proposition 1.15 (Upperbound on ε-mixing time in terms of the mixing time)**.** *For any ergodic Markov chain, we have*

$$t_{\text{mix}}(\varepsilon) \leq \left\lceil \log_2 \frac{1}{\varepsilon} \right\rceil t_{\text{mix}}$$

Bound d by $\bar{d}$ and apply sub-multiplicativity of $\bar{d}$.

Proof. In fact we have for any ℓ the following useful inequality just by directly applying inductively Lemma 1.13

$$d(\ell t_{\text{mix}}(\varepsilon)) \leq \bar{d}(\ell t_{\text{mix}}(\varepsilon)) \leq \bar{d}(t_{\text{mix}}(\varepsilon))^{\ell} \leq (2\varepsilon)^{\ell}.$$

By picking $\varepsilon = 1/4$ this gives

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

One can actually slightly improve the bound on $d(\ell t_{\text{mix}}(\varepsilon))$ using sub-multiplicativity of $2d$ and get

$$d(\ell t_{\text{mix}}(\varepsilon)) \leq \varepsilon(2\varepsilon)^{l-1}.$$

□

If a finite, irreducible chain is not aperiodic (and therefore not ergodic), then $t_{\text{mix}}(1/4)$ would be ∞, and the above result would still be true but would not be useful. We now explain how one could show that $t_{\text{mix}}(1/4) = \infty$. Let's say that the chain is periodic and that one state has a period $p > 1$. It can be checked now that (as the chain is irreducible) all states have the same period and that the state space S can be decomposed into sets $C_1, \dots C_p$ which are such that $P(x, y) > 0$ if and only if $x \in C_k$ and $y \in C_{k+1}$ (where $k+1$ is taken modulo p). This decomposition is found by taking any state x and setting C_k to be the states which can be reached from x after times of form $p\ell + k$. Now as $p \geq 2$ we have that for some k has $\pi(C_k) \leq \frac{1}{2}$ and one can check that for $x \in C_k$ we have that $|\mathbb{P}_x(X_{p\ell} \in C_k) - \pi(C_k)| = 1 - \pi(C_k) \geq \frac{1}{2}$ for all $\ell > 0$ which implies (as $d(t)$ is decreasing) that $\forall t\, d(t) \geq \frac{1}{2}$ and so for all $\varepsilon < \frac{1}{2}$ we have $t_{\text{mix}}(\varepsilon) = \infty$.

We will now study some techniques which give us a way to bound the mixing times. Usually, we will not want to determine the exact value of the mixing time for one given finite chain (this might be a

[1] Statement and proof of convergence using Fekete's lemma is not examinable. We denote by * non examinable sections of the notes.

hard question if the chain is large but one could in principle write a computer program which just computes a total variation distance to stationarity after any time for any given chain). Instead, the setup will usually be like the following: We will have a Markov chain depending on a parameter n. This parameter n might be the number of states, e.g. we can consider a random walk on a complete graph on n vertices, or a random walk on a line of length n, but it doesn't have to be, e.g. in the shufflings of n cards the state space is of size $n!$. Usually, the size of the state space will increase as n increases, or at least as $n \to \infty$ the size of the state space will also go to infinity. We will then want to know something about the behaviour of the mixing time of this sequence of chains with parameter n. Usually, it is hard to determine the exact value of the mixing time of the nth chain as a function of n. If we can not determine this exact value (which will almost always be the case), we at least hope to be able to find its order and, in some cases, we can even hope to determine the constant of the leading order term in the expression for the mixing time of the nth chain (the mixing time will usually also increase to infinity as $n \to \infty$). We will see many techniques which allow us to have tight upper and lower bounds on the mixing times and we will see some examples of this soon. We will start the following chapter with a method that gives an upper bound on the mixing time for many different chains.

2 Markovian coupling

We have seen the probability that the two random variables in the coupling disagree upper bounds the total variation distance and that, in principle, that technique has no limits as there is always an optimal coupling, so if we know how to pick an optimal coupling or at least some which is close to optimal we would get a sharp upper bound on the total variation distance. When studying mixing times the goal would be to control the total variation distance of $P^t(x,\cdot)$ and π, or as $d(t) \leq \bar{d}(t) \leq 2d(t)$ it would also be useful if we could control the TV distance of $P^t(x,\cdot)$ and $P^t(y,\cdot)$ for any $x, y \in S$. We want to use couplings for this. We know that any coupling of $P^t(x,\cdot)$ and $P^t(y,\cdot)$ gives an upper bound on the TV distance, but so far we have not seen any way in which we can construct couplings of the tth step of Markov chains. The issue with constructing coupling of $P^t(x,\cdot)$ and $P^t(y,\cdot)$ is that we do not understand these distributions well (if we did we would probably be able to bound the TV distance directly without the need for coupling). However, for any Markov chain we want to study we know the exact distribution of one step of that chain (starting from anywhere). In this section, we see a type of coupling, which is useful for us, as it gives us a way to couple two Markov chains at times t by coupling them one step at a time.

Definition 2.1 (Markovian coupling)**.** A coupling of Markov chains with transition matrix P is a process $(X_t, Y_t)_{t\geq 0}$ 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, $(X_t, Y_t)_{t\geq 0}$, which is also itself a Markov chain and which satisfies that for all x, x', y, y'

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

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 to become coalescent without changing the law of the first time the Markov chains meet (coalesce). Simply run the Markov chains using their Markovian coupling until they meet for the first time and then continue them together.*

More precisely, the above remark holds because of the following. We know that the vector process $(X_t, Y_t)_{t\geq 0}$ is a Markov chain, so we can run it step by step and its given by some transition matrix. Moreover, no matter what its transition matrix was from states of form (z, z) we can just change it to have a transition matrix sending (z, z) to (u, u) with probability $P(z, u)$ without affecting the part of the process which did not go through any state of the form (z, z). This is because we know that conditional on being in state (z, z) the next step of the processes X and Y each have a transition according to P which is somehow coupled for the two processes, but we can just change that part of the coupling, i.e. change how the transition matrix of chain (X_t, Y_t) looks only from vertices of the form (z, z). This shows why the condition $\mathbb{P}(X_1 = x' \mid X_0 = x, Y_0 = y) = P(x, x')$ and $\mathbb{P}(Y_1 = y' \mid X_0 = x, Y_0 = y) = P(y, y')$ is important. If we did not have it we would not know that the transitions of X and Y conditional on $(X_0, Y_0) = (z, z)$ are the same, (and one could come up with the example of coupling where these are not the same) so we can not amend the joint transition matrix from (z, z) to take X and Y to the same states and still have the correct marginals $(X_t)_{t\geq 0}$ and $(Y_t)_{t\geq 0}$.

One way to make an example of a coupling of Markov chains which is not Markovian would be to find a coupling of Markov chains which is not itself a Markov chain. This would mean that there is a dependence of (X_t, Y_t) as a pair on the full past and not just on the previous step, but as it is a coupling of Markov chains, the marginals X_t and Y_t still have to be independent of the past of the corresponding marginals, given the previous step. However, we give here explicit example of a non-Markovian coupling where (X_t, Y_t) is a Markov chain, but the conditional probability of $X_1 = x_1$ on both $X_0 = x$ and $Y_0 = y$ is not $P(x, x_1)$. Consider the transition matrix on $\{0, 1\}$ given by $P(x, y) = \frac{1}{2}$ for all $x, y \in \{0, 1\}$, corresponding to a sequence of i.i.d. Bernoulli $1/2$ bits. Let $(Y_t)_{t\geq 0}$ be a Markov chain with transition matrix P started with a fair coin toss, and set $X_0 = 0$ and $X_{t+1} = Y_t$ for $t \geq 0$. Both (X_t) and (Y_t) are Markov chains with transition matrix P, so (X_t, Y_t) is a coupling. Moreover, the sequence $(X_t, Y_t)_{t\geq 0}$ is itself a Markov chain, but it is not a Markovian coupling. In this chain the transition from $(0, 0)$ to $(1, \cdot)$ has probability 0 so it is not possible to amend the matrix from points $(0, 0)$ to only have non-negative probabilities to send the processes to $(1, 1)$ or $(0, 0)$ without changing either what the matrix does for some of the other states or the marginal of X_t.

However, we do not have to worry much about what non-Markovian couplings could look like, as Markovian ones are often more useful and almost all couplings used in this course will be Markovian.

We now state the first result which will be helpful for upper bounding the mixing times in various cases.

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

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

with the usual convention that $\inf \emptyset = \infty$. Then

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

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

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

This means that if we know that for some time t we have that for all x, y that $\mathbb{P}_{x,y}(\tau_{\text{couple}} > t) \leq 1/4$ then $t_{\text{mix}} \leq t$ and from Markov's inequality this also gives that $t_{\text{mix}} \leq 4 \max_{x,y} \mathbb{E}_{x,y}[\tau_{\text{couple}}]$

This follows from TV distance upper bound in terms of coupling as X_t, Y_t is a coupling of $P^t(x,\cdot), P^t(y,\cdot)$ and as from coalescence $\mathbb{P}_{x,y}(X_t \neq Y_t) = \mathbb{P}_{x,y}(\tau_{\text{couple}} > t)$.

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 using the general bound of total variation distance by coupling, i.e. Theorem 1.7

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

The final assertion of the theorem follows now as $d(t) \leq \bar{d}(t)$ from Lemma 1.12. □

We will soon see an application of this result. First, we notice that actually, the only thing we have used is that the coupling we are considering is coalescent and the above result would be true for any coalescent coupling of Markov chains, as $X_t \neq Y_t$ would imply that $\tau_{\text{couple}} > t$ for any coalescent coupling. However, if we consider couplings which are not Markovian, it is not easy to modify them to make them coalescent (as unlike the Markovian one, we can not generate them step by step, as we do not know that the next step of the coupling process is independent of the past and is just according to the exact transition matrices we need for X_t and Y_t coupled in some way). So, while the above result holds for coalescent couplings more generally, its main use is when the couplings are Markovian, as those couplings are most natural to come up with and can always be modified to be coalescent (without increasing the coupling time). *See Appendix A for an example of non-Markovian coupling which is not coalescent and so for which the coupling time can not be used to bound the mixing time[2]

Lecture 4

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 $f(n) \gtrsim g(n)$ hold. We also write $g(n) \ll f(n)$ or $f(n) \gg g(n)$ if $g(n) = o(f(n))$ i.e. if $g(n)/f(n) \to 0$ as $n \to \infty$.

Claim 2.4 (Mixing time of lazy random walk on cycle $\mathbb{Z}_n$)**.** *A lazy simple random walk on $\mathbb{Z}_n$ is a lazy version of a simple random walk on a cycle, which is a graph with vertices $\mathbb{Z}_n = \{0, \ldots, n-1\}$ with edges between i and j iff they differ by $1 \mod n$, i.e. it is a Markov chain with transitions $(P+I)/2$ where only non zero transitions in P are $P(i, (i \pm 1) \mod n) = 1/2$ (recall the definitions of lazy chain and simple random walk from definitions 0.15 and 0.11). Then the mixing time of lazy simple random walk on $\mathbb{Z}_n$ satisfies*

$$t_{\text{mix}} \asymp n^2$$

For the upper bound construct a Markovian coupling in which walks have the positive probability to collide if they are on neighbouring vertices, i.e. they do not avoid colliding forever by jumping over each other (and use Gambler's ruin). For the lower bound show that at some time of order n^2 the walk has a small probability to reach some fixed large set (particularly, it has a probability at most $1/4$ to be further than $n/4$ from its starting point).

Before proving the theorem, we give some informal motivation, which should hopefully help us understand better how one constructs useful optimal couplings. As one can expect, in the proof below, we are going to construct a coupling of two lazy walks X and Y starting from x and y

[2]All examples of non-Markovian coupling are non-examinable and they are just presented in the notes in order to explain better why we only want to work with the Markovian couplings when bounding mixing times.

respectively, and then we will use Theorem 2.3. We want to find a coupling in which the coupling time is as quick as possible, and so we ideally want the walks to approach each other at every step. We could try to always either keep both walks in place or, if they jump, couple it such that when one jumps clockwise, the other one jumps anticlockwise. However, unless the walks have distance $n/2$, in one of these cases, the shortest distance between the walks will actually increase. We can see that unless we are moving the two walks such that their distance always stays the same (in which case they never couple), we can not avoid having situations where the distance of the walks increases. Moreover, the problem with the coupling we proposed above is that if the distance between the walks is 1 at some time, then the walks will jump over each other and will not coalesce, and if n is even, their distance will always be odd, which means that the walks will never coalesce so Theorem 2.3 doesn't give us anything useful. So to construct a useful Markovian coupling we will want a coupling that doesn't have this parity issue and is also such that when the walks are at distance 1 they have a chance of coalescing in the next step. We do this by only letting one walk move at a time. [Only allowing one chain to move at any given time is actually a common way of coupling lazy chains as it avoids issues of chains jumping over each other.]

Proof. Consider a coupling defined as follows: 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. 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 runs until it hits 0 or n for the first time (and then gets absorbed there). By gambler's ruin (or mean time to absorption problem, which can be solved by writing appropriate recurrences for $\tau_k = \mathbb{E}_k\big[T_{\{0,n\}}\big]$), 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, as the coupling we constructed is clearly coalescent and it is easy to see it is Markovian as well, we obtain

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

where we used Markov's inequality for the second inequality and then bound $k(n-k) \le n^2/4$ for $0 \le k \le n$. Taking $t = n^2$ gives $d(n^2) \le 1/4$, and hence this proves the upper bound $t_{\text{mix}} \le n^2$.

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 $X_t = S_t \mod n$ and by Chebyshev's inequality

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

Taking now $t = n^2/33$ (or any $t < n^2/32$) gives $\mathbb{P}_0(X_t \in A) < 1/4$, where $A = \{\lceil n/4\rceil + 1, \dots, \lceil 3n/4\rceil\}$. But $\pi(A) \ge 1/2$, and hence we deduce

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

which shows that $t_{\text{mix}} \ge n^2/33$ and completes the proof of the lower bound. □

In the above proof, we have used a common strategy for determining a lower bound on the mixing time. We have just found a set which has a large π measure but which the chain likely does not reach by the time t.

Claim 2.5 (Lazy random walk on a binary tree)**.** *We consider a lazy simple random walk on a finite rooted tree on n vertices with the property that the root has degree two, leaves have degree 1 and every other vertex has degree 3 (for this to be possible n has to be of the form $2^k - 1$, where $k - 1$ is the largest distance of any vertex from the root). Then for this Markov chain, we have*

$$t_{\text{mix}} \asymp n.$$

Construct a coupling which makes sure walks reach the same level first without jumping over each other and then move through the same level until they meet (which has to happen by the time they reach the root). For the lower bound use the ES1 result saying that mixing time is bounded by the maximal hitting time of a large enough set and apply it to one side of the tree when started from the leaf in the other side.

Proof. 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. By a level of a tree we mean a collection of vertices which all have the same distance from the root, and the level i corresponds to the ones on distance i and therefore has 2^i vertices. We first want the two walks to be in the same level of the tree, and starting from there we will move them together up and down the tree and they will couple (coalesce) the first time they reach the root. More precisely, 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 X jumps to a neighbour chosen uniformly at random and Y stays in place. If tails, then Y moves while X stays in place. The first time they reach the same level, we move them up or down together. Then if we wait for the first time one of them has visited the root after having previously visited any of the leaves, the walks must have coupled. This is because if one of the walks starts from a leaf, regardless of where the other walks starts from, the walks have to be at the same level before this walk reaches the root, simply as only one walk moves at a given time until they are in the same level, so the walks can not jump over each other.

By reducing to a lazy version of a lazy biased simple random walk (with a probability to jump to the left being 1/6, to stay in place being 1/2 and to jump to the right being 1/3) on the line segment of length k (using that the expected hitting time of 0 starting from $k-1$ is of order 2^k) , if τ is the first time the walks couple, then we have $\mathbb{E}_{x,y}[\tau] \leq Cn$ for a positive constant C. Therefore, using Markov's inequality again 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 $t_{\text{mix}} \leq 4Cn$. For a lower bound, we use Example Sheet 1 Question 8 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 $t_{\text{mix}} \asymp n$. □

2.1 Strong stationary times

We now present another technique which allows us to upper bound the mixing time. For the Markov chain X the goal is to find a random time τ with the property that τ is independent of X_τ and X_τ is distributed according to π. We will show how to use the expectation of such a time in order to bound the mixing time and then we will give some examples where strong stationary times can be found.

Definition 2.6 (Strong stationary times)**.** We know that 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. More generally a (randomised) stopping time is a random time T such that $\{T \leq t\}$ is determined by the filtration $\mathcal{F}_t$ to which X is adapted and with respect to which it is still a Markov chain [Filtration is just an increasing sequence of sigma-algebras and we say that $(X_t)_{t\geq 0}$ is adapted to $(\mathcal{F}_t)_{t\geq 0}$ if X_t is $\mathcal{F}_t$ measurable for all $t \geq 0$ and we say that it is a Markov chain with respect to $\mathcal{F}_t$ if we have that $\mathbb{P}(X_{t+1} = y \mid \mathcal{F}_t) = P(X_t, y)$ for all $t \geq 0$ and $y \in S$].

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

A stationary time τ is called a *strong stationary time* (starting from x) if X_τ is independent of τ, i.e. if it satisfies that for all $y \in S$

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

//

When we use the term stopping time, we mean the randomised stopping time, i.e. the more general definition of a stopping time. It is a more general definition as, for example, we can have that $\{T \leq t\}$ depends on the states $X_0, \ldots, X_t$ and also on some other process which is independent of X_t. An example of a (randomised) stopping time would be a time T which depends on repeated independent coin tosses which are independent of X_t and is defined by $T = \inf\{t \geq 0 : \text{ coin } t \text{ is H and } X_t \in A\}$ where A is some fixed subset of S. An example of (randomised) stopping time is also any time T, which is a stopping time according to the less general definition. Another useful example is one where we can use a more general process to generate moves of our Markov chain and then define T to depend on that process. This would also be adding extra bits of randomness which do not change how the chain behaves and we will see an example of how this can be used.

The goal will now be to show that the distance to equilibrium at time t is at most the probability that the strong stationary time T is at least t. We should not be surprised that this happens as at the strong stationary time, T, the chain is distributed according to π and therefore it should be distributed according to π at all later times as well. However, we need to be a bit careful as T is a random time, but as we are looking at a strong stationary time, not just stationary time, we also know that conditional on the exact value T takes, say $T = u$ the chain is at time u distributed according to π and so it should be distributed like π at all later times as well. Therefore the distance to equilibrium should be bounded by the probability that strong stationary time has not happened yet. To give formal proof, we start by defining a separation distance.

Definition 2.7 (Separation distance)**.** We define the separation distance $s(t)$ to be

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

//

Lemma 2.8 (Separation distance upper bounds total variation distance)**.** *For all $x \in S$ we have*

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

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

Just use the expression for TV with $(\mu - \nu)^+$.

Proof. Using the definition of total variation distance we have

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

and this concludes the proof. □

Proposition 2.9 (Strong stationary time upper bound on distance to equilibrium)**.** *If τ is a strong stationary time when $X_0 = x$, then for all t we have*

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

As $\mathbb{P}_x(X_t = y, \tau \le t) \le \mathbb{P}_x(X_t = y)$ it is enough to check that $\mathbb{P}_x(X_t = y, \tau \le t) = \pi(y)\mathbb{P}_x(\tau \le t)$.

Proof. Using Lemma 2.8 it suffices to show $s_x(t) \le \mathbb{P}_x(\tau > t)$.

For all $x \in S$ we have that for some $y \in S$

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

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

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

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

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

where for the last equality we used the stationarity of π. This implies that $s_x(t) \le 1 - \mathbb{P}_x(\tau \le t)$ and concludes the proof. □

Before giving an example of strong stationary time and its use for bounding the mixing time, we quickly show that at least for reversible chains, we can also upper bound separation distance by the function of the distance to equilibrium. One can then also define a separation mixing time by replacing $d(t)$ with $s(t)$ in the mixing time definition and in the case of reversible chains bound the separation mixing time in terms of suitable ε-mixing time from both sides using the following result and the bound we have already shown.

Lecture 5

Lemma 2.10 (Separation is upper bounded by distance to equilibrium for reversible chains)**.** *For reversible chains we have*

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

Common idea for reversible chains: write $P^{2t}(x,y)$ by summing over all states $z \in S$ the chain could go through at step t and then reverse the second half of the path (i.e. use $\frac{P^t(z,y)}{\pi(y)} = \frac{P^t(y,z)}{\pi(z)}$) and then suitably apply Cauchy-Schwarz (and use expression with minimum for TV in this case).

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

$$\frac{P^{2t}(x,y)}{\pi(y)} = \sum_{z\in S} P^t(x,z)\frac{P^t(z,y)}{\pi(y)} = \sum_{z\in S}\frac{P^t(x,z)P^t(y,z)}{\pi(z)} = \sum_{z\in S}\frac{P^t(x,z)P^t(y,z)}{\pi(z)^2}\pi(z)$$

Applying Cauchy-Schwarz inequality [or Jensen's on x^2] to the expression above, we get

$$\begin{aligned}\frac{P^{2t}(x,y)}{\pi(y)} &\geq \left(\sum_{z\in S}\sqrt{\frac{P^t(x,z)P^t(y,z)}{\pi(z)^2}}\cdot\pi(z)\right)^2 \\ &\geq \left(\sum_{z\in S} P^t(x,z)\wedge P^t(y,z)\right)^2 = \left(1-\left\|P^t(x,\cdot)-P^t(y,\cdot)\right\|_{\mathrm{TV}}\right)^2,\end{aligned}$$

where for the final equality we have used one of the expressions for total variation from Proposition 1.3. Rearranging the above and taking the maximum over all x and y in S proves that $s(2t) \leq 1-(1-\bar{d}(t))^2$ and using the Lemma 1.12 completes the proof. □

We will now present two examples where strong stationary times are used to get the value of the mixing time. For both of these, we will need to use the following coupon collector problem:

Proposition 2.11 (Coupon collector)**.** *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 to be any of the n types. Let τ be the number of coupons he acquires until he obtains a full set. Then*

$$\mathbb{E}[\tau] = n\sum_{k=1}^{n}\frac{1}{k}$$

and for any $c>0$

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

Write τ as the sum of geometric and for the second part use the union bound on the events that a specific coupon has not appeared by time $\lceil n\log n + cn\rceil$.

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

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

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

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

Since the probability of not drawing coupon i in a given trial is $1-\frac{1}{n}$ 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. □

Claim 2.12 (Mixing time of a lazy random walk on a 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. For all ε there is a constant $C(\varepsilon) > 0$ such that a lazy simple random walk on an n-dimensional hypercube satisfies*

$$t_{\text{mix}}(\varepsilon) \leq n \log n + C(\varepsilon) n.$$

Think of the walk as at every time step picking a uniform coordinate which gets refreshed, and use the random time which is the first time all coordinates were refreshed at least once.

Proof. We note that using the symmetry of the chain the invariant distribution for this chain is uniform [Note that in general, to see that an invariant distribution is uniform, it is enough to check that for every state the sum of incoming probabilities to that state is 1, i.e. that a transpose of transition matrix is stochastic]. 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 it is not hard to check that 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 Propositions 2.11 and 2.9 that

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

and hence by taking $C(\varepsilon)$ large, we get that $t_{\text{mix}}(\varepsilon) \leq n \log n + C(\varepsilon) n$. □

Claim 2.13 (Random-to-top shuffle mixing time)**.** *Recall from our first Example 1.1 that a random-to-top shuffling of cards consists of picking a random card and moving it to the top of the deck. For all ε there is a constant $C(\varepsilon) > 0$ such that the ε-mixing time of a random-to-top shuffling of n cards satisfies*

$$n \log n - C(\varepsilon) n \leq t_{\text{mix}}(\varepsilon) \leq n \log n + C(\varepsilon) n.$$

For an upper bound show that the first time all cards were moved is a strong stationary time and us a coupon collector problem and for a lower bound show that the probability that j cards on the bottom of the uniform deck are in the decreasing order is small while for some C at time $n \log n - Cn$ this happens with sufficiently larger probability.

Proof. We note that it is easy to check that the invariant distribution is uniform.

We first show the *lower bound*: let A_j be the event that the j bottom cards of the deck are in their original relative order (that is, if these cards are in order from the bottom $i_1, \ldots, i_j$ then we have $i_1 > i_2 > \ldots > i_j$). For a uniform permutation,

$$\pi(A_j) = \frac{1}{j!}$$

as any arrangement of the j bottom cards is equally likely. Thus if j is reasonably large, this has a small probability for a uniform permutation. However, if we are significantly before τ then the event A_j has a large chance to hold for some high value of j. Indeed, more formally, let τ_j be the first time that we have picked $n - j + 1$ different cards, then similarly to the coupon collector proof, as τ_j is a sum of $n - j + 1$ independent geometric variables, we obtain

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

It is clear that if $\tau_j \geq t$, then the event A_{j-1} holds. We thus deduce using Chebyshev's inequality

$$P^{n\log n - Cn}(\mathrm{id}, A_{j-1}) \geq \mathbb{P}(\tau_j \geq n\log n - Cn) \geq 1 - \mathbb{P}(|\tau_j - \mathbb{E}[\tau_j]\,| \geq n(C - \log j)) \geq 1 - \frac{1}{j-1}$$

for $C \geq \log j + 1$. [The above indeed holds as follows: Notice that $\{|\tau_j - \mathbb{E}[\tau_j]\,| < n(C - \log j)\}$ just means that $\{\tau_j \in (\mathbb{E}[\tau_j] - Cn + n\log j, \mathbb{E}[\tau_j] + Cn - n\log j)\}$ which, as $\mathbb{E}[\tau_j] \geq n\log n - n\log j$ implies that $\{\tau_j \geq n\log n - Cn\}$ and so

$$\begin{aligned} 1 - \mathbb{P}(|\tau_j - \mathbb{E}[\tau_j]\,| \geq n(C - \log j)) &= \mathbb{P}(|\tau_j - \mathbb{E}[\tau_j]\,| < n(C - \log j)) \\ &\leq \mathbb{P}(\tau_j \geq n\log n - Cn)\,. \end{aligned}$$

We then use that Chebychev's inequality gives

$$\mathbb{P}(|\tau_j - \mathbb{E}[\tau_j]\,| \geq n(C - \log j)) \leq \frac{\mathrm{Var}\,(\tau_j)}{n^2(C - \log j)^2} \leq \frac{n^2}{(j-1)n^2(C - \log j)^2} \leq \frac{1}{j-1}$$

when $C \geq \log j + 1$ and combining the previous two display equations gives the claimed inequality.] Now,

$$d(n\log n - Cn) \geq P^{n\log n - Cn}(\mathrm{id}, A_{j-1}) - \pi(A_{j-1}) \geq 1 - \frac{2}{j-1}.$$

Taking $j = \lfloor e^{C-1} \rfloor$ and using this for $n \geq j$, we get

$$d(n\log n - Cn) \geq 1 - \frac{2}{e^{C-2} - 1},$$

and hence taking $C(\varepsilon)$ sufficiently large gives that $t_{\mathrm{mix}}(\varepsilon) \geq n\log n - C(\varepsilon)n$. [To be very precise, we mention that we have shown that for all ε for all sufficiently large n, i.e. for all $n > N_\varepsilon$, for some fixed constant N_ε depending only on ε, we have $t_{\mathrm{mix}}(\varepsilon) \geq n\log n - C(\varepsilon)n$. But by setting $C'(\varepsilon) = \max\{C(\varepsilon), \log N_\varepsilon\}$, we get that the same bound on the mixing time holds for all n with $C'(\varepsilon)$ in place of $C(\varepsilon)$ (as for $n \leq N_\varepsilon$ we are just using that the mixing time is at least 0).]

We now show the *upper bound*: We first show that a first time when cards with all labels have been picked at least once, which we label τ, is a strong stationary time. We prove by induction on time the following: for $k \leq \min\{t, n\}$ conditional on k being the number of different cards which have been picked by time t and conditional on the number of different cards which have been picked by time $t-1$ (this can be either k or $k-1$), that the set C_k of k cards on the top of the deck is a uniform subset of $\{1, \dots n\}$ of size k, and the order of k cards on the top of the deck is uniformly random order of C_k. The statement is clearly true for $t = 1$ as there $k = 1$ (and also by time 0 we have picked 0 cards) and the card on the top of the deck is a uniform card. Assume the statement is true at time t and any k. Notice that this also implies that if we know that k cards have been picked by time t then regardless of how many cards have been picked by time $t-1$ we know that the k cards on top are uniform cards in uniform order (as in both cases we have that top k cards are uniform cards in uniform order). Assume that k cards have been picked by time t. Then at time $t+1$ we have either picked one of the already picked k cards (i.e. one of the k cards which are currently at the top of the deck) and moved it to the top of the deck, in which case we moved a random card from the uniform order of C_k to the top, which still gives us a uniform permutation of C_k and so the claim holds in this case. In the other case, we have picked a new card and moved it to the top. This new card is equally likely any cards from $\{1, \dots, n\} \setminus C_k$ so the set C_{k+1} is a uniform subset of $\{1, \dots n\}$ of size $k+1$ and moreover, conditional of the values in C_{k+1}, we see that C_k is a uniform subset of size k of C_{k+1} as any order of picking these given $k+1$ cards was equally likely. This means that the card on the top is a uniform card from C_{k+1} and the next k

cards are in a uniformly random relative order as they were a uniform permutation of C_k. This means that in both cases all cards which have been picked so far are in a uniform order. So we have shown that conditional on k' being the number of cards which were picked by time $t+1$ and either $k'-1$ or k' being the number of cards picked by time t, we have that at time $t+1$ the top k' cards are uniform cards and have equally likely any of the $k'!$ orders. This completes the induction. Therefore, conditional on $\tau = t$ (which is just the event that by time t all n cards were picked while by time $t-1$ exactly $n-1$ cards have been picked) we know that at time t the order of cards is uniform giving that τ is a strong stationary time. The time τ is distributed as the coupon collector time and therefore using Propositions 2.11 and 2.9 we have that for any C

$$d(n\log n + Cn) \leq \mathbb{P}(\tau > n\log n + Cn) \leq e^{-C},$$

and so picking $C(\varepsilon)$ large enough gives $t_{\text{mix}}(\varepsilon) \leq n\log n + C(\varepsilon)n$. □

We can notice that we have shown that when shuffling n cards then if we shuffle for time $(1+\delta)n\log n$ we get arbitrarily close to uniform as the size of the deck gets larger, while if we shuffle for time $(1-\delta)n\log n$ the deck is still "arbitrarily far" from the uniform deck (i.e. $d((1-\delta)n\log n) \to 1$ as $n \to \infty$) and this tells us that somehow all of the mixing happens at the time $n\log n$, i.e. at some window of time around this time which has a smaller order than $n\log n$. This is a nice property to know, as it tells us that we know "exactly" (i.e. up to smaller order terms) how long we should shuffle the deck when the number of cards n is large enough. This motivates the following definition.

Definition 2.14 (Cutoff)**.** A sequence of Markov chains X^n is said to exhibit *cutoff*, if for all $\varepsilon \in (0,1)$

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

where $t^n_{\text{mix}}(\varepsilon)$ corresponds to the ε-mixing time of the nth chain. Equivalently, writing $d_n(t)$ for $d(t)$ defined with respect to X^n, there is a sequence t_n such that for all $\delta > 0$

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

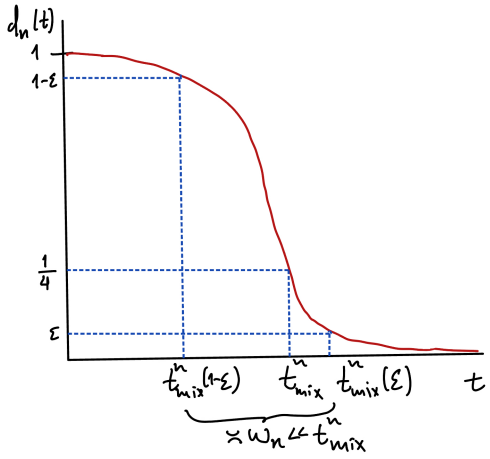
//

Definition 2.15 (Cutoff window)**.** A sequence of Markov chains is said to have cutoff with a *cutoff window* of order w_n if $w_n \ll t^n_{\text{mix}}$ (this means that $w_n/t^n_{\text{mix}} \to 0$ as $n \to \infty$) and for any ε there is a constant $C(\varepsilon) > 0$ such that

$$t^n_{\text{mix}} - C(\varepsilon)w_n \leq t^n_{\text{mix}}(\varepsilon) \leq t^n_{\text{mix}} + C(\varepsilon)w_n$$

//

When thinking about cutoff property we should have the following plot in mind which explains how $d_n(t)$ decreases for the nth chain.



We see that from the Claim 2.13 that a random-to-top card shuffling exhibits cutoff at a time $n \log n$ with a window of order n. The cutoff was observed to hold for many other card shufflings and some other standard Markov chains. It however also doesn't hold for many examples of Markov chains. No condition which would be simple to check (even for just some types of Markov chains such as random walks on graphs) and which is equivalent to cutoff is known. However, we will see some simple conditions which would tell us that the cutoff doesn't hold in case of reversible chains and throughout the course, we will show the existence or lack of cutoff for some particular chains.

3 $\mathcal{L}^p$ distance

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

$$\|f\|_p = \|f\|_{p,\pi} = \begin{cases} \left(\sum_{x\in S} |f(x)|^p \pi(x)\right)^{1/p} & \text{if } 1 \le p < \infty, \\ \max_{x\in S} |f(x)| & \text{if } p = \infty. \end{cases}$$

For functions f, g, we define the scalar product $\langle f, g\rangle_\pi = \sum_{x\in S} f(x)g(x)\pi(x)$. Finally, we define $q_t(x, y) = \frac{P^t(x,y)}{\pi(y)}$. When the chain is reversible, then from detailed balance we have $q_t(x, y) = q_t(y, x)$. We define the $\mathcal{L}^p$ distance via

$$d_p(t) = \max_{x\in S} \|q_t(x, \cdot) - 1\|_p = \begin{cases} \max_{x\in S} \left(\sum_{y\in S} \left|\frac{P^t(x,y)}{\pi(y)} - 1\right|^p \pi(y)\right)^{1/p} & \text{if } 1 \le p < \infty, \\ \max_{x\in S} \max_{y\in S} \left|\frac{P^t(x,y)}{\pi(y)} - 1\right| & \text{if } p = \infty. \end{cases}$$

Using Proposition 1.3 and Jensen's inequality [as $x^{p/r}$ is concave for $p < r$ so $\sum_{x\in S} |f(x)|^p \pi(x) = \sum_{x\in S} (|f(x)|^r)^{p/r} \pi(x) \le \left(\sum_{x\in S} |f(x)|^r \pi(x)\right)^{p/r}$], we have 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 3.1 (Relating d_2 and d_∞ for reversible chain)**.** *For reversible Markov chains, we have*

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

Show $\frac{P^{2t}(x,y)}{\pi(y)} - 1 = \sum_{z\in S} \left(\frac{P^t(x,z)}{\pi(z)} - 1\right)\left(\frac{P^t(y,z)}{\pi(z)} - 1\right)\pi(z)$ by reversing the path from x to y of length $2t$ starting from the vertex z which the chain goes through at time t, and using Cauchy-Schwarz upper bound this by $\leq \sqrt{\left[\sum_{z\in S}\left(\frac{P^t(x,z)}{\pi(z)} - 1\right)^2 \pi(z)\right] \cdot \left[\sum_{z\in S}\left(\frac{P^t(y,z)}{\pi(z)} - 1\right)^2 \pi(z)\right]}$.

Proof. By reversibility we have

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

Using that $\sum_{z\in S} P^t(x,z) = 1$ and $\sum_{z\in S}\pi(z) = 1$ and the above gives

$$\frac{P^{2t}(x,y)}{\pi(y)} - 1 = \sum_{z\in S}\left(\frac{P^t(x,z)}{\pi(z)} - 1\right)\left(\frac{P^t(y,z)}{\pi(z)} - 1\right)\pi(z) \tag{3.1}$$

and then taking $x = y$ and maximising over x proves the second equality in the proposition, as by definition $\|q_t(x,\cdot) - 1\|_2 = \sqrt{\sum_{z\in S}\left(\frac{P^t(x,z)}{\pi(z)} - 1\right)^2 \pi(z)}$. Applying Cauchy-Schwarz inequality, we obtain

$$\left|\frac{P^{2t}(x,y)}{\pi(y)} - 1\right| \leq \sqrt{\left[\sum_{z\in S}\left(\frac{P^t(x,z)}{\pi(z)} - 1\right)^2 \pi(z)\right] \cdot \left[\sum_{z\in S}\left(\frac{P^t(y,z)}{\pi(z)} - 1\right)^2 \pi(z)\right]}$$

Taking the maximum over all x and y shows that $d_\infty(2t) \leq (d_2(t))^2$. Taking $x = y$ in the equation (3.1) then proves using above that

$$\begin{aligned} d_\infty(2t) = \max_{x,y}\left|\frac{P^{2t}(x,y)}{\pi(y)} - 1\right| &\leq \max_{x,y}\sqrt{\left[\sum_{z\in S}\left(\frac{P^t(x,z)}{\pi(z)} - 1\right)^2 \pi(z)\right] \cdot \left[\sum_{z\in S}\left(\frac{P^t(y,z)}{\pi(z)} - 1\right)^2 \pi(z)\right]} \\ &= \max_{x,y}\sqrt{\left(\frac{P^{2t}(x,x)}{\pi(x)} - 1\right)\cdot\left(\frac{P^{2t}(y,y)}{\pi(y)} - 1\right)} \leq d_\infty(2t). \end{aligned}$$

and so inequalities have to be equalities and $d_\infty(2t) = (d_2(t))^2 = \max_x \frac{P^t(x,x)}{\pi(x)} - 1$ □

Lecture 6

Corollary 3.2 ($\mathcal{L}^2, \mathcal{L}^\infty$ and TV mixing for a reversible chain)**.** *For a reversible chain we have that*

$$t_{\text{mix}}(\varepsilon) \leq t_{\text{mix}}(\varepsilon/2) = t^{(1)}_{\text{mix}}(\varepsilon) \leq t^{(2)}_{\text{mix}}(\varepsilon) \leq t^{(\infty)}_{\text{mix}}(\varepsilon) \leq 2t^{(2)}_{\text{mix}}(\sqrt{\varepsilon}) \text{ and } t^{(2)}_{\text{mix}}(\varepsilon) = \left\lceil \frac{1}{2} t^{(\infty)}_{\text{mix}}(\varepsilon^2)\right\rceil$$

Proof. All but the last two inequalities hold in general from the monotonicity of $\mathcal{L}^p$ and the fact that the total variation distance is a half of $\mathcal{L}^1$ distance. The last two results are direct applications of the above proposition and hold only for a reversible chain. □

4 Spectral techniques

4.1 Spectral decomposition and relaxation time

In this section, we explore the connection between the mixing times and eigenfunctions (i.e. eigenvectors) and eigenvalues of the transition matrix. 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 $\langle f, g \rangle_\pi = \sum_{x \in S} f(x)g(x)\pi(x)$.

Theorem 4.1 (Spectral decomposition of reversible chains)**.** *Let P be reversible with respect to π. The inner product space $(\mathbb{R}^S, \langle \cdot, \cdot \rangle_\pi)$ has an orthonormal basis of real-valued eigenfunctions $(f_j)_{j \le |S|}$ corresponding to real eigenvalues $(\lambda_j)_{j \le |S|}$ 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*

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

Use that from the spectral theorem for symmetric matrices as $\frac{\sqrt{\pi(x)}}{\sqrt{\pi(y)}} P(x,y)$ is symmetric it has orthonormal basis of eigenfunctions.

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 and so we can set $g_1(x) = \sqrt{\pi(x)}$ (also note that $\sum_{y \in S} g_1(x)^2 = \sum_{y \in S} \pi(x) = 1$). 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 $\langle f_j, f_i \rangle_\pi = \mathbb{1}_{i=j}$. So we have $P^t f_j = \lambda_j^t f_j$ and hence

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

where we have used that $\mathbb{1}_y = \sum_{j=1}^{|S|} f_j \langle f_j, \mathbb{1}_y \rangle_\pi$ as f_j give an orthonormal basis and any function can be decomposed over the orthonormal basis. Using that $f_1 = (1, \dots, 1)$ and $\lambda_1 = 1$ gives the desired decomposition. □

Remark 4.2. *This provides another proof of the ergodic theorem for reversible chains as if the chain is aperiodic one can check that -1 is not an eigenvalue and then the desired convergence holds as $t \to \infty$.*

Definition 4.3 (Spectral gap and absolute spectral gap)**.** Let P be a reversible matrix with respect to π. We order its eigenvalues

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

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 $\gamma = 1 - \lambda_2$. //

Remark 4.4. *From "Useful results for Mixing times of Markov chains" Question 6 we know that all eigenvalues of a stochastic matrix are in $[-1, 1]$ and from Example Sheet 2 Question 1 we will see that if the chain is lazy then $\gamma_* = \gamma$.*

Definition 4.5 (Relaxation time)**.** The relaxation time[3] for a reversible Markov chain is defined to be

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

//

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

To use the exact expression given in Theorem 4.1, we need to know (or be able to bound) all eigenvalues and eigenfunctions of P, which is rarely possible. Fortunately, the expression can be bounded by a function of λ_* only, giving the following simple and general bound on the $\mathcal{L}^2$ distance to stationarity for a reversible chain.

For the rest of this section, unless otherwise stated we will only talk about aperiodic chains (so ergodic, as we assume finite and irreducible anyway).

Theorem 4.6 (Poincaré inequality for reversible chains)**.** *Let P be a reversible matrix with respect to the invariant distribution π. Then for all starting distributions ν we have*

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

This proof is just algebra. Notice that $\sum_{x \in S} \nu(x) P^t(x, y)$ squared appears in $\mathcal{L}^2$ distance and write this instead as a product of sums over x and x'. Use spectral decomposition and that the functions f_i are orthogonal. Use that Theorem 4.1 gives for $t = 0$ that $\sum_{i=2}^n f_i(x) f_i(x') = \frac{\mathbb{1}_{x=x'}}{\pi(x)} - 1$.

Before proving the proof of the Poincaré inequality, we note that we should think about it as a way to upper bound the $\mathcal{L}^2$ distance from π after t steps of the chain in terms of the $\mathcal{L}^2$ distance of the starting distribution from π. It specifically tells us that in the worst case the $\mathcal{L}^2$ distance drops by a multiplicative factor of $1 - \gamma_*$ after every additional step of the Markov chain.

Proof. The proof of this consists of algebraic manipulations using the decomposition of P with respect to the orthonormal eigenfunction basis. We have

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

Using the decomposition of $P^t(x, y)$ from Theorem 4.1, and setting $n = |S|$ we obtain

$$\begin{aligned} \sum_{y \in S} \frac{1}{\pi(y)} \left(\sum_{x \in S} \nu(x) \mathbb{P}_x(X_t = y) \right)^2 &= \sum_{y \in S} \frac{1}{\pi(y)} \left(\sum_{x \in S} \nu(x) P^t(x, y) \right) \left(\sum_{x' \in S} \nu(x') P^t(x', y) \right) \\ &= \sum_{x, x' \in S} \nu(x) \nu(x') \sum_{i,j=1}^n \lambda_i^t \lambda_j^t f_i(x) f_j(x') \sum_{y \in S} f_i(y) f_j(y) \pi(y). \end{aligned}$$

[3] some mixing times notes/papers might call $\frac{1}{\gamma_*}$ absolute relaxation time and call $\frac{1}{\gamma}$ the relaxation time but in these notes (and in many other notes/books/papers) we will call $\frac{1}{\gamma_*}$ the relaxation time and $\frac{1}{\gamma}$ just the inverse of the spectral gap.

As (f_j) is an orthonormal basis of the eigenfunctions, we obtain that in this last sum, the sum over y is equal to $\mathbb{1}_{i=j}$ so above is

$$\begin{aligned}\sum_{x,x'\in S}\nu(x)\nu(x')\sum_{i,j=1}^{n}\lambda_i^t\lambda_j^t f_i(x)f_j(x')\mathbb{1}_{i=j} &= \sum_{x,x'\in S}\nu(x)\nu(x')\sum_{i=1}^{n}\lambda_i^{2t}f_i(x)f_i(x')\\ &= 1+\sum_{x,x'\in S}\nu(x)\nu(x')\sum_{i=2}^{n}\lambda_i^{2t}f_i(x)f_i(x').\\ &\leq 1+\lambda_*^{2t}\sum_{i=2}^{n}\left(\sum_{x,x'\in S}\nu(x)\nu(x')f_i(x)f_i(x')\right).\end{aligned}$$

Using that $P^0(\cdot,\cdot)=I$ and Theorem 4.1, we deduce

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

and hence this finally gives using that $1-x\leq e^{-x}$ that

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

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 4.7 (Eigenvalue decomposition bound on TV distance for reversible chains)**.** *Let P be reversible with respect to π and let $1=\lambda_1\geq\lambda_2\geq\cdots\geq\lambda_n\geq -1$, where $n=|S|$ be its eigenvalues and (f_j) the corresponding orthonormal eigenfunctions. Then for all x we have*

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

In the proof of Poincaré just set ν to be a measure with weight 1 on some fixed state x.

Proof. We had from Theorem 4.6 that

$$\left(\sum_{y\in S}\frac{1}{\pi(y)}\left(\sum_{x\in S}\nu(x)\mathbb{P}_x(X_t=y)\right)^2\right)-1=\sum_{x,x'\in S}\nu(x)\nu(x')\sum_{i=2}^{n}\lambda_i^{2t}f_i(x)f_i(x').$$

and so taking $\nu=\mathbb{1}_x$ we get that the above is exactly the wanted equality. □

Before showing how the Poincaré inequality helps us upper bound the mixing time in terms of the relaxation time we quickly apply first the above result for particular types of chains, the transitive chains. Transitive chains are simply the chains which "look the same" from every element in the state space. We define it formally as follows:

Definition 4.8 (Transitive chain)**.** A Markov chain with transition matrix P is *transitive* if for all $x, y \in S$ there is bijection $\phi = \phi_{x,y}$ such that $\phi(x) = y$ and $P(z, w) = P(\phi(z), \phi(w))$ for all $z, w \in S$. //

Lemma 4.9 (Eigenvalue decomposition for a transitive reversible chain)**.** *Let P be reversible and transitive. Then for all $x \in S$ we have*

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

where $n = |S|$ and $1 = \lambda_1 \geq \lambda_2 \geq ... \geq \lambda_n \geq -1$ are eigenvalues of P.

Notice that for a transitive chain, $P^{2t}(x, x)$ does not depend on x and that π is uniform and take the average over all states in S of the general eigenvalue decomposition.

Proof. First of all, we check that the uniform distribution, i.e., $\pi(x) = \frac{1}{n}$ for all x, is invariant for P. Indeed, to have uniform π, it is enough to check that for every state $x \in S$ we have $\sum_{z \in S} P(z, x) = 1$ but as $\sum_{x,y \in S} P(x, y) = n$ this is the same as showing that for any two states $x, y \in S$ we have $\sum_{z \in S} P(z, x) = \sum_{z \in S} P(z, y)$. This now holds from the definition of transitivity as $\sum_{z \in S} P(z, x) = \sum_{z \in S} P(\phi_{x,y}(z), \phi_{x,y}(x)) = \sum_{z \in S} P(\phi_{x,y}(z), y) = \sum_{z \in S} P(z, y)$ where the last equality holds as $\phi_{x,y}$ is a bijection. Next, recall that from the proof of Proposition 3.1 for all $x \in S$ 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 [one can check inductively that $\forall \ell > 0$ we have $P^\ell(z, w) = P^\ell(\phi_{x,y}(z), \phi_{x,y}(w))$ and so $P^{2t}(x, x) = P^{2t}(\phi_{x,y}(x), \phi_{x,y}(x)) = P^{2t}(y, y)$]. Therefore, by Lemma 4.7 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 (so $\sum_{x \in S} f_j^2(x)\pi(x) = 1$ i.e. for transitive chains $\sum_{x \in S} f_j^2(x) = n$) implies that

$$\sum_{j=2}^{n} f_j(x)^2 \lambda_j^{2t} = \frac{1}{n} \sum_{x \in S} \sum_{j=2}^{n} f_j(x)^2 \lambda_j^{2t} = \sum_{j=2}^{n} \lambda_j^{2t}$$

where for the last equality we have just swapped the sums. This concludes the proof. □

We now show an important result saying how one uses Poincaré inequality to bound the mixing time.

Theorem 4.10 (Upper bound on mixing time in terms of the relaxation time for reversible chains)**.** *Let P be reversible with respect to the invariant distribution π and let $\pi_{\min} = \min_{x \in S} \pi(x)$. Then for all $\varepsilon \in (0, 1)$ we have*

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

Direct consequence of the Poincaré inequality and upper bound on mixing in terms of $\mathcal{L}^2$ mixing.

Proof. By the monotonicity of the $\mathcal{L}^p$ norms it suffices to prove the second inequality above. Also, using that from Corollary 3.2 $t_{\text{mix}}^{(\infty)}(\varepsilon) \leq 2t_{\text{mix}}^{(2)}(\sqrt{\varepsilon})$, it suffices to prove that

$$t_{\text{mix}}^{(2)}(\sqrt{\varepsilon}) \leq \frac{1}{2} t_{\text{rel}} \log\left(\frac{1}{\varepsilon \pi_{\min}}\right). \tag{4.1}$$

To this end, fix x in the state space and by the Poincaré inequality (Theorem 4.6) we have

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

Taking $t = t_{\text{rel}} \log(1/(\varepsilon \pi_{\min}))/2$ in the above inequality shows that $t_{\text{mix}}^{(2)}(\sqrt{\varepsilon}) \leq t$ and thus proves (4.1) and completes the proof of the theorem. □

Remark 4.11. *We remark that, if we define the relaxation time in the case of non-reversible chains in the same way as in the case of reversible chains, the above result would* **not** *hold without the reversibility assumption (see Example Sheet 2 Question 4 for counterexample). The definition of relaxation time was intentionally stated only for reversible chains, as most of the results related to it do not hold without reversibility.*

It is actually also possible to lower bound the mixing time by the relaxation time in the case of reversible chains. We have the following useful result.

Theorem 4.12 (Lower bound on the mixing time in terms of the relaxation time for a reversible chain)**.** *Let P be a reversible matrix with respect to π. Let λ be an eigenvalue with $\lambda \neq 1$. Then $2d(t) \geq |\lambda|^t$. Moreover, for all $\varepsilon \in (0,1)$ we have*

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

For eigenpair (f, λ) where $\lambda \neq 1$ show using that $\mathbb{E}_\pi[f] = 0$ that $|\lambda|^t |f(x)| \leq \max_{y \in S} |f(y)| 2d(t)$ ($\mathbb{E}_\pi[f] = 0$ holds as f needs to be orthogonal to $f_1 = (1, \ldots, 1)$).

Proof. Let f be the eigenfunction corresponding to the eigenvalue λ. Then, by the orthogonality of the eigenfunctions, f is orthogonal to $f_1 = (1, \ldots, 1)$ corresponding to $\lambda_1 = 1$ (as $\lambda \neq 1$). Therefore, $\mathbb{E}_\pi[f] = \langle f, 1 \rangle_\pi = 0$. Using $P^t f = \lambda^t f$ for all $t \geq 0$ gives

$$|\lambda^t f(x)| = |P^t f(x)| = \left| \sum_{y \in S} (P^t(x,y) f(y) - \pi(y) f(y)) \right| \leq \max_{y \in S} |f(y)| \cdot 2d(t).$$

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

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

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

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

and this completes the proof. □

Corollary 4.13 (λ_* is limit of $d(t)^{\frac{1}{t}}$ for reversible chain)**.** *Let P be reversible with respect to π. Then we have*

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

Proof. From the proof of Theorem 4.12 we see that $\liminf_{t\to\infty} d(t)^{\frac{1}{t}} \geq \lambda_*$. For the other direction we use again the monotonicity of $\mathcal{L}^p$ norms in p to get

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

where the last inequality follows from the Poincaré inequality (Theorem 4.6). This completes the proof. □

Lecture 7

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

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

We will now show that for a sequence of reversible Markov chains if the relaxation time and mixing time are of the same order (i.e. for any fixed ε $t^n_{\text{mix}}(\varepsilon) \asymp t^n_{\text{rel}}$, where n denotes the nth chain) then cutoff can not happen. Notice that from Theorem 4.12 we already know that for all ε the mixing time is at least as large as the relaxation time times a function of ε, so this result will tell us that cutoff can only happen when the relaxation time is of strictly smaller order than the mixing time (i.e. $t^n_{\text{rel}} \ll t^n_{\text{mix}}$). To show this we show that a weaker condition than cutoff fails when the mixing and relaxation time are of the same order.

Definition 4.14 (Pre-cutoff)**.** A sequence of chains satisfies a weaker condition than cutoff, called *pre-cutoff* if

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

//

Definition 4.15 (Product condition)**.** Let P_n be a sequence of reversible Markov chains with mixing times t^n_{mix} and relaxation times t^n_{rel}. Then we say that P_n satisfies the *product condition* if

$$t^n_{\text{mix}} \gg t^n_{\text{rel}},$$

i.e. if $t^n_{\text{rel}}/t^n_{\text{mix}} \to 0$ as $n \to \infty$. //

Proposition 4.16 (No cutoff without the product condition, reversible chain)**.** *Let P_n be a sequence of reversible Markov chains with mixing times t^n_{mix} and relaxation times t^n_{rel}. If $t^n_{\text{mix}} \to \infty$ as $n \to \infty$ but the product condition doesn't hold, i.e. if $t^n_{\text{rel}}/t^n_{\text{mix}}$ doesn't converge to 0, then there is no pre-cutoff and so there is also no cutoff.*

Directly implied by the lower bound on the mixing time in terms of the relaxation time.

Proof. Dividing both sides of the statement of Theorem 4.12 by t^n_{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 $\frac{t^n_{\text{rel}}}{t^n_{\text{mix}}}$ doesn't converge to 0, we have that there is some $c > 0$ such that there is an infinite set $J \subset \mathbb{N}$ such that for $n \in J$ we have $\frac{t^n_{\text{rel}}}{t^n_{\text{mix}}} > c$. We then obtain that the right-hand side above is lower bounded by $c/2 \log(1/(2\varepsilon))$ for all large enough $n \in J$ (as $-1/t^n_{\text{mix}} \to 0$ as $n \to \infty$). Letting $\varepsilon \to 0$ proves the proposition. □

4.2 Applications and Wilson's method

Claim 4.17 (No cutoff for lazy random walk on a cycle)**.** *Recall from Claim 2.4 the definition of a lazy simple random walk on a cycle $\mathbb{Z}_n$ and that its mixing time satisfies $t^n_{\text{mix}} \asymp n^2$. If X^n is the sequence of lazy simple random walks on $\mathbb{Z}_n$ with the nth chains mixing time t^n_{mix} and relaxation time t^n_{rel} then we have that*

$$t^n_{\text{mix}} \asymp t^n_{\text{rel}} \asymp n^2$$

and therefore by Proposition 4.16 for this sequence of chains there is no pre-cutoff and so there could be no cutoff.

Looking at the roots of unity we can find all eigenvalues and eigenfunctions for this chain (which by definition means we can also exactly find t_{rel}).

Proof. We will find the eigenfunctions and the corresponding eigenvalues for a lazy simple random walk on $\mathbb{Z}_n$ and as we will find all eigenvalues we will be able to directly find the relaxation time. Let f be an eigenfunction with eigenvalue λ. Then it must satisfy $Pf = \lambda f$ which rewrites to

$$\frac{f(x)}{2} + \frac{f(x+1)}{4} + \frac{f(x-1)}{4} = \lambda f(x) \tag{4.2}$$

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

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

Then it is straightforward to check that for each k, the function f_k satisfies (4.2) with

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

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 we get

$$\varphi_k(x) = \cos\left(\frac{2\pi k x}{n}\right) \quad \text{and} \quad \psi_k = \sin\left(\frac{2\pi k x}{n}\right)$$

for real eigenfunctions. Taking $k = 0$ gives (as expected) $\lambda_0 = 1$ and taking $k = 1$ gives the second largest eigenvalue, which by laziness also corresponds to λ_*. So we get

$$\lambda_* = \lambda_1 = \frac{1 + \cos(2\pi/n)}{2} \approx 1 - \frac{\pi^2}{n^2} + O(n^{-4}).$$

Therefore, this implies that

$$t_{\text{rel}} \sim \frac{n^2}{\pi^2} \text{ as } n \to \infty.$$

Since $t_{\text{rel}} \asymp n^2 \asymp t_{\text{mix}}$ from Claim 2.4, it follows that the lazy random walk on the cycle does not exhibit pre-cutoff and, therefore, also doesn't exhibit cutoff. □

We will see in Example Sheet 2, Question 7 that $d(\alpha n^2)$ converges to a smoothly decreasing function in α as $n \to \infty$. By the definition of cutoff if the chain had a cutoff around $c \cdot n^2$ for some constant c then the above would have to converge to a step function. The limit of $d(\alpha t^n_{\text{mix}})$ as $n \to \infty$ is also

known as a limit profile of a sequence of chains and, as mentioned, it looks like a step function (it is 1 for $\alpha < 1$ and 0 for $\alpha > 1$) if and only if there is cutoff. If there is no cutoff the limit profile is a decreasing function which could have many different shapes.

We now present a method which lets us lower bound $d(t)$ and therefore also the mixing time, if we know some eigenpair of P and then we use it to show cutoff for a lazy simple random walk on a hypercube (recall the definition and an upper bound on mixing time from 2.12, we will see shortly that the upper bound was of the correct order but the constant in it was not sharp). We emphasise that for the following result, reversibility is not required.

Lemma 4.18 (Wilson's method)**.** *Let X be an irreducible and aperiodic Markov chain and let f be an eigenfunction corresponding to eigenvalue λ with $\frac{1}{2} < \lambda < 1$. Suppose there exists $R > 0$ such that $\forall x \in S$*

$$\mathbb{E}_x\left[(f(X_1) - f(X_0))^2\right] \leq R.$$

Then we have that for all $x \in S$

$$\forall t \in \mathbb{N}, \quad d(t) \geq \frac{(1-\lambda)\lambda^{2t} f(x)^2}{2R + (1-\lambda)\lambda^{2t} f(x)^2},$$

and so

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

Example Sheet 1 Question 5 gives us a way to lower bound the total variation distance for two distributions if we can find a function $S \to \mathbb{R}$ such that the average variance of the function with respect to these distributions is small while the difference of the means is large enough. Therefore, we bound the variance of f with respect to both π and $P^t(x, \cdot)$ and use that its mean is 0 and $\lambda^t f(x)$, respectively with respect to these two distributions. $\text{Var}_{P^t(x,\cdot)}(f)$ can be bounded by inductively upper bounding the expectation of the $f(X_{t+1})^2$ given the value of X_t (in terms of the parameter R). The bound on $\text{Var}_\pi(f)$ is then obtained by taking limits and using the ergodic theorem.

Proof. As (f, λ) is an eigenpair we have that $\mathbb{E}_x[f(X_t)] = \lambda^t f(x)$ so using this and the condition of the lemma we have that for all $z \in S$

$$\mathbb{E}[f(X_{t+1}) - f(X_t) \mid X_t = z] = (\lambda - 1)f(z) \quad \text{and} \quad \mathbb{E}\left[(f(X_{t+1}) - f(X_t))^2 \,\middle|\, X_t = z\right] \leq R.$$

Therefore,

$$\mathbb{E}\left[f(X_{t+1})^2 \,\middle|\, X_t = z\right] = \mathbb{E}\left[((f(X_{t+1}) - f(X_t)) + f(X_t))^2 \,\middle|\, X_t = z\right] \leq R + 2(\lambda - 1)f(z)^2 + f(z)^2$$

and taking expectation over X_t gives

$$\mathbb{E}_x\left[f(X_{t+1})^2\right] \leq R + (2\lambda - 1)\mathbb{E}_x\left[f(X_t)^2\right].$$

Inductively, and using in the last step that $\lambda > 1/2$ we have

$$\mathbb{E}_x\left[f(X_t)^2\right] \leq R\frac{1 - (2\lambda - 1)^t}{1 - (2\lambda - 1)} + (2\lambda - 1)^t f(x)^2 \leq \frac{R}{2 - 2\lambda} + (2\lambda - 1)^t f(x)^2.$$

This gives, using that $\mathbb{E}_x[f(X_t)] = \lambda^t f(x)$ and that $2\lambda - 1 \leq \lambda^2$ that

$$\text{Var}_x(f(X_t)) \leq ((2\lambda - 1)^t - \lambda^{2t})f(x)^2 + \frac{R}{2 - 2\lambda} \leq \frac{R}{2 - 2\lambda}.$$

Taking limit as $t \to \infty$ and using that, since the chain is aperiodic, the ergodic theorem (Theorem 1.10) also gives the convergence in total variation [this is enough as $|S| < \infty$ gives that $\forall g : S \to \mathbb{R}\ \exists M > 0$ such that $\sup_{x \in S} |g(x)| \leq M$ and so using the existence of optimal coupling $|\mathbb{E}_\mu[g] - \mathbb{E}_\nu[g]\,| \leq 2M\|\mu - \nu\|_{\mathrm{TV}}$ and applying this to f and f^2 gives that convergence of variances is implied by convergence in total variation] gives that

$$\mathrm{Var}_\pi\,(f) \leq \frac{R}{2 - 2\lambda}.$$

As $\lambda \neq 1$ and $\lambda \mathbb{E}_\pi[f] = \mathbb{E}_\pi[Pf] = \mathbb{E}_\pi[f]$ we have $\mathbb{E}_\pi[f] = 0$ so

$$|\mathbb{E}_{P^t(x,\cdot)}[f] - \mathbb{E}_\pi[f]\,| = \lambda^t f(x).$$

For $\sigma^2 = \frac{\mathrm{Var}_{P^t(x,\cdot)}(f) + \mathrm{Var}_\pi(f)}{2} \leq \frac{R}{2-2\lambda}$ we have $|\mathbb{E}_{P^t(x,\cdot)}[f] - \mathbb{E}_\pi[f]\,| \geq r\sigma$ where $r^2 = \frac{2(1-\lambda)\lambda^{2t} f(x)^2}{R}$ and therefore from Example Sheet 1 Question 5 applied to distributions $P^t(x,\cdot)$ and π we have

$$\|P^t(x,\cdot) - \pi\|_{\mathrm{TV}} \geq \frac{r^2}{4 + r^2} = \frac{(1-\lambda)\lambda^{2t} f(x)^2}{2R + (1-\lambda)\lambda^{2t} f(x)^2}.$$

Above gives that $t_{\mathrm{mix}}(\varepsilon) \geq t$ where t is any number which satisfies $(1-\lambda)\lambda^{2t} f(x)^2 \frac{1-\varepsilon}{\varepsilon} \geq 2R$. Taking log of this equation and rearranging gives the bound of the mixing time from the statement of the lemma. □

Remark 4.19. **A very similar lower bound can be obtained in a more general case when eigenvalues and eigenfunctions are complex and without a condition $\lambda > \frac{1}{2}$.*

Remark 4.20. *Instead of using the bound from Example Sheet 1 Question 5 one can use the following weaker bound, which has an easy proof and which would give a similar result to the one above (which would still be good enough for most applications). Notice that for $f : S \to \mathbb{R}$, $\|\mu - \nu\|_{\mathrm{TV}} = \mathbb{P}(X \neq Y) \geq \mathbb{P}(f(X) \neq f(Y))$ where X and Y is an optimal coupling of μ and ν. We can bound $\mathbb{P}(f(X) \neq f(Y)) \geq \mathbb{P}\big(f(X) \in \big(\mathbb{E}_\mu[f] - \frac{\sigma r}{2}, \mathbb{E}_\mu[f] + \frac{\sigma r}{2}\big)\big) - \mathbb{P}\big(f(Y) \notin \big(\mathbb{E}_\nu[f] - \frac{\sigma r}{2}, \mathbb{E}_\nu[f] + \frac{\sigma r}{2}\big)\big)$, as from the definition of r the two sets we have in the previous expression are disjoint and for any $A \subset \mathbb{R}$ we have $\mathbb{P}(f(X) \neq f(Y)) \geq \mathbb{P}(f(X) \in A) - \mathbb{P}(f(Y) \in A)$. This is lower bounded using Chebychev's inequality and the definition of σ^2 by $1 - \frac{\mathrm{Var}_\mu(f) + \mathrm{Var}_\nu(f)}{(\sigma r/2)^2} \geq 1 - \frac{8}{r^2}$. To solve the example sheet question one needs to improve the lower bound on $\mathbb{P}(f(X) \neq f(Y))$.*

Lecture 8

We now apply this result to the lazy simple random walk on a hypercube.

Claim 4.21 (Lazy simple random walk on a hypercube has cutoff)**.** *Recall the definition of lazy simple random walk on the hypercube from Claim 2.12. If X^n is lazy simple random walks on n-dimensional hypercube then the sequence $(X^n)_{n \geq 1}$ exhibits cutoff at time $\frac{1}{2} n \log n$ with a window of order n. Moreover, the relaxation time is $t_{\mathrm{rel}} = n$.*

Upper bound: find all eigenvalues and eigenfunctions (think of the chain as acting on each coordinate with probability $1/n$ and get that the eigenfunctions are a product of eigenfunctions for the two-state chain on each coordinate), use that the hypercube is a transitive chain so we can bound the $\mathcal{L}^2$ as sum of powers of eigenvalues. Lower bound: apply Wilson's method to a suitable eigenfunction with eigenvalue $1 - \frac{1}{n}$.

Proof. Let functions $g_1(x) = 1$ and $g_2(x) = 1 - 2x$. We can then check that for $x \in \{0,1\}^n$ the function

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

where f_i is either g_1 or g_2 for all i is an eigenfunction of a lazy simple random walk on a hypercube corresponding to the eigenvalue $\lambda_I = \frac{n-|I|}{n}$ where $I = \{i \leq n : f_i = g_2\}$ is the set of indices i such that $f_i = g_2$. We can check that $f_I = \prod_{i \in I} g_2(x_i)$ are actually orthonormal eigenfunctions indexed over all subsets I of $\{1, \ldots, n\}$ (and as mentioned that the corresponding eigenvalues are $\lambda_I = \frac{n-|I|}{n}$). As the size of the state space is 2^n and as there are 2^n subsets of $\{1, \ldots n\}$, this gives us all eigenfunctions and the full orthonormal basis. When $I = \emptyset$, this gives the eigenvalue $\lambda_\emptyset = 1$ and for $|I| = 1$ we get $\lambda_* = \lambda_2 = 1 - \frac{1}{n}$, which implies that $t_{\text{rel}} = n$ (as the chain is lazy so $\gamma_* = \gamma$). Note that $\pi_{\min} = 2^{-n}$, and hence applying Theorem 4.10 gives

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

This is not a good bound, since we have already obtained a better upper bound in Claim 2.12 of order $n \log n$ using strong stationary times. However, using the full spectrum and not just the second eigenvalue we will now manage to 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 4.9 to get $x \in S$

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

Taking now $t = \frac{n \log n}{2} + Cn$ gives

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

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 need to show the matching lower bound. For this, we will use Wilson's method. We will consider eigenfunction $\widetilde{f}(x) = \sum_{i=1}^{n} (1 - 2x_i)$ which has eigenvalue $1 - \frac{1}{n} \geq \frac{1}{2}$. The difference in the value of this function for two neighbouring vertices has magnitude 2 and therefore we can use $R = 2$ in Wilson's method (as the chain moves with probability $1/2$ and stays in place otherwise) and using $x = (0, \ldots 0)$ gives that $\widetilde{f}(x) = n$ and so from Wilson's method we have

$$t_{\text{mix}}(\varepsilon) \geq \frac{1}{2 \log(1/(1 - 1/n))} \left(\log \left(\frac{(1/n)n^2}{4} \right) + \log \left(\frac{1 - \varepsilon}{\varepsilon} \right) \right).$$

Expanding $\log(1/(1 - 1/n)) = 1/n + O(1/n^2)$ the above gives that

$$t_{\text{mix}}(\varepsilon) \geq \frac{n}{2} \left(\log n - \log 4 + \log \left(\frac{1 - \varepsilon}{\varepsilon} \right) \right) + O(\log n) \geq \frac{1}{2} n \log n - C(\varepsilon) n$$

for large enough constant $C(\varepsilon)$ which doesn't depend on n and this completes the proof. □

5 Dirichlet form and the bottleneck ratio

5.1 Dirichlet form, variational characterisation and Poincaré constant

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

$$\langle f, g\rangle_\pi = \sum_{x\in S} f(x)g(x)\pi(x).$$

Definition 5.1 (Dirichlet form)**.** Let P be a transition matrix with invariant distribution π. The *Dirichlet form* associated to P and π is defined for all $f, g : S \to \mathbb{R}$ by

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

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

Claim 5.2 (Alternative expression of Dirichlet form)**.** *For all P we have*

$$\mathcal{E}(f) = \frac{1}{2}\mathbb{E}_\pi\big[(f(X_1) - f(X_0))^2\big] = \frac{1}{2}\sum_{x,y\in S} \pi(x)P(x, y)(f(x) - f(y))^2.$$

Moreover, if P is reversible we have that

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

Just expand the definition and use that $\pi P = \pi$ and $\sum_{y\in S} P(x, y) = 1$ and relabel the sums appropriately.

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

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

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

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

Therefore, in the reversible case, averaging the two expressions, we get

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

Even without reversibility, we have, using that $\pi P = \pi$ and $\sum_{y\in S} P(x, y) = 1$, that when $f = g$

$$\begin{aligned}
\mathcal{E}(f) &= \sum_{x,y\in S}(f(x)^2 - f(y)f(x))\pi(x)P(x, y) = \sum_{x\in S} f(x)^2\pi(x) - \sum_{x,y\in S} f(x)f(y)\pi(x)P(x, y)\\
&= \frac{1}{2}\left(\sum_{x\in S} f(x)^2\pi(x) + \sum_{y\in S} f(y)^2\pi(y) - 2\sum_{x,y\in S} f(x)f(y)\pi(x)P(x, y)\right)\\
&= \frac{1}{2}\left(\sum_{x,y\in S} f(x)^2\pi(x)P(x, y) + \sum_{x,y\in S} f(y)^2\pi(x)P(x, y) - 2\sum_{x,y\in S} f(x)f(y)\pi(x)P(x, y)\right)\\
&= \frac{1}{2}\sum_{x,y\in S}(f(x) - f(y))^2\pi(x)P(x, y)
\end{aligned}$$

which completes the proof. □

Theorem 5.3 (Variational characterisation of spectral gap for reversible chains)**.** *Let P be a reversible matrix with respect to π. Then the spectral gap $\gamma = 1 - \lambda_2$ can be expressed as*

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

Show that the second and third expressions are the same as the first one as we can do shifting and scaling and lower bound the first expression by decomposing f in the orthonormal basis of eigenfunctions. Notice that f_2 achieves $\mathcal{E}(f_2) = \gamma$.

Proof. From the definition of the Dirichlet form, we can see that $\mathcal{E}(f + c) = \mathcal{E}(f)$ for any constant $c \in \mathbb{R}$. Using this and the fact that $\|f - \mathbb{E}_\pi[f]\|_2 = \mathrm{Var}_\pi(f)$ gives the third equality and taking $\widehat{f}(x) = f(x)/\|f\|_2$ gives the second one. So we now prove the first equality. Let (f_j) be an orthonormal basis of eigenvectors of P for eigenvalues (λ_i) for the space $(\mathbb{R}^S, \langle\cdot,\cdot\rangle_\pi)$. 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

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

Taking f with $\|f\|_2 = 1$ gives that the last sum appearing above is equal to 1 as $1 = \langle f, f\rangle_\pi = \left\langle \sum_{j=2}\langle f, f_j\rangle_\pi f_j, f \right\rangle_\pi = \sum_{j=2}^{n}\langle f, f_j\rangle_\pi^2$, and hence proves that

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

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

We now want to generalise the Poincaré inequality to chains which are not time reversible. Instead of trying to get a bound of the $\mathcal{L}^2$ distance in terms of the relaxation time (i.e. the absolute spectral gap), we will define a Poincaré constant which in the case of reversible chains will correspond to the spectral gap.

Definition 5.4 (Poincaré constant)**.** Poincaré constant of a Markov chain is defined to be

$$\gamma(P) := \inf\left\{ \frac{\mathcal{E}(f)}{\mathrm{Var}_\pi(f)} \mid f : S \to \mathbb{R} \text{ is not constant} \right\}.$$

Since the ratio $\frac{\mathcal{E}(f)}{\mathrm{Var}_\pi(f)}$ is invariant under translation and scaling, we also have

$$\gamma(P) = \inf\{\mathcal{E}(f) \mid \|f\|_2 = 1, \mathbb{E}_\pi[f] = 0\},$$

which shows that the infimum is actually attained. //

Remark 5.5. *For the time reversal P^* we have that $\pi(x)P(x,y) = \pi(y)P^*(y,x)$ so*

$$\mathcal{E}_P(f) = \mathcal{E}_{P^*}(f) = \mathcal{E}_{\frac{P+P^\star}{2}}(f),$$

for all $f : S \to \mathbb{R}$, where $\mathcal{E}_Q$ is the Dirichlet form corresponding to the transition matrix Q. In particular, we deduce that

$$\gamma(P) = \gamma(P^*) = \gamma\left(\frac{P+P^\star}{2}\right).$$

In other words, the Dirichlet form and the Poincaré constant are invariant under time reversal.

Lemma 5.6 (Spectral interpretation of the Poincaré constant)**.** *We always have*

$$\gamma(P) = 1 - \lambda_2\left(\frac{P+P^*}{2}\right),$$

where $\lambda_2(Q)$ denotes the second largest eigenvalue of a reversible transition matrix Q.

Proof. This is a direct consequence of Theorem 5.3 as the chain $\frac{P+P^*}{2}$ is reversible. □

Definition 5.7 (Additive and multiplicative symmetrizations of the Markov chain)**.** For a Markov chain with transition matrix P and time reversal P^* *additive symmetrization* of P is the chain with transition matrix $\frac{P+P^*}{2}$ while the *multiplicative symmetrization* of P is the chain with transition matrix PP^*. It is easy to check that both symmetrizations are time reversible. //

From Example Sheet 2 Question 5b we have the following.

Lemma 5.8 (Variational contraction)**.** *For all $f : X \to \mathbb{R}$, we have*

$$\mathrm{Var}\,(Pf) \le [1 - \gamma(P^*P)]\mathrm{Var}\,(f).$$

Moreover, if P is lazy, then $\gamma(P^*P) \ge \gamma(P)$.

The above lemma shows that the variance of any observable decays exponentially fast under the repeated action of P, with rate $-\gamma(P^*P)$. We now show an upper bound on the mixing time which does not require reversibility.

Theorem 5.9 (Poincaré bound on the mixing time)**.** *If P is lazy, then for all $\varepsilon \in (0,1)$,*

$$t_{\mathrm{mix}}(\varepsilon) \le \frac{2}{\gamma(P)}\log\left(\frac{1}{2\varepsilon\sqrt{\pi_{\min}}}\right),$$

where $\pi_{\min} = \min_{x\in S}\pi(x)$ Moreover, for any P we have

$$t_{\mathrm{mix}}(\varepsilon) \le \frac{2}{\gamma(PP^*)}\log\left(\frac{1}{2\varepsilon\sqrt{\pi_{\min}}}\right).$$

Use the Lemma 5.8 and show that $\mathrm{Var}\left(P^{*t}f_x\right) = \|P^t(x,\cdot) - \pi\|_2^2$ where $f_x(z) = \frac{\mathbb{1}_{x=z}}{\pi(x)}$.

Proof. We recall that the $\mathcal{L}^p$ norm is increasing so as total variation can be expressed as $\mathcal{L}^1$ norm we have

$$\left\|P^t(x,\cdot) - \pi\right\|_{\mathrm{TV}} \le \frac{1}{2}\left\|\frac{P^t(x,\cdot)}{\pi(\cdot)} - 1\right\|_2.$$

In the reversible case, the existence of an orthonormal basis of eigenfunctions of P allowed us to prove the exponential decay of the right-hand side. Without reversibility, we no longer have an orthonormal basis of eigenfunctions at our disposal, but we can write

$$\frac{P^t(x,y)}{\pi(y)} = \frac{P^{*t}(y,x)}{\pi(x)} = \sum_{z\in S} P^{*t}(y,z)\frac{\mathbb{1}_{x=z}}{\pi(x)} = P^{*t}f_x(y),$$

where we have introduced the function $f_x : z \mapsto \frac{\mathbb{1}_{x=z}}{\pi(x)}$. Since we can check that $\pi P^{*t} f_x = \pi f_x = 1$, we see that

$$\begin{aligned}\left\| \frac{P^t(x,\cdot)}{\pi(\cdot)} - 1 \right\|_2^2 &= \sum_{y\in S}\left(\frac{P^t(x,y)}{\pi(y)}\right)^2 \pi(y) - 2\sum_{y\in S}\frac{P^t(x,y)}{\pi(y)}\pi(y) + \sum_{y\in S}\pi(y) \\ &= \sum_{y\in S}\left(\frac{P^t(x,y)}{\pi(y)}\right)^2 \pi(y) - 1 = \mathrm{Var}(P^{*t}f_x), \end{aligned} \tag{5.1}$$

and the right-hand side can be estimated using Lemma 5.8 and that the time reversal of a time reversal gives back the original chain (i.e. $P^{**} = P$) gives:

$$\mathrm{Var}(P^{\star t}f_x) \le \mathrm{Var}\left(f_x\right)\left(1-\gamma(PP^*)\right)^t = \left(\frac{1}{\pi(x)} - 1\right)\left(1-\gamma(PP^*)\right)^t \le \frac{e^{-\gamma(PP^*)t}}{\pi_{\min}}.$$

This all gives us that

$$\|P^t(x,\cdot) - \pi\|_{\mathrm{TV}} \le \frac{e^{-\frac{\gamma(PP^*)t}{2}}}{2\sqrt{\pi_{\min}}} \quad \forall t \in \mathbb{N},$$

which completes the proof in the case of a general P by using the definition of the mixing time. Now the fact that the Poincaré constant agrees for the chain and its time reversal together with the final statement of Lemma 5.8 gives that when P is lazy, which implies that P^* is lazy (as their diagonal terms are the same) we have

$$\|P^t(x,\cdot) - \pi\|_{\mathrm{TV}} \le \frac{e^{-\frac{\gamma(PP^*)t}{2}}}{2\sqrt{\pi_{\min}}} \le \frac{e^{-\frac{\gamma(P^*)t}{2}}}{2\sqrt{\pi_{\min}}} = \frac{e^{-\frac{\gamma(P)t}{2}}}{2\sqrt{\pi_{\min}}} \quad \forall t \in \mathbb{N}.$$

Again using the definition of mixing time completes the proof of the statement for lazy P. □

Remark 5.10. *In the reversible case the method used above could give us the proof of Theorem 4.10 if we used the fact that* $1-\gamma(PP^*) = \lambda_*^2$.

The previous results conclude that in the reversible case, the variational characterisation of the spectral gap gives that the Poincaré constant is the same as the spectral gap, which, when the chain is lazy, is the same as the inverse of the relaxation time. In the non-reversible case, the Poincaré constant corresponds to the spectral gap of the chain $(P+P^*)/2$, and while it does not bound the relaxation time of the chain, we can use it instead of the relaxation time to get a similar bounds on the mixing time.

We now show how one could bound the Poincaré constant in some cases and we note that in the reversible case, this also bounds the spectral gap.

Theorem 5.11 (Comparison of Poincaré constants)**.** *Let P and $\widetilde{P}$ be two transition matrices with invariant distributions π and $\widetilde{\pi}$ and Dirichlet forms $\mathcal{E}$ and $\widetilde{\mathcal{E}}$, respectively. Suppose that there exists a positive A such that $\widetilde{\mathcal{E}}(f) \le A\mathcal{E}(f)$ for all functions $f : S \to \mathbb{R}$. Let γ and $\widetilde{\gamma}$ be the Poincaré constants of P and $\widetilde{P}$, respectively. Then we have*

$$\widetilde{\gamma} \le \left(\max_{x\in S}\frac{\pi(x)}{\widetilde{\pi}(x)}\right) A\gamma.$$

Use that the variance of any variable X is a minimum of $\mathbb{E}\big[(X-a)^2\big]$ over a to upper bound the ratio of variances of f with respect to π and $\widetilde{\pi}$ by $\max_{x\in S}\frac{\pi(x)}{\widetilde{\pi}(x)}$.

Proof. From the definition of Poincaré constant (Definition 5.4) and the assumption we have

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

Since the variance of a random variable X is the minimum of $\mathbb{E}\big[(X-a)^2\big]$ over $a \in \mathbb{R}$ (because $-2a\mathbb{E}[X]+a^2$ is minimised when $a=\mathbb{E}[X]$), it follows that

$$\begin{aligned}\mathrm{Var}_{\pi}(f) &= \mathbb{E}_{\pi}\big[(f-\mathbb{E}_{\pi}[f])^2\big] \leq \mathbb{E}_{\pi}\big[(f-\mathbb{E}_{\widetilde{\pi}}[f])^2\big] = \sum_{x\in S}\pi(x)(f(x)-\mathbb{E}_{\widetilde{\pi}}[f])^2 \\ &= \sum_{x\in S}\frac{\pi(x)}{\widetilde{\pi}(x)}\widetilde{\pi}(x)(f(x)-\mathbb{E}_{\widetilde{\pi}}[f])^2 \leq \left(\max_{x\in S}\frac{\pi(x)}{\widetilde{\pi}(x)}\right)\mathrm{Var}_{\widetilde{\pi}}(f)\end{aligned}$$

Substituting this into the earlier inequality finishes the proof. □

Lecture 9

5.2 Canonical Paths

Here we present a technique for lower bounding the Poincaré constant, i.e. the spectral gap in the case of reversible chains.

Definition 5.12 (E-path)**.** For each x and y in the state space S we call $e=(x,y)$ an edge if $P(x,y)>0$ and we let $E=\{(x,y):P(x,y)>0\}$ be the set of all edges. For each $x,y\in S$ we call $\Gamma_{xy}=(e_1,\ldots,e_k)$ an *E-"path"* from x to y if $e_i=(x_{i-1},x_i)$ for $1<i<k$, $e_1=(x,x_1)$ and $e_k=(x_{k-1},y)$ and $e_i\in E$ for all $1\leq i\leq k$. We write $|\Gamma_{xy}|=k$ for the length of the path, i.e., the number of edges in it. For an edge $e=(a,b)$ we also write $Q(e)=\pi(a)P(a,b)$. //

Theorem 5.13 (Canonical path method)**.** *Let P be a transition matrix with invariant distribution π. For each $x,y\in S$ fix an E-"path" Γ_{xy}. 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} and the above sum is over all x,y such that $e\in\Gamma_{xy}$. Then the Poincaré constant of P, $\gamma=\gamma(P)$ satisfies $\gamma\geq 1/B$. In the case of the reversible chain, we know that γ is also the spectral gap.

Use Cauchy-Schwarz to show that $(f(x)-f(y))^2\leq|\Gamma_{xy}|\sum_{e\in\Gamma_{xy}}(\nabla f(e))^2$ where $\nabla f(e)$ is the difference of f along the endpoints of the edges. Express variance and Dirichlet form as a suitable sum of terms containing $(f(x)-f(y))^2$

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

$$\mathrm{Var}_{\pi}(f)=\frac{1}{2}\mathbb{E}\big[(f(X)-f(Y))^2\big]=\frac{1}{2}\sum_{x,y\in S}\pi(x)\pi(y)(f(x)-f(y))^2.$$

Using Cauchy-Schwarz we have that

$$(f(x)-f(y))^2 = \left(\sum_{e\in\Gamma_{xy}} \nabla f(e)\right)^2 \le \left(\sum_{e\in\Gamma_{xy}} 1^2\right)\left(\sum_{e\in\Gamma_{xy}} (\nabla f(e))^2\right) = |\Gamma_{xy}| \sum_{e\in\Gamma_{xy}} (\nabla f(e))^2,$$

which gives after swapping the order of the two sums that

$$\mathrm{Var}_\pi(f) \le \frac{1}{2}\sum_{e\in E} Q(e)(\nabla f(e))^2 \left(\frac{1}{Q(e)} \sum_{x,y:e\in\Gamma_{xy}} |\Gamma_{xy}|\pi(x)\pi(y)\right) \le \mathcal{E}(f)B.$$

The above holds from the definition of B and as an alternative expression of the Dirichlet form is $\mathcal{E}(f) = \frac{1}{2}\sum_{e\in E} Q(e)(\nabla f(e))^2$ and this completes the proof. □

Claim 5.14 (Relaxation time of a lazy SRW on the box)**.** *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 $t_{\mathrm{rel}} \le C(dn)^2$.*

Use Theorem 5.13 for paths which connect two vertices by first moving in the first coordinate, then in the second then in the third and so on.

Proof. For each x, y we choose Γ_{xy} to be the path that first changes the first coordinate of x until it becomes the same as the first coordinate of y and then changes the second coordinate until it becomes the same as the second coordinate of y and then the third one and so on until we reach y. 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}. This is because if $e = (a, b)$ where a and b differ only in coordinate i then we know that for $e \in \Gamma_{xy}$ to hold we need $x_j = a_j$ for $j > i$ and $y_j = b_j$ for $j < i$ so this gives us n^{d-1} options for coordinates of x, y which are not the i-th one and as there are at most n^2 options for the i-th coordinate the n^{d+1} bound follows. Also, the invariant distribution satisfies $\pi(x) \asymp 1/n^d$ and $Q(e) \asymp (dn^d)^{-1}$. Using also that $|\Gamma_{xy}| \le dn$ we bound the quantity B from Theorem 5.13 by

$$B \lesssim dn^d \cdot n^{d+1} \cdot nd \cdot \frac{1}{n^d} \cdot \frac{1}{n^d} = d^2n^2,$$

and this concludes the proof. □

5.3 Comparison technique

Similar to the variational characterisation of the second largest eigenvalue we have the following result for the reversible chains.

Theorem 5.15 (Characterisation of jth largest eigenvalue, reversible chain)**.** *Let P be reversible with respect to the invariant distribution π and let λ_j be its eigenvalues with $1 = \lambda_1 > \lambda_2 \ge \ldots \ge \lambda_n$. Then for all $j \in \{1, \ldots, n\}$ we have*[4]

$$1 - \lambda_j = \max_{\phi_1,\ldots,\phi_{j-1}} \min\left\{\mathcal{E}(f) : \|f\|_2 = 1, f \perp \phi_1, \ldots, \phi_{j-1}\right\}.$$

For any ϕ_i there is a function in the span of $f_1 \ldots f_j$ which is orthogonal to $\phi_1, \ldots \phi_{j-1}$ and that function (normalised) has Dirichlet form upper bounded by $1-\lambda_j$. Maximum is attained by $\phi_i = f_i$.

[4] $\perp$ means perpendicular with respect to the inner product with respect to π, i.e. $f \perp g$ means $\langle f, g\rangle_\pi = 0$

Proof. Let (f_j) be the eigenfunctions corresponding to the eigenvalues (λ_j). Let $\phi_1, \ldots, \phi_{j-1}$ be arbitrary functions. Consider $W = \operatorname{span}(\phi_1, \ldots, \phi_{j-1})^\perp$. Then $\dim(W) \geq n - j + 1$, and hence $W \cap \operatorname{span}(f_1, \ldots, f_j) \neq \emptyset$. So there exists g in the intersection. By normalizing, we can assume that $\|g\|_2 = 1$. Let $g = \sum_{i=1}^{j} a_i f_i$. Then $\sum_{i=1}^{j} a_i^2 = 1$ holds as $\|g\|_2 = 1$ and we have

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

Finally, taking $\phi_i = f_i$ for all $i \leq j-1$ gives the equality, as orthonormality of f_i means that $W = \operatorname{span}(f_j, \ldots, f_n)$ so any function g in W with $\|g\|_2 = 1$ can be written as $g = \sum_{i=j}^{n} a_i f_i$ with $\sum_{i=j}^{n} a_i^2 = 1$ and using the calculations as above has $\mathcal{E}(g) = \sum_{i=j}^{n} a_i^2(1-\lambda_i) \geq 1 - \lambda_j$. □

In case the chain is not reversible, as additive symmetrisation $\frac{P+P^*}{2}$ is always reversible then we instead have

$$1 - \lambda_j\left(\frac{P+P^*}{2}\right) = \max_{\phi_1, \ldots, \phi_{j-1}} \min\left\{\mathcal{E}(f) : \|f\|_2 = 1, f \perp \phi_1, \ldots, \phi_{j-1}\right\}.$$

where $\lambda_j(\frac{P+P^*}{2})$ is the jth largest eigenvalue of $\frac{P+P^*}{2}$.

Corollary 5.16 (Comparison of the jth largest eigenvalues)**.** *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 (λ_i) and $(\widetilde{\lambda}_i)$ their respective eigenvalues (in decreasing order). If there exists a positive constant A such that for all $f : S \to \mathbb{R}$ we have $\widetilde{\mathcal{E}}(f) \leq A\mathcal{E}(f)$, then $1 - \widetilde{\lambda}_j \leq A(1-\lambda_j)$ for all j.*

We have seen that if we can compare Dirichlet forms corresponding to two transition matrices which are reversible with respect to the same invariant distribution then we can also compare all eigenvalues of them. We now state the result which gives a general strategy for comparing Dirichlet forms:

Theorem 5.17 (Comparison technique)**.** *Let P and $\widetilde{P}$ be two transition matrices with invariant distributions π and $\widetilde{\pi}$ respectively, and let edges corresponding to these transition matrices be E and $\widetilde{E}$, respectively. Suppose that for each $(x,y) \in \widetilde{E}$ we fix a path Γ_{xy} in E and we set*

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

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

Use Cauchy-Schwarz to bound $(f(x) - f(y))^2$ by sum of $|\Gamma_{x,y}| \sum_{e \in \Gamma_{x,y}} (\nabla f(e))^2$.

Proof. This proof is very similar to the proof of Theorem 5.13 and one can actually prove Theorem 5.13 by applying this result to $\widetilde{P}(x,y) = \pi(y)$.We have

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

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

5.4 Bottleneck ratio

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 5.18 (Bottleneck ratio)**.** The *bottleneck ratio* (also known as *Cheeger constant*) is defined to be

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

We also define a *conductance* of a set $A \subset S$ to be $\Phi(A) = \frac{Q(A,A^c)}{\pi(A)}$. //

Lecture 10

Theorem 5.19 (Inverse of bottleneck ratio lower bounds the mixing time)**.** *For any irreducible transition matrix P we have*

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

Lower bound $\pi(A^c) - \mathbb{P}_{\pi_A}(X_t \in A^c) \ge 1/2 - t\Phi(A)$ by upper bounding $\mathbb{P}_{\pi_A}(X_t \in A^c)$ using the union bound and bounding the probability that the chain starting from π goes from A to A^c in two consecutive steps.

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

$$\mathbb{P}_\pi(X_0 \in A, X_t \in A^c) \le \sum_{i=0}^{t-1} \mathbb{P}_\pi(X_i \in A, X_{i+1} \in A^c) = t\mathbb{P}_\pi(X_0 \in A, X_1 \in A^c) = tQ(A,A^c).$$

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

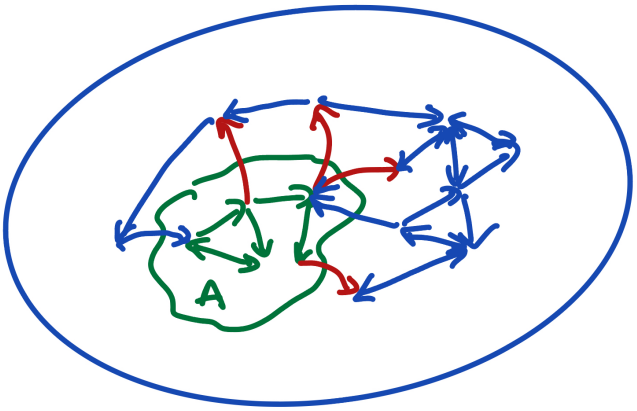
$$\mathbb{P}_{\pi_A}(X_t \in A^c) = \frac{\mathbb{P}_\pi(X_0 \in A, X_t \in A^c)}{\pi(A)} \le t\frac{Q(A,A^c)}{\pi(A)} = t\Phi_*,$$

where π_A is π restricted to A and normalised to be a distribution. Taking now $t = (4\Phi^*)^{-1}$ gives $\mathbb{P}_{\pi_A}(X_t \in A^c) \le 1/4$, and therefore [using that from convexity of TV distance $d(t) \ge \|\mu P^t - \pi\|$ for any distribution μ on S],

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

which completes the proof. □

One should think about the bottleneck ratio as a way to describe how hard it is to exit a set if you start from π within that set. In the following picture, the conductance of A would correspond to the probability of crossing a red edge when the chain starts from π conditional on starting in the set A (the directed edges going from vertices x to y indicate that $P(x,y) > 0$ and one should think like this about Markov chains in general). Indeed, if we write $\pi_A(x) = \frac{\pi(x)}{\pi(A)}\mathbb{1}_A(x)$ then we have that $\mathbb{P}_{\pi_A}(X_1 \in A^c) = \sum_{x\in S, y\in A^c} \pi_A(x)P(x,y) = \sum_{x\in A, y\notin A} \frac{\pi(x)P(x,y)}{\pi(A)} = \frac{Q(A,A^c)}{\pi(A)} = \Phi(A)$.



Theorem 5.20 (Cheeger's inequality)**.** *Let P be a reversible transition matrix with respect to the invariant distribution π. Let γ be the spectral gap. Then we have*

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

Moreover, for non-reversible matrix, by considering additive symmetrization $\frac{P+P^*}{2}$ and recalling from Example Sheet 3 Question 4 that it has the same bottleneck ratio as P, we have that

$$\frac{\Phi_*^2}{2} \leq \gamma(P) \leq 2\Phi_*,$$

where we know that $\gamma(P) = \gamma\left(\frac{P+P^*}{2}\right)$ is the Poincaré constant.

Upper bound: show that $f(x) = -\pi(A^c)\mathbb{1}_A(x) + \pi(A)\mathbb{1}_{A^c}(x)$ satisfies $\frac{\mathcal{E}(f)}{\mathrm{Var}_\pi(f)} = \frac{Q(A,A^c)}{\pi(A)\pi(A^c)}$. Lower bound: The proof follows easily after proving *step 1*, where you find a non-negative function f such that $\pi(f > 0) \leq \frac{1}{2}$ and $\gamma \geq \frac{\mathcal{E}(f)}{\|f\|_2^2}$, and *step 2*, where you show that for any function $\psi \geq 0$ we have $\mathcal{E}(\psi) \geq \frac{h(\psi)^2}{2}\|\psi\|_2^2$ where $h(\psi) = \inf_{\emptyset \neq A \subseteq \{x:\psi(x)>0\}} \frac{Q(A,A^c)}{\pi(A)}$. For the *step 1* consider $f = \max(0, f_2)$ where the eigenvector f_2 can WLOG be taken to have $\pi(f_2 > 0) \leq 1/2$. For the *step 2* one bounds $h(\psi)\|\psi\|_2^2 = h(\psi)\sum_{x\in S}\pi(x)\int_0^{\psi(x)} 2t dt \leq \sum_{\psi(x)<\psi(y)}(\psi(y)^2 - \psi(x)^2)Q(x,y) \leq \sqrt{2\mathcal{E}(\psi)}\|\psi\|_2$ where the second inequality follows by Cauchy-Schwarz and the first one follows by suitably exchanging sum and the integral and using that $\pi(S_t)h(\psi) \leq Q(S_t^c, S_t)$ where $S_t = \{x : \phi(x) > t\}$.

Proof. We start with the easy direction which is the upper bound. Using the variational characterisation of γ from Theorem 5.3 we get

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

Let $A \subset S$ be a set with $\pi(A) \leq 1/2$. Taking now $f(x) = -\pi(A^c)$ for $x \in A$ and $f(x) = \pi(A)$ for $x \in A^c$, we obtain using that $\pi(A^c) > \frac{1}{2}$ and that $\pi(A) + \pi(A^c) = 1$ [we are also using that for reversible chain from detailed balance equations we have $Q(a,b) = Q(b,a)$ for any two states $a, b \in S$ and so also $Q(A,B) = Q(B,A)$]

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

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

We next turn to the lower bound. The proof will consist of two main ideas which we will prove below as two steps. We first say what these two steps show and see how the proof follows assuming these.

In *step 1* we will show that there is a non-negative function f such that $\pi(f > 0) \leq \frac{1}{2}$

$$\gamma \geq \frac{\mathcal{E}(f)}{\|f\|_2^2}. \tag{5.2}$$

In the *step 2* we will show that for any function $\psi \geq 0$

$$\mathcal{E}(\psi) \geq \frac{h(\psi)^2}{2}\|\psi\|_2^2, \tag{5.3}$$

where for any non-negative function ψ define $S(\psi) = \{x : \psi(x) > 0\}$ and

$$h(\psi) = \inf_{\emptyset \neq A \subseteq S(\psi)} \frac{Q(A, A^c)}{\pi(A)}.$$

Since from *step 1* $\pi(f > 0) \leq 1/2$, using the definition of the bottleneck ratio and the definition of $h(f)$ we deduce $h(f) \geq \Phi_*$. This now combined with (5.2) and (5.3) gives

$$\gamma \geq \frac{\mathcal{E}(f)}{\|f\|_2^2} \geq \frac{h(f)^2}{2} \geq \frac{\Phi_*^2}{2}$$

which completes the proof of the theorem upon showing that the claims from the two steps hold. We now prove the claims from the two 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

$$((I - P)f)(x) \leq \gamma f(x).$$

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

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

where the inequality follows since $f \geq f_2$. Using that $f \geq 0$ and $((I - P)f)(x) \leq \gamma f(x)$ gives

$$\langle (I - P)f, f\rangle_\pi \leq \gamma \langle f, f\rangle_\pi \implies \gamma \geq \frac{\langle (I - P)f, f\rangle_\pi}{\langle f, f\rangle_\pi} = \frac{\mathcal{E}(f)}{\|f\|_2^2}$$

which as $\pi(f > 0) = \pi(f_2 > 0) \leq 1/2$ proves the claim from *step 1*.

2nd step: 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 $S_t \neq \emptyset$ we have (using that for reversible chain $Q(A, A^c) = Q(A^c, A)$ clearly holds and in Example Sheet 3 Question 4 we will see it holds without reversibility assumption)

$$\pi(S_t)h(\psi) \leq Q(S_t, S_t^c) = Q(S_t^c, S_t) = \sum_{\substack{\psi(x) \leq t \\ \psi(y) > t}} \frac{Q(x, y) + Q(y, x)}{2}.$$

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

$$\begin{aligned}
h(\psi)\|\psi\|_2^2 &= h(\psi)\sum_{x\in S}\psi(x)^2\pi(x) = h(\psi)\sum_{x\in S}\pi(x)\int_0^{\psi(x)}2tdt = h(\psi)\int_0^\infty 2t\sum_{x\in S:\psi(x)>t}\pi(x)dt \\
&= h(\psi)\int_0^\infty 2t\pi(\{x\in S:\psi(x)>t\})dt \le \int_0^\infty 2t\sum_{\substack{\psi(x)\le t\\ \psi(y)>t}}\frac{Q(x,y)+Q(y,x)}{2}dt \\
&= \sum_{\psi(x)<\psi(y)}(\psi(y)^2-\psi(x)^2)\frac{Q(x,y)+Q(y,x)}{2} = \frac{1}{2}\sum_{x,y\in S}|\psi(x)^2-\psi(y)^2|Q(x,y) \\
&\le \frac{1}{2}\sqrt{\sum_{x,y\in S}(\psi(x)-\psi(y))^2Q(x,y)\cdot\sum_{x,y\in S}(\psi(x)+\psi(y))^2Q(x,y)} \\
&\le \frac{1}{2}\sqrt{2\mathcal{E}(\psi)}\cdot\sqrt{2\left(\sum_{x,y\in S}\psi(x)^2\pi(x)P(x,y)+\sum_{x,y\in S}\psi(y)^2\pi(x)P(x,y)\right)} \\
&= \frac{1}{2}\sqrt{2\mathcal{E}(\psi)}\cdot 2\|\psi\|_2 = \sqrt{2\mathcal{E}(\psi)}\|\psi\|_2,
\end{aligned}$$

where for the inequality on the fourth line, we used Cauchy-Schwarz and for the final inequality, we used $(a+b)^2\le 2(a^2+b^2)$. Rearranging proves (5.3) and completes the proof the second step and of this theorem. □

Remark 5.21. *The bounds in Theorem 5.20 are the best one could hope for. Indeed, both bounds are achieved at least up to a constant factor in the following two simple examples.*

The lower bound can be seen to be sharp up to a constant factor for the lazy simple random walk on the cycle $\mathbb{Z}_n$. Indeed, Claim 4.17 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 following set with $\pi(A)=1/2$

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

then $\Phi(A)=1/(2n)$ (from every vertex in A there is a probability $1/(2n)$ to change the first coordinate to 1), which gives that $\Phi_*\le 1/(2n)$. We proved (Claim 4.21) that $\gamma=1/t_{\text{rel}}=1/n$, and hence $\Phi_*\ge 1/(2n)$, showing that $\Phi_*=1/(2n)$, and therefore the upper bound is sharp in this case.

5.5 Expander graphs

We will call $\Phi_*(G)$ a bottleneck ratio of a simple random walk on a graph G. We notice that if $G=(V,E)$ for $A\subset V$ we have that for P, a simple random walk on G, conductance of A is

$$\Phi(A)=\frac{\sum_{x\in A,y\notin A}\pi(x)P(x,y)}{\sum_{x\in A}\pi(x)}=\frac{\sum_{x\in A y\notin A}\frac{\deg(x)}{2|E|}\frac{\mathbb{1}_{\{x\sim y\}}}{\deg(x)}}{\sum_{x\in A}\frac{\deg(x)}{2|E|}}=\frac{|\{(x,y):x\sim y,x\in A,y\notin A\}|}{\sum_{x\in A}\deg(x)}$$

where $x\sim y$ means that x and y are neighbours. This tells us that $\Phi(A)$ is simply the number of edges between A and A^c divided by the sum of degrees of all vertices in A.

Expanders are just sequences of graphs which increase in size, have degrees bounded by some absolute constant and have bottleneck ratio lower bounded away from 0. The formal definition is as follows.

Definition 5.22 (Expanders)**.** A sequence of graphs $G_n = (V_n, E_n)$ is called a (d, α)-*regular expander family* if $\lim_{n\to\infty} |V_n| = \infty$, for each n the graph G_n is d-regular and the bottleneck ratios satisfy $\Phi_*(G_n) \geq \alpha$ for all n, where $\Phi_*(G)$ is the bottleneck ratio of the simple random walk on graph G. We call a sequence of graphs $G_n = (V_n, E_n)$ a (Δ, α)-*expander family* if that degrees of all graphs are bounded by Δ, if $\lim_{n\to\infty} |V_n| = \infty$ and if the bottleneck ratios satisfy $\Phi_*(G_n) \geq \alpha$ for all n. We say that a sequence of graphs is α-*expander family* or α-*expanders* if there is a constant Δ such that the sequence is a (Δ, α)-*expander family* and we simply call a sequence of graphs *expanders* if there is a constant α such that the graphs are α-expanders. //

From the above definition, it is not clear that such a family of graphs even exists, as it could be the case that all large enough graphs which have a degree bounded by some absolute constant (not depending on the size of the graph) need to have bottleneck ratio converging to 0 [i.e. it is not clear that one can construct an arbitrarily large bounded degree graph for which all sets of size roughly at most one half of the number of vertices have constant proportion of neighbours outside of the set]. We will first show a bound on the mixing times of expanders, to see why it is useful in the context of mixing times to know that something is an expander. Then we will find an example of an expander family, thus showing that such a family exists.

Proposition 5.23 (Mixing time of expanders)**.** *Let G_n be a α-expander family. Then the mixing time of the lazy simple random walk on G_n satisfies $t_{\text{mix}} = O(\log |V(G_n)|)$.*

Use Cheegers inequality and the bound of the mixing time in terms of the relaxation time as well as that we know how to relate the spectral gap for a chain and its lazy version.

Proof. Applying Theorem 5.20 we get that for $\alpha-$expanders $\gamma_{\text{SRW}} \geq \alpha^2/2$ (where γ_{SRW} is the spectral gap for a simple random walk on G_n). As the spectral gap of a lazy simple random walk is just half of the spectral gap for a simple random walk [$\gamma_{\text{SRW}} = 1 - \lambda_2^{\text{SRW}} = 2(1 - \frac{1+\lambda_2^{\text{SRW}}}{2}) = 2\gamma_{\text{LSRW}}$, where γ_{LSRW} is the spectral gap for the lazy simple random walk], and as the spectral gap and the absolute spectral gap are the same for any lazy Markov chain, we get that the upper bound follows directly from Theorem 4.10 [using that $\pi_{\min} \geq \frac{1}{\Delta |V_n|}$ where Δ is the absolute constant from the definition of expanders bounding the degree]. □

Lecture 11

Claim 5.24 (Expanders have fastest mixing times)**.** *Expander graphs of bounded degree have the fastest mixing time (up to constants) among all regular graphs.*

Diameter lower bound from ES1 Question 7 gives this, together with the fact that bounded degree graphs have at most exponential growth.

Proof. The diameter of a graph of degree bounded by Δ is at least (up to constants) $\log |V(G_n)|$. This is because for a fixed vertex x the number of vertices on distance r from x is at most Δ^r [as we can look at vertices on distance k from x and see that each of them gives at most Δ new neighbours with distance $k+1$ from x] so the diameter D needs to satisfy that $\Delta^D \geq |V(G_n)|$ (as any vertex is at most D away from x). The claim then follows from Example Sheet 1, Question 7 (diameter lower bound on the mixing time). □

Here we will show that there is a constant α (which does not depend on n) such that a certain randomly generated 3-regular graph on $\geq n$ vertices is an α-expander with high probability (i.e. has bottleneck ratio lower bounded by α). This in fact means that for any n there are graphs on

at least n vertices with bottleneck ratio lower bounded by this α and thus an expander sequence exists.

Theorem 5.25 (Existence of expander graphs)**.** *There exists a positive constant c such that there is a random (multi) graph on n vertices with $\mathbb{P}(\Phi_* > c) = 1 - o(1)$ as $n \to \infty$. Moreover, there is a positive constant $c > 0$ such that for all $n \in \mathbb{N}$ there exists a simple 3 regular graph without multiple edges on at least n vertices for which $\Phi_* > c$.*

One gets a random graph by considering two sets A and B of size n and then taking three uniformly at random picked perfect matchings between them (i.e. make every vertex $a_i \in A$ correspond to a vertex $b_i \in B$, connect the corresponding vertices and then add connections $(a_i, b_{\sigma_1(i)})$ and $(a_i, b_{\sigma_2(i)})$ for two independent uniform permutations σ_1, σ_2). By union bound over the sets $X \subset A$ of size at most $n/2$ and possible neighbours of X show that the probability that there is a subset of A of size $k \leq n/2$ that has $\leq (1+\delta)k$ neighbours goes to 0 as $n \to \infty$ for small enough $\delta > 0$. In the same way this holds for subsets of B and now working on these events one can show that the obtained graph is a $\frac{\delta}{6}$-expander [WLOG for some $|X| \leq \frac{n}{2}$ if $|A \cap X| \geq |B \cap X|$ analyse the number of neighbours of $A \cap X$ in $B \cap X^c$ separately in the two cases where $|A \cap X| \leq \frac{n}{2}$ and when $|A \cap X| \geq \frac{n}{2}$]. One then modifies this graph to be 3 regular simple graph by showing there are likely no triple edges in it and very few double edges and that subdividing double edges and connecting their midpoint can not change the bottleneck ratio significantly.

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 uniform permutations of $\{1, \ldots, n\}$. For every i we place three edges

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

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

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

Indeed, $|N(X)|$ is at least k (as it has the vertices of the same index in B). 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 which includes vertices in B with the same index as the ones in X and additional δk vertices which we can pick in at most $\binom{n}{\delta k}$. So we have

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

Using a union bound again over all subsets X of size $\leq n/2$ we obtain

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

One now has to estimate the above quantity and show that for a sufficiently small constant δ it converges to 0 as $n \to \infty$ and this is left as an exercise (see Question 5 from Example Sheet 3).

So this showed that all subsets $X \subseteq A$ 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 for this graph $\Phi_* > \delta/6$.

We will now work on the high probability event above that both subsets of A and subsets of B of size $k \leq n/2$ have at least $(1+\delta)k$ neighbours for small enough constant δ which does not depend on n.

Let $X \subseteq A \cup B$ of size $|X| \leq n$. We write $A' = X \cap A$ and $B' = X \cap B$. Wlog suppose that $|A'| \geq |B'|$. Then this implies that $|A'| \geq |X|/2$.

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 $|B'| \leq |X|/2 \leq |A'|$, it follows that A' will have at least $\delta|X|/2$ neighbours in $B \setminus B'$. These must be edges connecting X to X^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 $\delta|X|/2$ neighbours in $B \setminus B'$ and the corresponding edges will connect X to X^c.

Therefore, in either case, the bottleneck ratio is lower bounded by $\delta/2 \cdot 1/3$ as the probability of jumping along any edge is $1/3$.

The graph produced by Pinsker's method is a multigraph. The expected number of triple edges is $n \cdot 1/n^2 = 1/n$ as both σ_1 and σ_2 have to have i as a fixed point in order to have a triple edge between a_i and b_i. The expected number of double edges is $n\left(\frac{2}{n}\left(1-\frac{1}{n}\right) + (n-1)\frac{1}{n^2}\right) \leq 3$ where the first term comes from the cases where a_i has a double edge with b_i and the second term by finding the probability that a_i has a double edge with one of the possible $n-1$ vertices b_j for $j \neq i$. By Markov's inequality [$\mathbb{P}$(number of triple edges $\geq 1) \leq 1/n$ and $\mathbb{P}$(number of double edges $\geq 4) \leq 3/4$] and the high probability bound we have on the bottleneck ratio of this random graph, we have that

$$\mathbb{P}\left(\Phi_* \geq \frac{\delta}{6}, \text{ no triple edges and no more than 3 double edges}\right) \geq 1 - \frac{3}{4} - \frac{1}{n} - o(1) \geq \frac{1}{4} - o(1)$$

and so there exists a graph on n vertices having at most 3 double edges and where all vertices are of degree 3 and $\Phi_* \geq \frac{\delta}{6}$. We can now subdivide every double edge by placing a vertex in the middle of each of the two edges and connecting the two new vertices and creating a 3-regular graph. One now needs to check a few cases to see that the bottleneck ratio will be multiplied in the worst-case scenario by $2/5$. [Loosely speaking, the worst case for the change of $\Phi(X)$ is if the set X in the new graph contains both endpoints of previous double edges and the middle points of these previous double edges for some pair of double edges. This makes what was previously two edges (and so a sum of degrees of the two vertices 4) into 5 edges (and so a sum of degrees of these 4 vertices 10) while keeping the number of neighbours of our set the same. The bottleneck ratio is the total number of edges coming out of the set (so a number of neighbours, as it is now a simple graph) divided by the sum of degrees of all vertices in the set, and as we have increased the sum of degrees by a factor of at most $5/2$ and kept the number of neighbours the same the bottleneck ratio is multiplied by at least $2/5$. To have a full formal proof one needs to check the cases of sets containing any combination of endpoints and midpoints of previously double edges and compare that to the bottleneck ratio of the set with just one or both of the endpoints in the multigraph.] □

There are actually many other random graph models which generate expander sequences with large probabilities. One well-known such model is a configuration model with degrees at least 3 and at most some constant Δ, where one fixes a degree for each vertex and then generates a (multi) graph by making connections between vertices uniformly at random (while keeping the degrees fixed.

In the final part of the previous theorem, we have changed an expander sequence of graphs by adding at most 8 vertices (which also changed connections from additional 8 vertices) while keeping degrees 3 and the graph connected, and stated that this modification will still give us an α-expander sequence, but with a smaller constant α. One can check that for any expander sequence and a constant k, if each of the graphs in the sequence is changed in at most k vertices (vertex is changed if any of its connections are changed, if it is deleted or if it is a new vertex which was added), such that the graphs stay connected and still have bounded degrees then the sequence of graphs one gets after the modification is still an expander sequence (but possibly for smaller α which also depends on k).

6 Spectral and isoperimetric profile

Here we prove some results which generalise Theorem 4.10 and 5.20.

6.1 Spectral profile

We start by defining the spectral profile, which is a generalisation of a Poincaré constant, i.e. in the reversible case of the spectral gap. The idea is that in the definition of Poincaré constant (Definition 5.4) instead of taking infimum over all functions from the state space to $\mathbb{R}$ one could fix $r \leq 1$ and take infimum over non-negative functions with the support of π measure at most r. After stating the formal definition of the spectral profile, we will also prove a bound on the mixing time using it. It is standard to write Λ for the spectral profile and we will use the standard notation here, even though it might be confusing, as it generalises $\gamma = 1 - \lambda_2$.

Definition 6.1 (Spectral profile)**.** For any transition matrix P with invariant distribution π and a non-empty set $A \subset S$ we define

$$\lambda_P(A) = \lambda(A) = \inf_{f \in c_0^+(A)} \frac{\mathcal{E}(f)}{\mathrm{Var}_\pi(f)},$$

where $c_0^+(A) = \{f : \mathrm{supp}(f) \subset A, f \geq 0, f \text{ is not constant}\}$, where $\mathrm{supp}(f)$ denotes the support of f, i.e. $\mathrm{supp}(f) \subset A$ means that f is 0 on A^c. We then define the *spectral profile* (of P) to be the function $\Lambda_P : [\pi_{\min}, \infty) \to \mathbb{R}$ by

$$\Lambda_P(r) = \Lambda(r) = \inf_{A : \pi_{\min} \leq \pi(A) \leq r} \lambda(A).$$

//

We will drop P in the subscript whenever it is clear which transition matrix the spectral profile corresponds to. We can right away notice that $\Lambda(m) = \Lambda(1)$ for any $m \geq 1$, that Λ is non-increasing, that for any $A \subset S$

$$\lambda(A) \geq \Lambda(\pi(A))$$

and that for all r we have by the definition of Poincaré constant that

$$\gamma(P) \leq \Lambda(r).$$

In Example Sheet 3 Question 6 you will also show that

$$\Lambda\left(\frac{1}{2}\right) \leq 2\gamma(P)$$

and so if we want to get a bound on the mixing time using spectral profile rather than just simply the Poincaré constant (i.e. spectral gap for reversible chains) the improvement would come from the behaviour of spectral profile for small r.

Lecture 12

Theorem 6.2 (Bounding $\mathcal{L}^2$ distance via the spectral profile)**.** *For a transition matrix P with invariant distribution π satisfying that PP^* and P^*P are irreducible*[5] *we have that for all $x \in S, t \in \mathbb{N}$*

$$d_2(t)^2 \leq \frac{4}{V_{PP^*}\left(\frac{t}{2}\right)},$$

where for any transition matrix Q with invariant distribution μ the function $V_Q : [0, \infty) \to \mathbb{R}$ is given by

$$t = \int_{4\mu_{\min}}^{V_Q(t)} \frac{dv}{v\Lambda_Q(v)}.$$

To motivate the proof of this we notice that when proving the mixing time bound using the Poincaré constant we rely on Lemma 5.8 and we apply it for transition matrix P^* to $P^{*t}f_x$ (recall proof of Theorem 5.9) for different values of t. But this seems particularly wasteful as for smaller values of t the bound is not sharp (as the support of the function $P^{*t}f_x$ couldn't be too large). So we will first improve the result from Lemma 5.8.

Lemma 6.3 (Variational contraction by spectral profile)**.** *For every non-constant function $f : S \to \mathbb{R}_+$ we have*

$$\frac{\mathcal{E}(f)}{\mathrm{Var}_\pi(f)} \geq \frac{1}{2}\Lambda\left(\frac{4(\mathbb{E}_\pi[f])^2}{\mathrm{Var}_\pi(f)}\right).$$

This further implies that

$$\mathrm{Var}_\pi(P^*f) - \mathrm{Var}_\pi(f) \leq -\frac{1}{2}\mathrm{Var}_\pi(f)\,\Lambda_{PP^*}\left(\frac{4(\mathbb{E}_\pi[f])^2}{\mathrm{Var}_\pi(f)}\right).$$

For the second part, notice that the left-hand side is $\langle P^*f, P^*f\rangle_\pi - \langle f, f\rangle_\pi = -\mathcal{E}_{PP^*}(f)$. For the first part $\mathcal{E}(f) \geq \mathcal{E}((f-c)^+) \geq \mathrm{Var}_\pi((f-c)^+)\,\Lambda(\pi(f > c))$ and $\mathrm{Var}_\pi((f-c)^+) \geq \mathrm{Var}_\pi(f) - 2c\mathbb{E}_\pi[f]$ gives the result by taking $c = \frac{\mathrm{Var}_\pi(f)}{4\mathbb{E}_\pi[f]}$.

Proof. We first notice that the second statement of this lemma follows from the first one as PP^* has invariant distribution π and that from the definition $P^*(x,y)\pi(x) = P(y,x)\pi(y)$ so we have that $\langle f, P^*g\rangle_\pi = \sum_{x,y\in S} f(x)P^*(x,y)g(y)\pi(x) = \sum_{x,y\in S} f(x)P(y,x)g(y)\pi(y) = \langle Pf, g\rangle_\pi$, and so as $\mathbb{E}_\pi[f] = \mathbb{E}_\pi[P^*f]$ we have

$$\mathrm{Var}_\pi(P^*f) - \mathrm{Var}_\pi(f) = \langle P^*f, P^*f\rangle_\pi - \langle f, f\rangle_\pi = -\langle (I - PP^*)f, f\rangle_\pi = -\mathcal{E}_{PP^*}(f).$$

We now show the first inequality. Recall that $a^+ = \max\{a, 0\}$.

For any constant c we know that $\mathcal{E}(f, f) = \mathcal{E}(f - c, f - c)$. Also, $\forall a, b \in \mathbb{R} : (a-b)^2 \geq (a^+ - b^+)^2$, so $\mathcal{E}(f) \geq \mathcal{E}(f^+)$. It follows that when $0 \leq c < \max f$, then

$$\begin{aligned}\mathcal{E}(f) \geq \mathcal{E}((f-c)^+) &\geq \mathrm{Var}_\pi\left((f-c)^+\right) \inf_{g \in c_0^+(\{x : f(x) > c\})} \frac{\mathcal{E}(g)}{\mathrm{Var}_\pi(g)} \\ &\geq \mathrm{Var}_\pi\left((f-c)^+\right)\Lambda(\pi(f > c)).\end{aligned}$$

[5] for this result was only need that PP^* is irreducible so that $\Lambda_{PP^*}(r) > 0$ for all r, but we will apply this result to P and to P^* chain which will require that both PP^* and P^*P are irreducible

Now, $\forall a, b \geq 0 : (a-b)^{+2} \geq a^2 - 2ba$ and $(a-b)^+ \leq a$, so

$$\mathrm{Var}_\pi\left((f-c)^+\right) = \mathbb{E}_\pi\left[((f-c)^+)^2\right] - \left(\mathbb{E}_\pi\left[(f-c)^+\right]\right)^2 \geq \mathbb{E}_\pi\left[f^2\right] - 2c\mathbb{E}_\pi[f] - (\mathbb{E}_\pi[f])^2.$$

Let $c = \frac{\mathrm{Var}_\pi(f)}{4\mathbb{E}_\pi[f]}$, which is $\leq \frac{\max f}{4}$. By Markov's inequality $\pi(f > c) < \frac{\mathbb{E}_\pi[f]}{c}$, so

$$\mathcal{E}(f,f) \geq (\mathrm{Var}_\pi(f) - 2c\mathbb{E}_\pi[f])\Lambda\left(\frac{\mathbb{E}_\pi[f]}{c}\right) = \frac{1}{2}\mathrm{Var}_\pi(f)\,\Lambda\left(\frac{4(\mathbb{E}_\pi[f])^2}{\mathrm{Var}_\pi(f)}\right)$$

which completes the proof. □

We now prove Theorem 6.2. Recalling that $\left\|\frac{P^t(x,\cdot)}{\pi(\cdot)} - 1\right\|_2^2 = \mathrm{Var}_\pi\left(P^{*t}f_x\right)$ it is enough to bound the variance of $P^{*t}f_x$ which we denote by $I_x(t)$. We extend $I_x(t)$ to $\mathbb{R}_+$ such that it is piecewise linear between integers. The previous lemma gives $I_x'(t) \leq -\frac{1}{2}I_x(m)\pi\Lambda_{PP^*}\left(\frac{4}{I_x(m)}\right)$ for $m = \lfloor t \rfloor$ and as $I_x(t) \leq I_x(m)$ and Λ_{PP*} is non-increasing the proof follows by integrating this from 0 to t and using $I_x(0) \leq \frac{1}{\pi_{\min}}$.

Proof of Theorem 6.2. Recall that from the equation (5.1) if we set $f_x(z) = \mathbb{1}_{x=z}/\pi(x)$ we have

$$\left\|\frac{P^t(x,\cdot)}{\pi(\cdot)} - 1\right\|_2^2 = \mathrm{Var}_\pi\left(P^{*t}f_x\right). \tag{6.1}$$

As for all $t \in \mathbb{N}_0$ we have $\mathbb{E}_\pi\left[P^{*t}f_x\right] = 1$ then Lemma 6.3 we have that

$$\mathrm{Var}_\pi\left(P^*(P^{*t}f_x)\right) - \mathrm{Var}_\pi\left(P^{*t}f_x\right) \leq -\frac{1}{2}\mathrm{Var}_\pi\left(P^{*t}f_x\right)\Lambda_{PP^*}\left(\frac{4}{\mathrm{Var}_\pi\left(P^{*t}f_x\right)}\right).$$

We now think of $\mathrm{Var}_\pi\left(P^{*t}f_x\right)$ as a function of t and want to see what the maximal possible value of a function satisfying the above inequality for all $t \in \mathbb{N}_0$ could be. We define I_x to be the piecewise linear extension of this function, i.e. $I_x : [0,\infty) \to \mathbb{R}_+$ such that for all $m \in \mathbb{N}_0$ we set $I_x(m) = \mathrm{Var}_\pi\left(P^{*m}f_x\right)$ and for $t \in (m, m+1)$ we set $I_x(t) = I_x(m) + (t-m)(I_x(m+1) - I_x(m))$. Differentiating this expression and using the above inequality, that I_x is decreasing non-negative function and that Λ is non-increasing, gives that for $t \in [m, m+1)$

$$I_x'(t) = I_x(m+1) - I_x(m) \leq -\frac{1}{2}I_x(m)\Lambda_{PP^*}\left(\frac{4}{I_x(m)}\right) \leq -\frac{1}{2}I_x(t)\Lambda_{PP^*}\left(\frac{4}{I_x(t)}\right).$$

Rearranging and integrating this over $[0,t]$ gives

$$\int_{I_x(0)}^{I_x(t)} \frac{dI_x}{I_x\Lambda_{PP^*}\left(\frac{4}{I_x}\right)} = \int_0^t \frac{I_x'(t)}{I_x(t)\Lambda_{PP^*}\left(\frac{4}{I_x(t)}\right)}dt \leq -\frac{1}{2}t.$$

Change of variable $u = 4/I_x$ gives that

$$\frac{1}{2}t \leq \int_{4/I_x(0)}^{4/I_x(t)} \frac{du}{u\Lambda_{PP^*}(u)} \leq \int_{4\pi_{\min}}^{4/I_x(t)} \frac{du}{u\Lambda_{PP^*}(u)}$$

since $I_x(0) = \mathrm{Var}_\pi(f_x) = \frac{1}{\pi(x)} - 1 \leq \frac{1}{\pi_{\min}}$. As $V(t/2) = V_{PP^*}(t/2)$ is the value for which

$$\int_{4\pi_{\min}}^{V(\frac{t}{2})} \frac{du}{u\Lambda_{PP^*}(u)} = \frac{t}{2} \leq \int_{4\pi_{\min}}^{4/I_x(t)} \frac{du}{u\Lambda_{PP^*}(u)}$$

we see that

$$I_x(t) \leq \frac{4}{V(\frac{t}{2})}$$

and the proof of the theorem follows using equation (6.1) and the definition of $I_x(t)$. □

One way of thinking about the above result is that it gives us a way to relate the spectral profile of PP^* with the time it takes to decrease the $\mathcal{L}^2$ distance to π from $\frac{1}{\pi_{\min}}$ (which is upper bound on the largest possible $\mathcal{L}^2$ distance) to any target distance $\varepsilon > 0$ (i.e. with ε-$\mathcal{L}^2$ mixing time). The result simply says that in time $\int_{4\pi_{\min}}^{4/\varepsilon^2} \frac{du}{u\Lambda_{PP^*(u)}}$ the distance is dropped from maximal possible to at most ε. We recall here that in our definition of $\mathcal{L}^2$ distance for probability distributions we divide by π, so the distance can be greater than 1. We know that $\mathcal{L}^2$ distance also upper bounds total variation distance, so this method would give us a bound on the (total variation) mixing time. However, this bound is only useful if we can find a spectral profile for PP^* which might be trickier to do than finding the spectral profile for P. Therefore, it is useful to state the following consequence of this theorem, which gives us a way to bound mixing time in terms of the spectral profile of P when the chain we consider is lazy (and in fact we will state the result in the most general form where the laziness parameter doesn't have to be 1/2).

Corollary 6.4 (Bounding mixing time by the spectral profile)**.** *Assume that there exists a constant $\alpha > 0$ such that for all $x \in S$ we have $P(x,x) \geq \alpha$. Then for $\varepsilon > 0$ the $\mathcal{L}^\infty$ mixing time satisfies*

$$t_{\text{mix}}(\varepsilon) \leq t_{\text{mix}}^{(2)}(\varepsilon) \leq t_{\text{mix}}^{(\infty)}(\varepsilon) \leq 2\left\lceil \int_{4\pi_{\min}}^{4/\varepsilon} \frac{du}{\alpha u \Lambda(u)} \right\rceil.$$

Moreover, using that spectral profile is lower bounded by the Poincaré constant, i.e. $\Lambda(r) \geq \gamma(P)$ above directly implies that for any $M > \varepsilon$ we have

$$t_{\text{mix}}(\varepsilon) \leq t_{\text{mix}}^{(2)}(\varepsilon) \leq t_{\text{mix}}^{(\infty)}(\varepsilon) \leq 2\left\lceil \int_{4\pi_{\min}}^{4/M} \frac{du}{\alpha u \Lambda(u)} \right\rceil + 2\left\lceil \frac{1}{\alpha\gamma(P)} \log(M/\varepsilon) \right\rceil.$$

As $\pi(x)PP^*(x,y) \geq \alpha(\pi(x)P(x,y)) + \alpha\pi(y)P(y,x)$ gives $\mathcal{E}_{PP^*}(f) \geq 2\alpha\mathcal{E}_P(f)$ so $\Lambda_{PP^*}(r) \geq 2\alpha\Lambda(r)$ and $V_{PP^*}(t/2) \geq V_P(\alpha t)$ and similarly $V_{P^*P}(t/2) \geq V_P(\alpha t)$ then Cauchy-Schwarz and reversing the path from t give $\left|\frac{P^{2t}(x,y)}{\pi(y)} - 1\right| \leq \left\|P^t(x,\cdot) - \pi\right\|_2 \left\|P^{*t}(y,\cdot) - \pi\right\|_2 \leq \sqrt{\frac{4}{V_{PP^*}(t/2)} \frac{4}{V_{P^*P}(t/2)}} \leq \frac{4}{V_P(\alpha t)}$.

Proof. Observe that $P^*(x,x) = P(x,x) \geq \alpha$ and so for all $x, y \in S$

$$PP^*(x,y)\pi(x) \geq P(x,x)P^*(x,y)\pi(x) + P(x,y)P^*(y,y)\pi(x) \geq \alpha P(y,x)\pi(y) + \alpha P(x,y)\pi(x).$$

This means, as invariant distribution of both P and PP^* is π, that

$$\begin{aligned}\mathcal{E}_{PP^*}(f) &\geq \frac{1}{2}\sum_{x,y\in S} (f(x) - f(y))^2 \left(\alpha P(x,y)\pi(x) + \alpha P(y,x)\pi(y)\right) \\ &= 2\alpha\frac{1}{2}\sum_{x,y\in S} (f(x) - f(y))^2 P(x,y)\pi(x) = 2\alpha\mathcal{E}_P(f)\end{aligned}$$

implying $\Lambda_{PP^*}(r) \geq 2\alpha\Lambda(r)$ for all r and so

$$\alpha t = \int_{4\pi_{\min}}^{V_P(\alpha t)} \frac{dv}{v\Lambda_P(v)} \geq 2\alpha \int_{4\pi_{\min}}^{V_P(\alpha t)} \frac{dv}{v\Lambda_{PP^*}(v)}$$

which implies that $V_{PP^*}\left(\frac{t}{2}\right) \geq V(\alpha t)$ and similarly we can get that $V_{P^*P}\left(\frac{t}{2}\right) \geq V(\alpha t)$. We now have for any $x, y \in S, t \in \mathbb{N}$, similar as in the Proposition 3.1, that

$$\left|\frac{P^{2t}(x,y)}{\pi(y)} - 1\right| = \left|\sum_{z\in S} \frac{P^t(x,z)P^t(z,y)}{\pi(y)} - 1\right|$$
$$= \left|\sum_{z\in S} \left(\frac{P^t(x,z)}{\pi(z)} - 1\right)\left(\frac{P^{*t}(y,z)}{\pi(z)} - 1\right)\pi(z)\right| \leq \left\|P^t(x,\cdot) - \pi\right\|_2 \left\|P^{*t}(y,\cdot) - \pi\right\|_2,$$

where the inequality follows from Cauchy-Schwarz. Using the Theorem 6.2 for both P and P^* (and that the reversal of P^* is P) we get

$$\left|\frac{P^{2t}(x,y)}{\pi(y)} - 1\right| \leq \sqrt{\frac{4}{V_{PP^*}(t/2)}\frac{4}{V_{P^*P}(t/2)}} \leq \frac{4}{V_P(\alpha t)}$$

and so from the definition of $V_P(\alpha t)$ we have

$$\left|\frac{P^{2t}(x,y)}{\pi(y)} - 1\right| \leq \varepsilon \quad \text{for} \quad t \geq \left\lceil \int_{4\pi_{\min}}^{4/\varepsilon} \frac{du}{\alpha u \Lambda(u)} \right\rceil$$

where $\Lambda = \Lambda_P$, which completes the proof, as then $d_\infty(2t) = \max_{x\in S}\max_{y\in S}\left|\frac{P^{2t}(x,y)}{\pi(y)} - 1\right| \leq \varepsilon$. $\square$

Lecture 13

Remark 6.5. *Note that in the setup of above result, one can also get that if* $t \geq 2\left\lceil \int_{4\pi_{\min}}^{4/\varepsilon} \frac{du}{\alpha u \Lambda(u)} \right\rceil$ *then* $d_\infty(t) \leq \varepsilon$ *by bounding* $\left|\frac{P^{2t+1}(x,y)}{\pi(y)} - 1\right| \leq \left\|P^t(x,\cdot) - \pi\right\|_2 \left\|P^{*t+1}(y,\cdot) - \pi\right\|_2 \leq \frac{4}{V_P(\alpha t)}$ *for odd times .*

We have developed a very general upper bound on the mixing time, but the key question now is if this bound would give us sharp mixing time estimates in any example. Also taking infimum over all possible non-negative functions supported in a certain set doesn't seem like an easy thing to do. So how can we use this spectral profile bound? It turns out that as we had that spectral gap and Poincaré constant are the same for the reversible chain, we also have a spectral representation of the spectral profile (that is where the name spectral profile comes from). The following result is from Question 6 from Example Sheet 3.

Lemma 6.6 (Spectral representation of the spectral profile)**.** *For any* $A \subset S$ *and any transition matrix* P *on* S *define* $\lambda_0(A) = \inf_{f\in c_0(A)} \frac{\mathcal{E}_{P_A}(f)}{\|f\|_2^2} = \inf_{f\in c_0(A)} \frac{\langle (I-P_A)f, f\rangle_\pi}{\|f\|_2^2}$ *where* $P_A(i,j) = P(i,j)$ *for all* $i, j \in A$ *and* 0 *otherwise, and* $c_0(A) = \{f : \mathrm{supp}(f) \subset A, f$ *is not constant*$\}$*. For every chain we have*

$$\lambda_0(A) \leq \lambda(A) \leq \frac{1}{1-\pi(A)}\lambda_0(A).$$

For $r \geq \pi_{\min}$ *define* $\Lambda_0(r) = \inf_{\pi(A)\leq r} \lambda_0(A)$ *then*

$$\Lambda_0(r) \leq \Lambda(r) \leq \frac{1}{1-r}\Lambda_0(r).$$

Moreover, when P *is reversible for any* $A \subset S$ $\lambda_0(A)$ *is also the smallest eigenvalue of matrix* $(I - P_A)_{ij\in A}$ *where* $P_A(i,j) = P(i,j)$ *for all* $i, j \in A$ *and* I *is the identity matrix.*

Finding the smallest eigenvalue seems more manageable than taking the infimum over a certain set of functions, so we now give an example where we can get a sharp bound using the spectral profile. We first quickly prove another useful result.

Lemma 6.7 (Spectral profile is infimum over functions with connected support)**.** *Assume that a decomposition of set A into connected components is $A = A_1 \cup A_2 \cup \ldots \cup A_k$ (that means that you can not reach A_i from any vertex in A_j in one step, for all $i \neq j$). Then $\lambda(A) = \min_{i \leq k} \lambda(A_i)$, and so to find a value of the spectral profile function, it is enough to consider connected sets.*

Key is to notice that if f_i is f restricted to A_i then as A_i are disjoint $\mathcal{E}(f) = \sum_{i=1}^{k} \mathcal{E}(f_i)$.

Proof. From the definition it is clear that $\lambda(A) \leq \lambda(A_i)$ for all i so we need to show that $\lambda(A) \geq \min_{i \leq k} \lambda(A_i)$. For a non-negative function f we let $f_i = f \mathbb{1}_{A_i}$ and we have

$$\mathrm{Var}_\pi(f) = \mathrm{Var}_\pi\left(\sum_{i=1}^{k} f_i\right) = \sum_{i=1}^{k} \mathbb{E}_\pi\left[f_i^2\right] - \left(\sum_{i=1}^{k} \mathbb{E}_\pi[f_i]\right)^2 \leq \sum_{i=1}^{k} \mathrm{Var}_\pi(f_i),$$

where we have used that f is non-negative and that f_i have disjoint support. As A_i are disconnected we also have that

$$\mathcal{E}(f) = \sum_{i=1}^{k} \mathcal{E}(f_i)$$

This means that

$$\lambda(A) = \inf_{f \in c_0^+(A)} \frac{\sum_{i=1}^{k} \mathcal{E}(f_i)}{\mathrm{Var}_\pi(f)} \geq \inf_{f \in c_0^+(A)} \frac{\sum_{i=1}^{k} \lambda(A_i)\mathrm{Var}_\pi(f_i)}{\mathrm{Var}_\pi(f)} \geq \inf_{i \leq k} \lambda(A_i)$$

and this completes the proof □

We now discuss an example of a lazy random walk on an n-cycle and see another proof of the mixing time bound of $O(n^2)$. Recall results we have from Claims 4.17 and 2.4. We can not simply use the order of Poincaré constant (i.e. spectral gap) and Theorem 5.9 to get a correct order bound for mixing time, as we would get an extra $\log(n)$ term from $\log(1/\pi_{\min})$. So let's try to use the spectral profile. From the above result, we only need to bound $\lambda(A)$ for connected sets and it is enough to lower bound $\lambda_0(A)$. The connected set on a cycle is just a straight line and the matrix $I - P_A$ has $1/2$ on diagonals and $-1/4$ on the first upper and lower off-diagonals. It is an exercise in linear algebra (you can look up traditional Toeplitz matrix) to find that the smallest eigenvalue of this matrix is

$$\lambda_0(A) = \frac{1}{2}(1 - \cos(\pi/(|A| + 1))),$$

where $|A|$ is the size of A, i.e. the dimension of this matrix. So $\Lambda_0(r) = \frac{1}{2}(1 - \cos(\pi/(\lfloor rn \rfloor + 1)))$ and $\Lambda(r) \gtrsim 1/(rn)^2$. Applying Corollary 6.4 for $M = 8$, $\alpha = 1/2$ and using that $\pi \equiv 1/n$, and $t_{\mathrm{rel}} \asymp n^2$ we get

$$t_{\mathrm{mix}}(\varepsilon) \leq t_{\mathrm{mix}}^{(\infty)}(\varepsilon) \lesssim \int_{4/n}^{1/2} \frac{n^2 u^2 du}{u} + n^2 \log(1/\varepsilon) \lesssim n^2 \log(1/\varepsilon).$$

We might still think that the spectral profile is not very useful, as we already know the mixing time of a lazy walk on a cycle so this is not new to us, but we will now see a nice consequence of knowing that the spectral profile gives us this bound of correct order. We can recall that if we have two chains, P and $\widetilde{P}$ such that $\mathcal{E}(f) \leq C\widetilde{\mathcal{E}}(f)$, we can compare their Poincaré constants (see Theorem 5.11). In exactly the same way, we can compare their spectral profile. We use this to get the following result.

Claim 6.8 (Mixing time of the lazy simple random walk on cycle with added edges)**.** *Consider a lazy simple random walk on an n-cycle ($\mathbb{Z}_n$) to which we add any edges we want but such that any vertex has degree at most some fixed constant $\Delta > 2$. Then $t_{\text{mix}}(\varepsilon) \lesssim n^2 \log(1/\varepsilon)$.*

One uses the fact that the spectral profile gives $t_{\text{mix}} \lesssim n^2$ for a lazy walk on a cycle and then compares the spectral profile for the cycle and the cycle with added edges by comparing the variances and Dirichlet forms of functions.

Proof. We use the fact that the spectral profile gives us a correct order bound for a simple random walk on the cycle. We label by P a lazy random walk on the cycle, while $\widetilde{P}$ corresponds to the walk on the cycle with added edges. If we add edges such that degrees are at most Δ that means that for all $x \in \mathbb{Z}_n$ $1/(\Delta n) \leq \widetilde{\pi}(x) \leq \Delta/n$ and that $\widetilde{P}(x, x \pm 1) \geq 1/(2\Delta)$. We see that for any fixed f

$$2\mathcal{E}(f) = \sum_{x,y \in \mathbb{Z}_n} \pi(x)P(x,y)(f(x) - f(y))^2 \leq \Delta^2 \sum_{x,y \in \mathbb{Z}_n} \widetilde{\pi}(x)\widetilde{P}(x,y)(f(x) - f(y))^2 = 2\Delta^2 \widetilde{\mathcal{E}}(f).$$

From Theorem 5.11 this means that $\widetilde{\gamma} \gtrsim \gamma$ and as in the proof of that theorem we can compare the ratio of variances with respect to $\widetilde{\pi}$ and π up to the multiplicative factor $\max \frac{\widetilde{\pi}(x)}{\pi(x)} \leq \Delta$ we also have that for any fixed set A

$$\widetilde{\lambda}(A) \gtrsim \lambda(A).$$

Now as $\widetilde{\pi}(A) \leq r$ implies that $\pi(A) \leq \Delta r$ we have that for all $r \geq \widetilde{\pi}_{\min}$

$$1/(nr)^2 \lesssim \Lambda(\Delta r) \lesssim \widetilde{\Lambda}(r)$$

and so the spectral profile bound [as $\int_{4\widetilde{\pi}_{\min}}^{\frac{4}{M}} \frac{du}{u\widetilde{\Lambda}(u)} \lesssim \int_{4\widetilde{\pi}_{\min}}^{\frac{4}{M}} \frac{n^2u^2du}{u} \lesssim n^2$] gives the same result as for the lazy simple random walk on the cycle, and this completes the proof.

□

We note that from Corollary 6.4 that the above also gives a bound on $\mathcal{L}^\infty$ mixing time, which is an even stronger result than just bounding the mixing time.

We now move to another way we apply the spectral profile result, and that is by bounding it in terms of the isoperimetric profile, which is a generalisation of the bottleneck ratio. The isoperimetric profile is the easiest way in practice to get a bound on the spectral profile; however, the bound might not always be sharp. The relation between the two is the same as the relation between the spectral gap and charger constant, i.e. we will have a somewhat more general version of Cheeger's inequality.

6.2 Isoperimetric profile

Recall Definition 5.18 of bottleneck ratio and recall that we defined the conductance of set $A \subset S$ to be $\Phi(A) = \frac{Q(A,A^c)}{\pi(A)}$ where $Q(A,B) = \sum_{x \in A, y \in B} \pi(x)P(x,y)$. We now define the generalisation of the bottleneck ratio, known as the isoperimetric profile.

Definition 6.9 (Isoperimetric profile)**.** *Isoperimetric profile* of transition matrix P is a function $\Phi_* : [\pi_{\min}, \infty) \to \mathbb{R}$ defined by

$$\Phi_*(r) = \inf_{\pi_{\min} \leq \pi(A) \leq r} \Phi(A)$$

for all $r \in [\pi_{\min}, 1/2]$ and $\Phi_*(r) = \Phi_*(1/2)$ for $r > 1/2$. //

We recall that we have shown Cheeger's inequality (Theorem 5.20), which bounds Poincaré constant in terms of the bottleneck ratio, and now state the corresponding result for the bound on the isoperimetric profile in terms of the spectral profile.

Theorem 6.10 (Generalised Cheeger's inequality)**.** *For $r \in [\pi_{\min}, 1/2]$ spectral profile Λ and isoperimetric profile Φ_* satisfy*

$$\frac{\Phi_*^2(r)}{2} \leq \Lambda(r) \leq \frac{\Phi_*(r)}{1-r}.$$

The proof of this result follows directly from showing that

$$\frac{1}{2}\Phi_*^2(\pi(A)) \leq \lambda_0(A) \leq \Phi(A)$$

which can be done in a similar way as the proof of Cheeger's inequality and it will be done in Question 6 in Example Sheet 3.

Generalised Cheeger's inequality and spectral profile bound on the mixing time give us the following useful corollary.

Corollary 6.11 (Bounding mixing time by the isoperimetric profile)**.** *Assume that there exists a constant $\alpha > 0$ such that for all $x \in S$ we have $P(x,x) \geq \alpha$. Then for $\varepsilon > 0$ the $\mathcal{L}^\infty$ mixing time satisfies*

$$t_{\text{mix}}(\varepsilon) \leq t_{\text{mix}}^{(2)}(\varepsilon) \leq t_{\text{mix}}^{(\infty)}(\varepsilon) \leq 2\left\lceil \int_{4\pi_{\min}}^{4/\varepsilon} \frac{2du}{\alpha u \Phi_*^2(u)} \right\rceil.$$

Moreover, we have

$$t_{\text{mix}}^{(\infty)}(\varepsilon) \leq 2\left\lceil \int_{4\pi_{\min}}^{4/M} \frac{2du}{\alpha u \Phi_*^2(u)} \right\rceil + 2\left\lceil \frac{1}{\alpha\gamma(P)} \log(M/\varepsilon) \right\rceil \leq 2\left\lceil \int_{4\pi_{\min}}^{4/M} \frac{2du}{\alpha u \Phi_*^2(u)} \right\rceil + 2\left\lceil \frac{2}{\alpha\Phi_*^2} \log(M/\varepsilon) \right\rceil.$$

Proof. The proof follows directly from Corollary 6.4 by lower bounding the spectral profile and Poincaré constant by isoperimetric profile and bottleneck ratio using Theorems 5.20 and 6.10. □

We first give one obvious application of the above theorem, which talks about lazy random walks on graphs which have a certain *small set expansion* property.

Claim 6.12 (Small set expander mixing time)**.** *We say that a family of graphs G_n is a (Δ, α, c)-small set expander family if for all n we have that the maximal degree in G_n is at most Δ and that $\Phi_*(c) \geq \alpha$ and if the number of vertices in G_n diverges to ∞ as $n \to \infty$. We call a family of graphs G_n if there are constants $\Delta, c > 0, \alpha > 0$ such that the family is (Δ, α, c)-small set expander family. We now consider a lazy simple random walk on a small set expander family where G_n has n vertices. The mixing time of this chain satisfies*

$$t_{\text{mix}}(\varepsilon) \lesssim \log n + t_{\text{rel}} \log(1/\varepsilon),$$

where $t_{\text{rel}} = 1/\gamma$ is the relaxation time, i.e. the inverse of the spectral gap.

Before proving the claim we recall again that we often drop n from the notation, so when we write γ or t_{mix} in the above results we really mean the spectral gap of the nth chain, i.e. γ_n and the mixing time of the nth chain, i.e. t_{mix}^n.

Just apply the isoperimetric profile bound.

Proof. We apply Corollary 6.11 for $M = 4/c$, where c is a constant such that $\Phi_*(c) \gtrsim 1$ and, as the chain is lazy and reversible so inverse of the Poincaré constant is the relaxation time, we have

$$\begin{aligned} t_{\text{mix}}(\varepsilon) &\lesssim \int_{4\pi_{\min}}^{c} \frac{du}{u\Phi_*^2(u)} + \frac{1}{\gamma(P)} \log(4/(c\varepsilon)) \lesssim \int_{4\pi_{\min}}^{c} \frac{du}{u} + t_{\text{rel}} \log(1/\varepsilon) \\ &= \log(c) - \log(4\pi_{\min}) + t_{\text{rel}} \log(1/\varepsilon) \lesssim \log n + t_{\text{rel}} \log(1/\varepsilon) \end{aligned}$$

where in the last line we have used that due to the bounded degree $\pi_{\min} \gtrsim \frac{1}{n}$. □

The above claim also tells us that if we have a small set expander family satisfying $t_{\text{rel}} \lesssim \log n$ (if a graph family were expander family the relaxation time would, from Cheeger's inequality, be of constant order, so this covers a broader class of graphs than expanders) then the mixing time is as fast as it could be on a bounded degree graph, i.e. it is $\log n$.

Lecture 14

More examples with specific chains will be given in the example sheet. We finish this section by stating a simple application of this result, which gives a general bound on transition probabilities for lazy random walks on bounded degree graphs.

Proposition 6.13 (Bound on transition probabilities for random walks)**.** *Consider a connected graph on n vertices with degrees bounded by Δ and let P denote a transition matrix of a lazy simple random walk on this graph. Then for all $t \in \mathbb{N}$ and all vertices x and y we have*

$$|P^t(x,y) - \pi(y)| \lesssim \frac{\Delta^2}{\sqrt{t}}.$$

Just apply the isoperimetric profile bound, where you bound $\Phi(A) \geq \frac{1}{E(A)}$, where $E(A)$ is the sum of degrees of A, which holds by noticing that as the graph is connected so any set A has at least one edge leading out of it and using the definition of $\Phi(A)$.

Proof. We use a very trivial bound for any set $A \subset S$

$$\Phi(A) \geq \frac{1}{2E(A)}$$

where $E(A) = \sum_{x \in A} \deg(x)$ is the total number of edges coming out of A. This bound holds as the graph is connected, so every set A has at least one edge coming out of it, and we know that the conductance of A is just the number of edges between A and A^c divided by twice the total sum of degrees of vertices in A (note that $1/2$ comes from the fact that we are looking at the lazy rather than just the simple random walk). This means that if $\pi(A) = \frac{E(A)}{2|E|} \leq u$ then $\Phi(A) \geq \frac{1}{4u|E|}$ where $|E| \leq \Delta n$ is the total number of edges. We now use the isoperimetric profile bound from Corollary 6.11 and get that for any $M \geq 8$

$$t_{\text{mix}}^{(\infty)}(M) \leq 2 \left\lceil \int_{4\pi_{\min}}^{4/M} \frac{4du}{u\Phi_*^2(u)} \right\rceil \leq 2 \left\lceil \int_{4\pi_{\min}}^{4/M} \frac{4du}{u\left(\frac{1}{4\Delta n u}\right)^2} \right\rceil \lesssim \frac{\Delta^2 n^2}{M^2}.$$

We first note that the first inequality above holds for all M but that we can lower bound $\Phi_*(u) \geq \frac{1}{n\Delta}$ for $u \geq 1/2$ so we directly get $t_{\text{mix}}^{(\infty)}(\varepsilon) \lesssim \Delta^2 n^2 + \Delta^2 n^2 \log\left(\frac{8}{\varepsilon}\right) \lesssim \frac{\Delta^2 n^2}{\varepsilon^2}$ for $\varepsilon \leq 8$. Also, there is a constant C such that for the time at least $C\frac{\Delta^2 n^2}{M^2}$ the $\mathcal{L}^\infty$ distance at time t is below M (using Remark 6.5). So for a fixed t we can take $M = \frac{2\Delta n\sqrt{C}}{\sqrt{t}}$ then the inequality $t \geq C\frac{\Delta^2 n^2}{M^2}$ holds and

we therefore know that $\|P^t(x,\cdot) - \pi(\cdot)\|_\infty \leq \frac{2\Delta n\sqrt{C}}{\sqrt{t}}$ which, from the definition of $\mathcal{L}^\infty$ distance for probability distributions, means that

$$|P^t(x,y) - \pi(y)| \leq \pi(y)\frac{2\Delta n\sqrt{C}}{\sqrt{t}} \leq \frac{2\sqrt{C}\Delta^2}{\sqrt{t}} \lesssim \frac{\Delta^2}{\sqrt{t}},$$

where we have used that $\pi_{\max} \leq \Delta/n$. □

The above result tells us that in particular, for a bounded degree graph on n vertices, at time $t \lesssim n^2$ we have a probability $\lesssim \frac{1}{\sqrt{t}}$ to be at any given vertex, regardless of where we started from. This is a very general and useful result and we note that this can also be extended to infinite graphs but this will not be discussed in this course.

7 Geometric techniques

When we talk about geometric techniques, we mean any method which uses the geometry of the Markov chain (i.e. of the (weighted and directed) graph where vertices are states and edges are non-zero transitions) to control the value of the mixing time or some other property of the Markov chain. An example of the geometric method is the theory with bottleneck ratio, or bounding mixing time of random walks on graphs using the graph's diameter, so we have already seen some geometric techniques by now.

We will start this section by showing an upper bound on transition probability between any states in the reversible Markov chain in terms of their distance, allowing us to lower bound the mixing time. We will then continue this section by introducing another general technique to upper bound the mixing time.

7.1 Varopoulos-Carne bound

The goal is to get a lower bound on the mixing time by controlling how far the walk is likely to get after t steps. A simplest bound on the distance would be to say that after t steps the walk can not get further than distance t away from its starting position. But this bound is in most cases very weak and a typical displacement of the chain is a lot smaller. We formally define *distance* as

$$\operatorname{dist}(x,y) = \min\{t : P^t(x,y) > 0\}$$

and it is not hard to check that in a case of reversible Markov chain distance is symmetric and is a metric on the state space S. The distance corresponds to the usual graph distance on the graph with vertices S where two states $x, y \in S$ are connected with an edge iff $P(x,y) > 0$ which is, due to reversibility equivalent to $P(y,x) > 0$.

To control how far the walk gets after t steps we will show a bound on $P^t(x,y)$ in terms of $\operatorname{dist}(x,y)$ and t. We will start by showing a standard bound for a simple random walk on $\mathbb{Z}$ which will actually be useful in proving the general result.

Proposition 7.1 (Distance of simple random walk on $\mathbb{Z}$ from the origin)**.** *Let Z_t be a simple random walk on $\mathbb{Z}$ starting from the origin. Then for all $t, d \in \mathbb{N}$ we have*

$$\mathbb{P}(|Z_t| \geq d) \leq 2\exp\left(-\frac{d^2}{2t}\right)$$

Apply Markov's inequality to $e^{\lambda Z_t}$.

Proof. The proof of this is a classical application of Markov's inequality to the exponential of a random variable, which is sometimes called Chernoff's bound. For any $d, \lambda > 0$, Markov's inequality and computing the moment generating function of simple random walk (we can write it as the sum of i.i.d. ± 1 steps and use that MGF of the sum of k i.i.d. variables is a kth power of MGF of one of these variables), give

$$\mathbb{P}(Z_t \geq d) = \mathbb{P}\Big(e^{\lambda Z_t} \geq e^{\lambda d}\Big) \leq e^{-\lambda d}\mathbb{E}\Big[e^{\lambda Z_t}\Big] = e^{-\lambda d}\left(\frac{e^{\lambda} + e^{-\lambda}}{2}\right)^t \leq e^{-\lambda d + \frac{\lambda^2 t}{2}}.$$

Picking $\lambda = d/t$ minimizes the right-hand side and gives $e^{-d^2/(2t)}$ upper bound. The result now follows by recalling that by symmetry $\mathbb{P}(Z_t \leq -d) = \mathbb{P}(Z_t \geq d)$ which gives a factor 2 from the statement of the result. [We have used that $\frac{e^{\lambda}+e^{-\lambda}}{2} = \sum_{k=0}^{\infty} \frac{\lambda^{2k}}{(2k)!} \leq \sum_{k=0}^{\infty} \frac{\lambda^{2k}/2^k}{k!} = e^{\lambda^2/2}$.] □

We now state and prove a very general result for transitions of reversible chains.

Theorem 7.2 (Varopoulos-Carne estimate)**.** *For any reversible transition matrix P with invariant distribution π we have that for all $x, y \in S$ and $t \in \mathbb{N}$*

$$P^t(x,y) \leq \sqrt{\frac{\pi(y)}{\pi(x)}}\mathbb{P}(|Z_t| \geq \operatorname{dist}(x,y))$$

where Z_t is a simple random walk on $\mathbb{Z}$ starting from the origin. Proposition 7.1 then gives

$$P^t(x,y) \leq 2\sqrt{\frac{\pi(y)}{\pi(x)}} \exp\left(-\frac{(\operatorname{dist}(x,y))^2}{2t}\right).$$

Check inductively that polynomials defined by $q_{k+1}(z) := 2zq_k(z) - q_{k-1}(z)$ with $q_0(z) = 1$ and $q_1(z) = z$ satisfy $q_k(\cos\theta) = \cos(k\theta)$ and q_k has degrees at most k. Notice that polynomial identity $z^t = \sum_{k=0}^{t} \mathbb{P}(|Z_t| = k)\, q_k(z)$ holds as we can check that it holds for infinitely many values of z (check it for $\cos(\theta)$). Apply it to $P^t(x,y)$ and using reversibility get that $\pi(x)q_k(P)(x,y) = \langle q_k(P)\mathbb{1}_y, \mathbb{1}_x\rangle_\pi \leq \|\mathbb{1}_x\|_2\|\mathbb{1}_y\|_2 = \sqrt{\pi(x)\pi(y)}$ holds as eigenvalues of $q_k(P)$ are in $[-1,1]$.

Proof. The proof uses the Chebyshev polynomials $(q_k)_{k\geq 0}$, which are defined by the recursion

$$q_{k+1}(z) := 2zq_k(z) - q_{k-1}(z) \quad \text{for } k \geq 1,$$

with initial conditions $q_0(z) = 1$ and $q_1(z) = z$. The trigonometric identity

$$2\cos(k\theta)\cos(\theta) = \cos((k+1)\theta) + \cos((k-1)\theta),$$

shows that $q_k(\cos\theta) = \cos(k\theta)$ for all $\theta \in \mathbb{R}$ and $k \in \mathbb{N}$. Now, observe that for all $t \in \mathbb{N}$,

$$\begin{aligned}(\cos\theta)^t &= \left(\frac{e^{i\theta} + e^{-i\theta}}{2}\right)^t = \mathbb{E}\Big[e^{i\theta Z_t}\Big] = \mathbb{E}[\cos(\theta Z_t)] + \mathbb{E}[i\sin(\theta Z_t)] \\ &= \sum_{k=0}^{t} \mathbb{P}(|Z_t| = k)\cos(k\theta) + \sum_{k=0}^{t} \mathbb{P}(Z_t = k)\,(\sin(k\theta) + \sin(-k\theta)) = \sum_{k=0}^{t} \mathbb{P}(|Z_t| = k)\, q_k(\cos\theta).\end{aligned}$$

Since this is true for all $\theta \in \mathbb{R}$, that means that for infinitely many z we have

$$z^t = \sum_{k=0}^{t} \mathbb{P}(|Z_t| = k)\, q_k(z),$$

and so the left-hand side and the right-hand side are the same polynomials.

So we can apply this polynomial identity to the matrix P, and evaluate the (x, y)-entry to get

$$P^t(x, y) = \sum_{k=\mathrm{dist}(x,y)}^{t} \mathbb{P}(|Z_t| = k)\, q_k(P)(x, y),$$

where we have observed that $P^k(x, y) = 0$ for $k < \mathrm{dist}(x, y)$, and so $q_k(P)(x, y) = 0$ for $k \leq \mathrm{dist}(x, y)$ (since we can show inductively that q_k has degree k). It remains to show that $\forall k \geq 0$

$$q_k(P)(x, y) \leq \sqrt{\frac{\pi(y)}{\pi(x)}},$$

and this is where we use reversibility. When P is reversible $q_k(P)$ is self-adjoint with respect to $\langle \cdot, \cdot \rangle_\pi$ (as P is and q_k is a polynomial) and has a set of eigenvalues $\{q_k(\lambda) : \lambda \text{ is eigenvalue of } P\} \subset q_k([-1, 1]) \subset [-1, 1]$ where the last inclusion follows from the identity $q_k(\cos\theta) = \cos(k\theta)$. Thus, $q_k(P)$ is a contraction meaning that

$$|\langle q_k(P)f, g\rangle_\pi| \leq \|f\|_2 \|g\|_2,$$

for all vectors $f, g : S \to \mathbb{C}$ (we can see this by expressing f and g as the sum of the orthonormal basis of eigenvectors with suitable weights). Taking $f = \mathbb{1}_y$ and $g = \mathbb{1}_x$ gives

$$\pi(x) q_k(P)(x, y) \leq \sqrt{\pi(x)\pi(y)},$$

which gives the wanted inequality. [Simple linear algebra gives that $|\langle q_k(P)f, g\rangle_\pi| \leq \|f\|_2\|g\|_2$. We note that from spectral theorem, we have an orthonormal basis of eigenfunctions of $q_k(P)$, which we can denote by $\widetilde{f}_i$ and which have real eigenvalues $\widetilde{\lambda}_i$. Writing $f = \sum_{i=1}^{|S|} a_i \widetilde{f}_i$, $g = \sum_{i=1}^{|S|} b_i \widetilde{f}_i$ gives that $|\langle q_k(P)f, g\rangle_\pi| = |\sum_{i=1}^{|S|} \widetilde{\lambda}_i a_i \langle \widetilde{f}_i, g\rangle_\pi| \leq \sum_{i=1}^{|S|} |a_i b_i| \leq \|f\|_2\|g\|_2$ where the last step is just Cauchy-Schwarz and penultimate step uses that by definition $b_i = \langle \widetilde{f}_i, g\rangle_\pi$, that $|\widetilde{\lambda}_i| \leq 1$ and the triangle inequality.] □

Remark 7.3. *We remark that reversibility matters in this theorem. It is not just the case that we can not come up with a proof that doesn't require reversibility but the result is actually not true when the reversibility condition is dropped. A way to see this is to consider a biased simple random walk on a cycle X_t, e.g. with clockwise transition $2/3$ and anticlockwise $1/3$. Due to symmetry, π is uniform so if the theorem is true in this case we would have $\mathbb{P}_x(X_t = y) \leq 2\exp\left(-\frac{(\mathrm{dist}(x,y))^2}{2t}\right)$. This means that if the walk starts from the origin, for any constant c, by summing over all vertices of distance $\geq ct$ which a walk can reach by time t, we have $\mathbb{P}(X_t \geq ct) \lesssim t\exp(-c^2 t/2)$, but from SLLN when t is large the walk at time t is expected to be around $t/6$ so this inequality will not hold for $c < 1/6$ and the theorem can not be true in this case. We also recall that the biased random walk on a line (which just means that we remove one edge from a cycle) is reversible so the above result has to hold for it. The difference is that π is not uniform anymore (in fact it increases as a geometric series) so the ratio $\pi(y)/\pi(x)$ is exponential in t when x is the origin and y is a vertex of distance $\gtrsim t$ and if we do the computations as we did for the cycle, we wouldn't, of course, get a contradiction.*

We now lower bound the mixing time of a reversible chain using the Varopoulos-Carne estimate.

Corollary 7.4 (Diffusive bound)**.** *For any $\varepsilon \in (0, 1/2)$ we have that for all large enough n if P is a lazy simple random walk on a simple (there are no double edges or loops) n vertex graph then*

$$t_{\text{mix}}(\varepsilon) \geq \frac{\operatorname{diam}(P)^2}{16 \log n},$$

where $\operatorname{diam}(P)$ *is the largest distance between any two vertices in this graph.*

Take two vertices x and y which have a distance equal to the diameter and use that WLOG the π of the set of vertices which are at least diameter half away from x is at least $1/2$. Apply Varopoulos-Carne bound to walk starting from x and any vertex further than diameter half from it and take the union bound over all such vertices.

Proof. Set $d = \lceil \operatorname{diam}(P)/2 \rceil$, so as $\operatorname{diam}(P) > 2(d-1)$ that means that we can find vertices x, y in the graph such that the balls of radius $d-1$ around them, denoted $\mathcal{B}_{d-1}(x)$ and $\mathcal{B}_{d-1}(y)$, respectively, are disjoint (formally ball of radius r around x, $\mathcal{B}_r(x)$ is defined to be a collection of all vertices which have distance at most r from x). Without loss of generality, we can assume $\pi(\mathcal{B}_{d-1}(x)) \leq 1/2$ (as if it didn't hold for x it would hold for y and we could swap the labels x and y). The elements of $\mathcal{B}_{d-1}(x)^c$ are distance at least $d \geq \operatorname{diam}(P)/2$ from x so from Theorem 7.2 we know that

$$P^t(x, \mathcal{B}_{d-1}(x)^c) \leq 2n^{3/2} e^{-\operatorname{diam}(P)^2/(8t)}$$

where we have used a very crude bound $|\mathcal{B}_{d-1}(x)^c| \leq n$ and $\pi(y)/\pi(x) = \deg(y)/\deg(x) \leq n$. Therefore,

$$d(t) \geq \|P^t(x,\cdot) - \pi\|_{\text{TV}} \geq |P^t(x, \mathcal{B}_{d-1}(x)^c) - \pi(\mathcal{B}_{d-1}(x)^c)| \geq \frac{1}{2} - 2n^{3/2} e^{-\operatorname{diam}(P)^2/(8t)}.$$

For $t \leq \operatorname{diam}(P)^2/(16 \log n)$ we have that $d(t) \geq 1/2 - 2/\sqrt{n}$ which completes the proof. □

Lecture 15

7.2 Path coupling

We can notice that the total variation distance is "blind" to the geometric distance of states. Even if we have a large probability to reach x from y in one step, we have that the total variation distance between distributions concentrated on x and y is maximal possible, i.e. 1. In fact, if we look at any two elements of the state space and two distributions concentrated on each of them, they have a total variation distance 1. So a simple idea would be to consider some other distance between distributions which also takes into account the geometric distance between the states. This brings us to the following definition.

Definition 7.5 (Transportation metric)**.** Let ρ be a metric on a state space S. We define a *transportation metric* between two distributions, μ and ν on S by

$$\rho_K(\mu, \nu) = \inf\{\mathbb{E}[\rho(X, Y)] : (X, Y) \text{ is a coupling of } \mu \text{ and } \nu\}. \tag{7.1}$$

//

We note that if $\mu = \mathbb{1}_x$ and $\nu = \mathbb{1}_y$, then $\rho_K(\mu, \nu) = \rho(x, y)$ and if $\rho(x, y) = \mathbb{1}_{x \neq y}$, then

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

We also remark here that we actually do not need to have a metric, but could consider a distance which is not symmetric but satisfies the other two properties from the definition of metric, and what follows would work in the same way (except that the transportation metric would not be a metric but we would rather call it transportation distance and it would satisfy properties of a metric except symmetry). This more general version would particularly be useful if we want to consider a distance defined by $\rho(x, y) = \text{dist}\,(x, y) = \min\{t : P^t(x, y) > 0\}$ which doesn't have to be symmetric in general (but is symmetric for all reversible chains).

Lemma 7.6 (Optimal ρ-coupling)**.** *There exists a coupling q_* of μ and ν such that*

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

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

Think of coupling q as vector in $[0, 1]^{S \times S}$ with condition that the marginals sum to μ and ν. As the set of such vectors is compact the infimum is attained.

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 $|S|^2 - 1$ dimensional simplex. The function

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

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

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

This concludes the proof. □

Lemma 7.7 (Transportation metric is a metric)**.** *The function ρ_K defines a metric on the space of probability distributions on S.*

For triangle inequality use the coupling (X, Y, Z) given by $\frac{p(x,y)q(y,z)}{\nu(y)}$ where p and q are optimal ρ-couplings of μ and ν and of ν and η respectively.

Proof. It is obvious that $\rho_K(\mu, \nu) \geq 0$ and that ρ_K is symmetric as ρ is symmetric (this is the only place where we use the symmetric property of metric). Also, if $\rho_K(\mu, \nu) = 0$, then this means that for an optimal ρ-coupling q_*

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

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 $\{(x, x) : x \in S\}$. This immediately gives now that $\mu = \nu$.

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

$$\rho_K(\mu, \eta) \leq \rho_K(\mu, \nu) + \rho_K(\nu, \eta). \tag{7.2}$$

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

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

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, the above expression bounds their transportation metric distance and this proves (7.2). □

We now define a class of metrics for which we can get, under certain conditions, an upper bound on the total variation distance in terms of their corresponding transportation metrics.

Definition 7.8 (Path metric)**.** For a connected graph $G = (V,E)$ and $\ell : E \to \mathbb{R}_+$ a *length function* on its edges satisfying that for all edges $(x,y) \in E$ we have $\ell(x,y) \geq 1$ we define a *path metric* corresponding to distance ℓ to be metric

$$\rho(x,y) = \min\left\{ \sum_{i=0}^{r-1} \ell(x_i, x_{i+1}) : x_0 = x, x_1, x_2 \ldots, x_{r-1}, x_r = y \text{ is a path from } x \text{ to } y \right\},$$

where a path is a sequence of vertices in which consecutive vertices in the sequence are connected by an edge of the graph. We also call $\sum_{i=0}^{r-1} \ell(x_i, x_{i+1})$ a length of a path $x_0 = x, x_1, x_2 \ldots, x_{r-1}, x_r$. //

Remark 7.9. *It is easy to check that the path metric is indeed a metric. We also see that $\rho(x,y) \leq \ell(x,y)$ for all $(x,y) \in E$ but that the equality does not need to hold (for example if on a complete graph on $\{x,y,z\}$ we have $\ell(x,y) = 1$, $\ell(y,z) = 1$ and $\ell(x,z) = 10$ then this is a distance function for which $\rho(x,z) = 2 < 10$).*

Proposition 7.10 (Transportation metric of a path metric bounds the total variation distance)**.** *Let μ and ν be a distribution on state space V and let $G = (V,E)$ be a connected graph and let ρ be a path metric corresponding to some length function ℓ of G. Then*

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

This is direct from the coupling bound on TV distance as for a path metric $\mathbb{P}(X \neq Y) \leq \mathbb{E}[\rho(X,Y)]$.

Proof. We know from Theorem 1.7 that the total variation distance is the infimum over all couplings X, Y of μ and ν of $\mathbb{P}(X \neq Y)$. From the definition of path metric we also know that $\rho(x,y) \geq \mathbb{1}_{x \neq y}$ so $\mathbb{P}(X \neq Y) \leq \mathbb{E}[\rho(X,Y)]$ and so taking infimum of this completes the proof. □

We now state the following theorem which bound the mixing time under certain conditions.

Theorem 7.11 (Mixing time bound in terms of the transportation metric)**.** *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 and let $\alpha \in \mathbb{R}$. 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*

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

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,

$$d(t) \le e^{-\alpha t}\mathrm{diam}\,(V) \quad \textit{and if } \alpha > 0 \quad t_{\mathrm{mix}}(\varepsilon) \le \frac{1}{\alpha}(\log(\mathrm{diam}\,(V)) + \log(1/\varepsilon)),$$

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

We first note that when the metric considered is the usual graph distance and the edges of the graph represent non-zero transition probabilities then the optimal α (i.e. largest possible) from the theorem above is called curvature. We note that positive α satisfying above doesn't have to exist and curvature could be negative. In the context of mixing time the above result is mainly useful if it holds for some $\alpha > 0$. We now prove this result.

Show the inequality for μ and ν being concentrated on a single points x and y and using the triangle inequality for ρ_K along the path of shortest length from x to y. Then extend to all distributions by using the coupling which starts from an optimal ρ-coupling (X, Y) of (μ, ν) and then generating the next step of the chain using the optimal ρ-coupling conditional on $(X, Y) = (x, y)$. The mixing time bound follows directly from the first part together with the upper bound on TV distance on the transportation metric and the definition of diameter.

Proof. We first establish the inequality $\rho_K(\mu P, \nu P) \le e^{-\alpha}\rho_K(\mu, \nu)$ for $\mu = \mathbb{1}_x$ and $\nu = \mathbb{1}_y$. Let $x = x_0, x_1, \ldots, x_k = y$ 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. [As $\mathbb{P}(X' = x') = \sum_{x,y\in V}\mathbb{P}(X' = x' \mid (X, Y) = (x, y))\,\mathbb{P}((X, Y) = (x, y)) = \sum_{x,y\in V} P(x, x')\mathbb{P}((X, Y) = (x, y)) = \sum_{x\in V} P(x, x')\mathbb{P}(X = x) = \sum_{x\in V} P(x, x')\mu(x) = (\mu P)(x')$ and similarly $\mathbb{P}(Y' = y') = (\nu P)(y')$]. We then have

$$\begin{aligned}\rho_K(\mu P, \nu P) \le \mathbb{E}\big[\rho(X', Y')\big] &= \sum_{(x,y)\in V^2}\mathbb{P}((X, Y) = (x, y))\,\mathbb{E}\big[\rho(X', Y') \;\big|\; (X, Y) = (x, y)\big]\\ &= \sum_{(x,y)\in V^2}\mathbb{P}((X, Y) = (x, y)) \cdot \rho_K(P(x,\cdot), P(y,\cdot))\\ &\le e^{-\alpha}\sum_{(x,y)\in V^2}\mathbb{P}((X, Y) = (x, y))\,\rho(x, y) = e^{-\alpha}\mathbb{E}[\rho(X, Y)] = e^{-\alpha}\rho_K(\mu, \nu),\end{aligned}$$

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 $t \in \mathbb{N}$

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

Taking now $\mu = \mathbb{1}_x$ and $\nu = \pi$, and maximising over x gives the second inequality of the theorem from which a mixing time bound directly follows. □

We note that the above result does not require the Markov chain to be irreducible and be able to reach any vertex in V from any starting point but it requires unique π in order for the mixing time to be well defined. The result also holds for chains with a single closed communicating class (on which π would be supported). We now present a specific application of the above theory related to proper graph (vertex) colourings.

Definition 7.12 (Proper colourings and Glauber dynamics)**.** Let $G = (V, E)$ be a graph. A *proper (vertex) colouring* of G with q colours is an assignment of one of the q colours to each vertex of the graph such that no two neighbouring vertices have the same colour. So a set of all proper colourings is

$$S = \left\{x \in \{1, 2, \dots q\}^V : x(u) \neq x(v) \ \forall (u, v) \in E\right\}.$$

We define a Markov chain on the set S using the *Glauber dynamics* which is described as follows: starting from any colouring in S we pick a uniformly at random vertex $w \in V$ and choose its colour uniformly at random from a set of colours which are not taken by any of its neighbours. It is easy to check that this Markov chain is reversible with respect to uniform distribution on S. //

We will now use the above theory to bound the mixing time of the Glauber dynamics chain under certain conditions.

Theorem 7.13 (Mixing times of Glauber dynamic chain on proper colourings)**.** *Let G be a graph on n vertices with maximal degree Δ and let $q > 2\Delta$. Then the Glauber dynamics chain mixing time $t_{\text{mix}}(\varepsilon)$ we have*

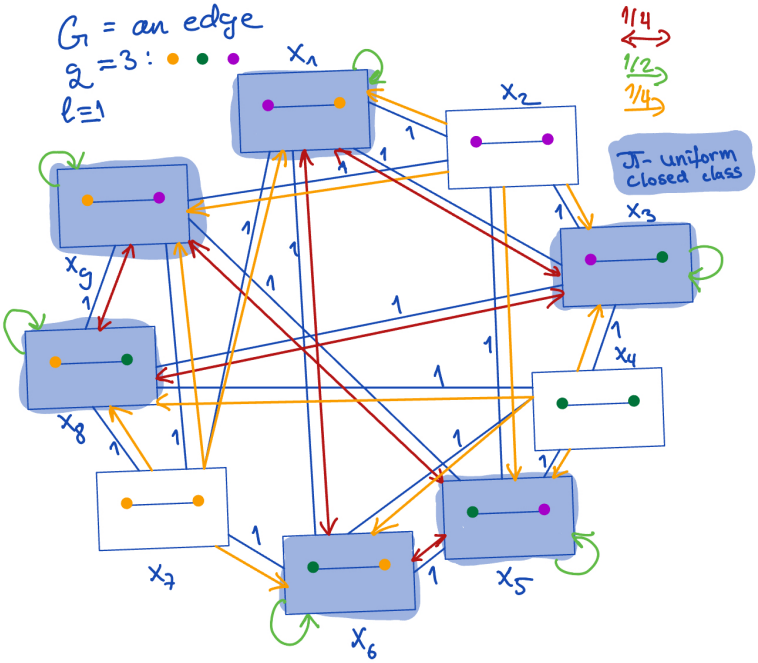
$$t_{\text{mix}}(\varepsilon) \leq \left\lceil \frac{q - \Delta}{q - 2\Delta} n(\log n - \log \varepsilon) \right\rceil.$$

Use Theorem 7.11 with the path metric defined to be the number of vertices where the two colourings disagree. For neighbouring colourings, use the same updating rule unless you have picked a neighbour of the vertex v where the two colourings differ and tried to update its colour to the one which is not allowed in the other colouring. Work out the updating rule in this case so that it gives a wanted coupling (consider cases when the number of allowed colours for the neighbour of v is the same or if it differs by one in the two colourings). Only the case where you pick a neighbour of v and try to update it to a colour which is allowed in exactly one of the colourings is going to increase the distance to 2, while changing the colour of v decreases it to 1, and one can now find α from the Theorem 7.11. We are using path metric on the set of all colourings and need to extend the Glauber dynamics to have the same updating rules on all (i.e. not necessarily proper) colourings.

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

We note that ρ is a path metric on the graph where the vertex set is a set of all colourings (not just proper ones, so this set is larger than the set on which the Markov chain we want to analyse moves) and where two colourings x and y are connected with an edge if they differ in the colour of one vertex of G (see figure below for a simple example where graph is one edge and we have 3 colours). Then the length function is just defined to have distance 1 on the edges and the corresponding path metric is ρ. We extend the Glauber dynamics to move on the set of all colourings using the same rule as for the proper colourings (i.e. pick a uniform vertex and update it to a uniform colour none of its neighbours have). We note that in the Theorem 7.11 it is actually not required that the chain is irreducible (however, we, in general, assume irreducibility and this example is probably the only

place in this course where we need to look at the non-irreducible chain) and has a positive transition probability to all possible vertices of the graph and along all edges, so we can apply the theorem here even though the chain started from a proper colouring will only move on proper colourings which is the subset of all colourings, and which we call set S. In general, for Theorem 7.11 it is only required that the invariant distribution is unique (as otherwise the mixing time definition of distance to stationarity and the mixing time does not make sense), which means that the chain needs to have exactly one closed communicating class, and we recall that the graph from that theorem does not need to be a graph induced by transitions of the chain.



A simple example where the graph G is just an edge and we have 3 colours. We simply apply Theorem 7.11 to the graph where vertices are "boxes" $x_1, \ldots, x_9$ from this picture, connected with blue edges which all have length 1. Transitions of Glauber dynamics are as labelled in the picture and we can see that the only closed class is the one shaded in blue $(x_1, x_3, x_5, x_6, x_8, x_9)$ i.e. the set of proper colourings S and π is uniform on it. The Glauber dynamics of the chain restricted to S will mix at least as fast as the chain which can also be started from outside of S. We note that we need to consider the larger graph as, for example, the ρ distance between x_1 and x_9 would be three if we remove x_2 and x_7 but we want this distance to be 2 as this is the number of vertices where the colours differ.

We define a coupling for (X_1, Y_1) when $X_0 = x$ and $Y_0 = y$ are colourings with $\rho(x, y) = 1$ (so X_1 and Y_1 are colourings after one step of the Glauber dynamics chains from x and y, respectively). Let v be the vertex where x and y differ. We pick the same uniform vertex w in both configurations and consider the following cases.

1. If w is not a neighbour of v or v, then we update both configurations in the same way since all such w's have the same allowed colours in both configurations. The distance of colourings after the update will stay 1.

2. If w is v we update the colour together in both configurations, which is allowed as both colourings have the same set of allowed colours for v since the colours of all vertices in them

agree. This decreases the distance of colourings to 0.

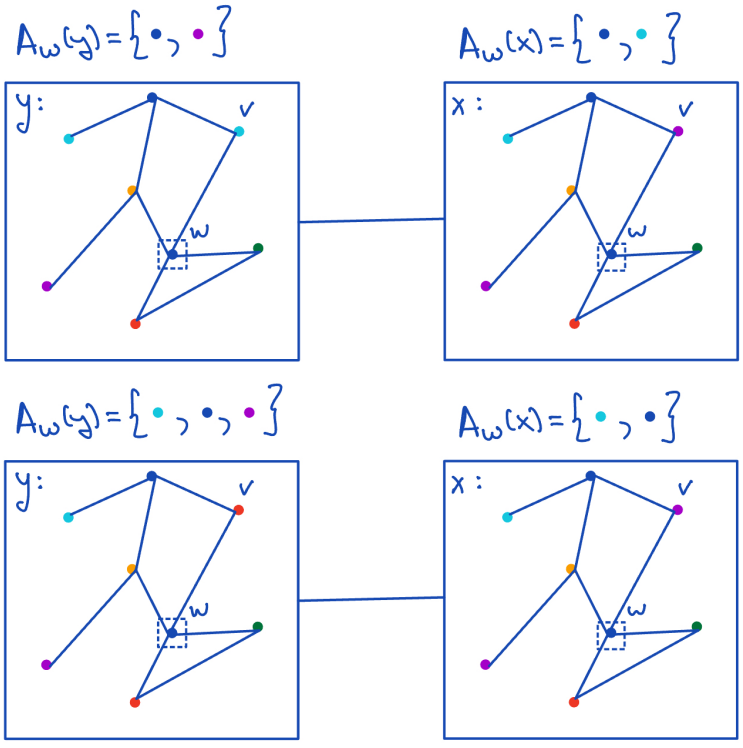
3. If $w \sim v$, then suppose without loss of generality that $|A_w(x)| \leq |A_w(y)|$ (differences of sizes is at most one as the colours of all vertices except v are same for the two colourings). We then pick a colour C uniformly at random from $A_w(y)$ and we have the following cases:

   (a) If $C \neq x(v)$, then we update $x(w)$ to C (this is allowed, as if another vertex $z \neq v$, $z \sim w$ had $x(z) = C$ then $y(z) = C$ which contradicts $C \in A_w(y)$). This distance of colourings after such an update stays 1.

   (b) If, however, $C = x(v)$, then we consider two further subcases:

   - if $|A_w(x)| = |A_w(y)|$, then we update $x(w)$ to $y(v)$ [in this case all other neighbours of w do not have colour which is either $x(v)$ or $y(v)$ and so we are updating v in x to a randomly picked colour from $A_w(x) = A_w(y) \cup y(v) \setminus x(v)$],
   - If $|A_w(x)| < |A_w(y)|$ then we update $x(w)$ to a uniform colour from $A_w(x)$. [in this case, we have $|A_w(y)| = |A_w(x)| + 1$ and the colour $y(v)$ is the same as the colour of some other neighbour of w (which are the same in x and y) while the colour $x(v)$ is not the same as the colour of any other neighbour of w].

   It is straightforward to check that $x(w)$ is updated to a uniform colour from $A_w(x)$ and we see that in both of these subcases, the distance of colourings has increased to 2 after the update.



An illustration of two different subcases in case (3b). In the first case, we identify purple in $A_w(y)$ with light blue in $A_w(x)$, while in the second case, if we pick purple in $A_w(y)$ as we do not have a corresponding colour for it in $A_w(x)$ we instead just take a uniform colour from $A_w(x)$ for an update of w in x [overall probability to pick light or dark blue in $A_w(x)$ is $\frac{1}{|A_w(y)|} + \frac{1}{|A_w(y)|} \cdot \frac{1}{|A_w(x)|} = \frac{|A_w(x)|+1}{|A_w(y)||A_w(x)|} = \frac{1}{A_w(x)}$ as wanted].

We now need to calculate $\mathbb{E}_{x,y}[\rho(X_1, Y_1)]$. We notice that the probability of picking v, i.e. case (2) has probability $\frac{1}{n}$ and was the only case where the distance decreased to 0. All other cases keep the distance at least 1 while only the case (3b), which happens with probability at most $\frac{\deg v}{n}\frac{1}{|A_w(y)|}$ increased it to 2. The case Putting all things together we obtain

$$\mathbb{E}_{x,y}[\rho(X_1, Y_1)] \leq \frac{1}{n}\sum_{w\sim v}\frac{1}{|A_w(y)|} + 1 - \frac{1}{n} = 1 - \frac{1}{n} + \frac{\deg(v)}{n}\max_{w\sim v}\frac{1}{|A_w(y)|}$$

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

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

where $\alpha(q, \Delta) = \frac{q-2\Delta}{q-\Delta}$, which is in $(0, 1)$ by the assumption $q > 2\Delta$. Theorem 7.11 together with the fact that the diameter of the graph of all colourings described above is n (as we can reach any colouring from any other by changing colours of vertices one by one, as we do not need to stay in the set of proper colourings) completes the proof.

□

If we have some not-too-complicated graph, such as a path, a cycle, or a star, we can, without too much effort, count the exact number of proper colourings with q colours. We can also, in general, sample proper colourings by doing a greedy method where we just colour one by one vertex with any of the colours we are allowed to use. However, if we consider a general graph with n vertices and maximal degree Δ there is no result which could tell us what the exact number of proper colourings is. As we know an upper bound on the mixing time of the Glauber dynamics chain and that its invariant distribution is a uniform colouring from the set of all proper colourings, we can actually sample a "close to uniform" proper colouring by starting from any proper colouring and then running the Glauber dynamics chain until time $\left\lceil \frac{q-\Delta}{q-2\Delta} n(\log n - \log \varepsilon)\right\rceil$ for some small ε. However, even if we can sample a "close to uniform" proper colouring that doesn't mean that we can get a good estimate on the total number of proper colourings (we can just get some upper bound). The below result tells us that there is a randomised algorithm which will with large probability give as an output a random variable which is a good estimate for a number of proper colourings. We call this approximate counting and the rest of this section is devoted to proving this result.

Lecture 16

Theorem 7.14 (Approximate counting of proper colourings)**.** ** Let G be a graph on n vertices with maximal degree Δ satisfying $q > 2\Delta$. Then, for any $\varepsilon, \eta > 0$ there exists a random variable W which can be simulated by running at most $n\left\lceil \frac{27qn}{\eta\varepsilon^2}\right\rceil\left\lceil \frac{n\log n + n\log(6eqn/\varepsilon)}{1-\frac{\Delta}{q-\Delta}}\right\rceil$ Glauber updates which satisfies*

$$\mathbb{P}\left((1-\varepsilon)|S|^{-1} \leq W \leq (1+\varepsilon)|S|^{-1}\right) \geq 1 - \eta,$$

where S is the set of all proper q-colourings of G.

The idea of the proof is to define a sequence of sets of proper colourings, S_k, run Glauber dynamics on them, and approximate $|S_{k-1}|/|S_k|$ and 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

$$S_k = \{x \in S : x(v_j) = x_0(v_j)\, \forall j > k\}$$

to be the set of all colourings where vertices with index $> k$ have the same colour as they have in the original colouring. We can notice that $S_0 = \{x_0\}$ and that $S_n = S$.

In the proof of Theorem 7.14, we will need to use that $|S_{k-1}|/|S_k|$ is not too small, so we will first prove that in the following lemma.

Lemma 7.15 (Bounding $\frac{|S_{k-1}|}{|S_k|}$)**.** ** Let $q > 2\Delta$. Then for above defined S_k and all $k \leq n$ we have*

$$\frac{|S_{k-1}|}{|S_k|} \geq \frac{1}{eq}.$$

Take a uniform element in S_k and run Glauber updates on neighbours of v_k and on v_k in order of increasing indices. Lower bound the probability that at the end you get a colouring in S_{k-1} (this is exactly the ratio we need as the final colouring is still uniform on S_k).

Proof. We bound the ratio by bounding the probability that the uniform element in S_k is also in S_{k-1} (which is exactly the ratio we want) and we do this as follows. Suppose that v_k has r neighbours in the set $\{v_1, \ldots, v_{k-1}\}$. Start with the uniform distribution on S_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. This preserves the uniform distribution on S_k.[6] Let Y be the configuration of colours at the end of this process. Then Y is uniform on S_k. Let A be the event that at the end of this process, the colour at each of the r neighbours is different from $x_0(v_k)$ and the colour of v_k is updated to $x_0(v_k)$. Then $Y \in S_{k-1}$ if and only if the event A occurs. So we have

$$\frac{|S_{k-1}|}{|S_k|} = \mathbb{P}(Y \in S_{k-1}) = \mathbb{P}(A) \geq \left(\frac{q-\Delta-1}{q-\Delta}\right)^r \frac{1}{q} \geq \left(\frac{\Delta}{1+\Delta}\right)^\Delta \frac{1}{q} \geq \frac{1}{eq},$$

where for the first inequality, we used that the size of the set of allowed colours for every neighbour of v_k is at least $q - \Delta$ and we can not use colour $x_0(v_k)$. For the penultimate inequality, we used the assumption that $q \geq 2\Delta + 1$ and $r \leq \Delta$. □

For each k to approximate $\frac{|S_{k-1}|}{|S_k|}$ run $a_n = \left\lceil \frac{27qn}{\eta\varepsilon^2} \right\rceil$ independent copies of Glauber dynamics on S_k each for $\left\lceil \frac{n\log n + n\log(6eqn/\varepsilon)}{1-\frac{\Delta}{q-\Delta}} \right\rceil$ steps independently for different k's and get one sample at the end of each run [a_n needs to be picked to be large enough such that the ratio of variance and mean of W_n is small enough]. Let $Z_{k,i} = \mathbb{1}_{\{i\text{-th sample is in } S_{k-1}\}}$ and $W_k = \frac{1}{a_n}\sum_{i=1}^{a_n} Z_{k,i}$. From the mixing time bound we have that $\mathbb{E}[Z_{k,i}] = \mathbb{E}[W_k]$ is close to $|S_{k-1}|/|S_k|$. As W_k is an average of i.i.d random variables it should be concentrated around its mean for large enough a_n (this is how we picked a_n). $W = \prod_{k=0}^n W_k$ approximates the quantity we want and we can show using Chebychev's that it is concentrated around its mean as well. To apply Chebychev's bound the ratio of variance and mean squared of W and for this you use Lemma 7.15 and that W_k are independent.

Proof of Theorem 7.14.* First recall that $|S_0| = 1$ and $|S_n| = |S|$. So we have

$$\prod_{i=0}^{n-1} \frac{|S_i|}{|S_{i+1}|} = \frac{1}{|S|}.$$

[6] If we are changing only the colour of vertex v, then for any proper colouring x there are $A_v(x)$ colourings from which, when changing only v, one can obtain x (because all vertices other than v have to have the same colour as their colour in x and v and be any of the $A_v(x)$ allowed colours). As for each of these $A_v(x)$ colourings the probability of changing the colour of v to exactly $x(v)$ is $1/A_v(x)$, it means that after one step from a uniform colouring, we are still in a uniform colouring.

The strategy of the proof is to define a random variable W_i which will be close to $|S_{i-1}|/|S_i|$ with high probability. Then we will define $W = \prod_{i=1}^{n} W_i$ and get that it will be close to $1/|S|$. The random variable W will be concentrated around its mean (the standard way of showing something like this for a random variable is using Chebychev's inequality and that is exactly what we will do) and we will just have to show that the mean is close enough to $1/|S|$.

To approximate $|S_{i-1}|/|S_i|$ we will sample many times a "close to uniform" random sample from S_i and then find the proportion of samples which also belong to S_{i-1}. The average of i.i.d. random variables is often a random variable which is concentrated around its expectation, so this proof strategy seems viable. We just have to figure out how to get a "close to random" sample from S_i. This will, of course, be done by running enough steps of the Glauber dynamics for long enough, and from the estimate on the mixing time, we will be able to decide how long we need to run it for. We now do these steps carefully:

Running Glauber dynamics on S_k (with frozen boundary conditions at the vertices $v_{k+1}, \dots, v_n$) will generate a uniform element of S_k. The same proof as in Theorem 7.13 gives that the mixing time of this chain is bounded as follows:

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

So starting from any colouring the distribution of Glauber dynamics on S_k at time t is within $\varepsilon/(6eqn)$ in total variation from the uniform distribution on S_k.

We now take $a_n = \left\lceil \frac{27qn}{\eta\varepsilon^2} \right\rceil$ and generate a_n samples by taking independent copies of Glauber dynamics on S_k each run for t steps independently for different k's. For $i = 1, \dots, a_n$ we let

$$Z_{k,i} = \mathbb{1}_{\{i\text{-th sample is in } S_{k-1}\}} \quad \text{and} \quad W_k = \frac{1}{a_n}\sum_{i=1}^{a_n} Z_{k,i}.$$

Note that to check if i-th sample is in S_{k-1} one just needs to check that the colour of v_k is $x_0(v_k)$. Using the mixing property at time t we get that if π_k is the invariant distribution for the Glauber dynamics on S_k and d^k is the distance to stationarity for this chain then as i-th sample is just this chain at time t we have

$$\left|\mathbb{E}[Z_{k,i}] - \frac{|S_{k-1}|}{|S_k|}\right| = |\mathbb{P}_{x_0}(i\text{-th sample } \in S_{k-1}) - \pi_k(S_{k-1})| \leq d^k(t) \leq \frac{\varepsilon}{6eqn}$$

$$\text{and so} \quad \left|\mathbb{E}[W_k] - \frac{|S_{k-1}|}{|S_k|}\right| \leq \frac{\varepsilon}{6eqn}. \tag{7.3}$$

The second inequality together with Lemma 7.15 now gives

$$1 - \frac{\varepsilon}{6n} \leq \mathbb{E}[W_k]\,\frac{|S_k|}{|S_{k-1}|} \leq 1 + \frac{\varepsilon}{6n}. \tag{7.4}$$

We now define $W = \prod_{i=1}^{n} W_i$. We will shortly show that each W_k is concentrated around its mean, which is close to $|S_{k-1}|/|S_k|$. So by taking the product of W_i's as the definition of W we will get that W is close to the product of $|S_{k-1}|/|S_k|$ for $k = 1, \dots, n$, which is equal to $1/|S|$, since $|S_0| = 1$ and $|S_n| = |S|$. Using (7.4) and that $\mathbb{E}[W] = \prod_{i=1}^{n}\mathbb{E}[W_i]$ holds by the independence of W_i we get

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

Therefore,

$$\left|\mathbb{E}[W] - \frac{1}{|S|}\right| \leq \frac{\varepsilon}{3|S|} \tag{7.5}$$

and so to approximate $|S|$ we just need to be able to approximate $\mathbb{E}[W]$ and as mentioned, we show that we can approximate it by W using Chebychev's (i.e. concentration of W around its mean).

To apply Chebychev's we need to control the variance. Using the independence of W_k's we get

$$\frac{\mathrm{Var}\,(W)}{(\mathbb{E}[W])^2} = \frac{\mathbb{E}\big[W^2\big] - (\mathbb{E}[W])^2}{(\mathbb{E}[W])^2} = \frac{\prod_{k=1}^n \mathbb{E}\big[W_k^2\big]}{\prod_{k=1}^n (\mathbb{E}[W_k])^2} - 1 = \prod_{k=1}^n \left(1 + \frac{\mathrm{Var}\,(W_k)}{(\mathbb{E}[W_k])^2}\right) - 1. \tag{7.6}$$

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

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

where in the second inequality we have used that

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

which follows from the Lemma 7.15 and (7.3). Plugging this bound into (7.6) we obtain

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

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

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

Using equation (7.5) we now see that on the event that $\{|W - \mathbb{E}[W]| \leq \varepsilon\mathbb{E}[W]\,/2\}$ we have

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

Therefore, 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 S_k for $k = 1, \ldots, n$ and this concludes the proof.

□

8 Coupling from the past* (non-examinable section)

8.1 Algorithm

Coupling from the past (by Propp and Wilson, 1996) is an algorithm for exact sampling from the invariant distribution π. To describe it we start with the random function representation of a Markov chain with transition matrix P.

Lemma 8.1 (Random function representation of Markov Chains)**.** *Let P be a transition matrix on a finite state space S. There exists a function $f : S \times [0,1] \to S$ such that if $(U_i)_{i\geq 0}$ are i.i.d. sequence of uniform random variables on $[0,1]$, and X_0 is some random variable on S, then a sequence*

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

is a Markov chain with transition matrix P.

If $S = \{x_1, \dots x_n\}$ for any fixed i subdivide $[0,1]$ into n intervals of lengths $P(x_i, x_j)$ for $j \leq n$ and use this to update the chain from x_i.

Proof. We label the state space as $S = \{x_1, \dots, x_n\}$ and let $F_{j,k} = \sum_{i=1}^{k} P(x_j, x_i)$ for all $j, k \leq n$. Define the function f by

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

We have that for a uniform random variable U on $[0,1]$

$$\mathbb{P}(f(x_j, U) = x_k) = \mathbb{P}(F_{j,k-1} < U \leq F_{j,k}) = P(x_j, x_k),$$

which, as U_i are i.i.d. variables with the same distribution as U completes the proof. □

We should think of 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 $\{U_{t,x} : x \in S, t \in \mathbb{Z}\}$ be a collection of uniform random variables in $[0,1]$ with the property that $(U_{t,x} : x \in S)_{t\in\mathbb{Z}}$ are i.i.d. Note that we do not specify the dependence between $U_{t,x}$ for different $x \in S$. For every $t \in \mathbb{Z}$ we let $f_t : S \to S$ be given by

$$f_t(x) = f(x, U_{t,x}),$$

where f is the function from Lemma 8.1. So, from Lemma 8.1 the process X which evolves as $X_{t+1} = f_t(X_t)$ is a Markov chain. We stress here that f_t is a random function that depends on a set of random variables $U_{x,t}$ and conditional on some fixed values of uniform random variables $U_{x,t}$ it is a deterministic function from S to S.

Definition 8.2 (Forward and backward maps)**.** For integers $s < t$ (which could be negative) we define

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

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* and for $t > 0$ the maps F_{-t}^0 are called the *backward maps* for the chain. 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 $F_{-s}^0 = F_{-t}^0 \circ (f_{-t-1} \circ \dots \circ f_{-s})$ (i.e. since to get F_{-s}^0 you first apply some other functions, and then you apply some function F_{-t}^0 which sends all states to one constant value then the output will be that constant value, i.e. if all states from some point in the past are sent to the same constant state then from further in the past all states need to be sent to this same constant state). //

The idea of *coupling from the past* algorithm 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 8.3 (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 for*

$$T = \min\{t \geq 0 : F_{-t}^0 \text{ is a constant function}\}$$

either $\mathbb{P}(T < \infty) = 1$ or $\mathbb{P}(T < \infty) = 0$. In the case when $\mathbb{P}(T < \infty) = 1$, then $F_{-T}^0(S) \sim \pi$, i.e., the unique value that F_{-T}^0 takes is distributed according to π.

Remark 8.4. *We note that it is crucial that in the algorithm above we consider time backwards. Indeed, if not, then if F_0^t is constant for $t > 0$, then of course F_0^{t+1} would also be a constant function, but it would not necessarily be equal to F_0^t (as to get F_0^{t+1} we apply a function to F_0^t and this function might send the constant value of F_0^t to another constant). So we could define a random time T from which the functions become constant, however, at that random time we would not necessarily have a chain which is distributed according to π.*

Random functions $F_{-L}^0, F_{-2L}^{-L}, F_{-3L}^{-2L}$ are i.i.d. so if there is L such that they have a positive probability of being constant then with probability 1 we have $T < \infty$. For any states $\mathbb{P}\big(F_{-t}^0(x) = y\big) \to \pi(y)$ as $t \to \infty$ and as F_{-t}^0 is the same constant value for $t > T$ the proof follows.

Proof of Theorem 8.3. We start by showing that if $\mathbb{P}(T < \infty) > 0$, then $\mathbb{P}(T < \infty) = 1$. Indeed, there must exist L sufficiently large and $\varepsilon > 0$ so that

$$\mathbb{P}\big(F_{-L}^0 \text{ is a constant function}\big) \geq \varepsilon. \tag{8.1}$$

Since the collection $\{U_{t,x} : x \in S\}_{t \in \mathbb{Z}}$ is i.i.d., it follows that the events

$$\{F_{-L}^0 \text{ is constant}\}, \{F_{-2L}^{-L} \text{ is constant}\}, \{F_{-3L}^{-2L} \text{ is constant}\} \dots$$

are independent and have the same probability. So by (8.1) we get that each of the events above has probability at least ε of happening and since they are also independent, we get that, with probability 1, at least one of them will eventually happen, proving that $P(T < \infty) = 1$.

Suppose now that $\mathbb{P}(T < \infty) = 1$. For all $t > 0$ and all x, y we have

$$\mathbb{P}\big(F_{-t}^0(x) = y\big) = P^t(x, y).$$

Since X is assumed to be irreducible and aperiodic, from the above it follows that

$$\lim_{t \to \infty} \mathbb{P}\big(F_{-t}^0(x) = y\big) = \pi(y). \tag{8.2}$$

As mentioned earlier, if there exists $t > 0$ such that 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. Therefore, we obtain that for all $t > T$, we have

$$F_{-t}^0 = F_{-T}^0.$$

We can now finish the proof by writing for every $t > 0$

$$\begin{aligned}\mathbb{P}\big(F_{-T}^0(S) = x\big) &= \mathbb{P}\big(F_{-T}^0(S) = x, T > t\big) + \mathbb{P}\big(F_{-T}^0(S) = x, T \leq t\big), \\ &= \mathbb{P}\big(F_{-T}^0(S) = x, T > t\big) + \mathbb{P}\big(F_{-t}^0(S) = x\big),\end{aligned}$$

where we used that if $F_{-t}^0(S)$ is a single point, then $T \leq t$. Taking the limit as $t \to \infty$ on the right-hand side above we get that

$$\mathbb{P}\big(F_{-T}^0(S) = x\big) = \pi(x),$$

where we used (8.2) for the second term and the assumption that $P(T < \infty) = 1$ for the first one. This concludes the proof. □

Remark 8.5. *If for every fixed t the variables $\{U_{t,x}\}_{x\in S}$ are also independent, then if the Markov chain is irreducible and aperiodic it will follow that there exists L sufficiently large and $\varepsilon > 0$ such that*

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

So in this case we get that $\mathbb{P}(T < \infty) = 1$. [This follows from "Useful results for Mixing times of Markov chains", Question 2 as there is a time $L > 0$ such that $P^L(x, y) > 0$ for all states x, y, so there is a positive probability that all uniforms were sampled in such a way to send all x to the same state.]

Claim 8.6 (T is distributed as coalescence time)**.** *Let C be a random time, known as the coalescence time which is the first time t such that F^t_0 is constant, i.e. $C = \min\{t \geq 0 : F^t_0 \text{ is a constant function}\}$. Then C and the time T from Theorem 8.3 have the same distribution.*

Follows directly from the fact that F^t_0 and F^0_{-t} have the same distribution.

Proof. As for all $t > 0$ we know that F^t_0 and F^0_{-t} have the same distribution we get for all $t > 0$

$$\mathbb{P}(T > t) = \mathbb{P}\big(F^0_{-t} \text{ is not constant}\big) = \mathbb{P}\big(F^t_0 \text{ is not constant}\big) = \mathbb{P}(C > t)$$

which completes the proof. □

In the next subsection, we will see that we can bound the average running time of the coupling from the past algorithm using the mixing time, in the case of certain chains, known as the monotone chains.

8.2 Monotone chains

Definition 8.7 (Monotone chains)**.** 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. //

We now consider a monotone chain on a partially ordered set $(S, \preceq)$ and suppose that it contains two elements $\widehat{0}$ and $\widehat{1}$ such that $\widehat{0} \preceq x \preceq \widehat{1}$ 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 8.8 (Bound on $\mathbb{E}[T]$ for monotone chains)**.** *For a monotone chain with a largest and smallest states $\widehat{0}$ and $\widehat{1}$ and function f from the definition of backward and forward map which preserves the monotonicity and using the same variables $U_{t,x}$ for all values of x, 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 for T from the Theorem 8.3 we have*

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

Show $\mathbb{P}(C > ik) \leq \mathbb{P}(C > k)^i$ and so bound $\mathbb{E}[C] = \sum_{i=0}^{\infty} \mathbb{P}(C > i) \leq \frac{k}{\mathbb{P}(C \leq k)}$ and also show that $\mathbb{P}(C > k) \leq \mathbb{E}[h(Y_k) - h(X_k)] \leq \ell \bar{d}(k)$ by considering optimal coupling of the law of X_k, Y_k.

Proof. Using Claim 8.6 it suffices to prove the bound of the statement for $\mathbb{E}[C]$. We start by proving that for all k

$$\frac{\mathbb{P}(C > k)}{\ell} \leq \bar{d}(k) \leq \mathbb{P}(C > k)\,, \tag{8.3}$$

where $\bar{d}$ is defined in Definition 1.8. 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 $X_t \preceq Y_t$ and $X_t \neq Y_t$, then $h(X_t) + 1 \leq h(Y_t)$. So we have

$$\mathbb{P}(C > k) = \mathbb{P}(X_k \neq Y_k) = \mathbb{P}(h(Y_k) - h(X_k) \geq 1) \leq \mathbb{E}[h(Y_k) - h(X_k)]\,,$$

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

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

where $\mathcal{L}(Z)$ denotes the law of a random variable Z. This proves the first inequality in (8.3). The second one follows immediately from the coupling upper bound on total variation distance from Theorem 1.7.

We now 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 the independence of these last two events gives the sub-multiplicativity.

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

$$\mathbb{E}[C] = \sum_{i=0}^{\infty} \mathbb{P}(C > i) \leq \sum_{i=0}^{\infty} k\mathbb{P}(C > ik) \leq \sum_{i=0}^{\infty} k\mathbb{P}(C > k)^i = \frac{k}{\mathbb{P}(C \leq k)}. \tag{8.4}$$

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

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

From (8.3) we obtain

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

and hence from (8.4) we deduce

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

and this completes the proof. □

The above theorem showed that for a monotone chain, we can bound the running time of the coupling from the past algorithm using the mixing time of the chain. To conclude this section, we give an example of a very well-known model, which also happens to be a monotone chain.

8.3 Ising model

We start by defining general Glauber dynamics.

Definition 8.9 (Glauber dynamics)**.** Let V and Q be two finite sets. Let S be a subset of Q^V and π a distribution on S. The *Glauber dynamics* on S is the Markov chain that evolves as follows: when at state $x \in S$, we pick a vertex v 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 $x \in S$ and $v \in V$ we define

$$A(x,v) = \{y \in S : y(w) = x(w), \forall w \neq v\} \quad \text{and} \quad \pi_{x,v}(y) = \frac{\pi(y)}{\pi(A(x,v))} \mathbb{1}_{\{y \in A(x,v)\}}.$$

When at state $x \in S$, 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 8.10. *It is easy to check that Glauber dynamics gives a reversible Markov chain with the invariant distribution π.*

We should also note that the above definition agrees with the one we saw in the context of proper graph colourings (Definition 7.12) but is just more general.

Definition 8.11 (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

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

where $\sigma \in \{-1, 1\}^V$ is called a *spin configuration*. The parameter $\beta \geq 0$ 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 $\beta = 0$, 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

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

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

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 (as $\frac{1+\tanh(\beta S_v(\sigma))}{2} = \frac{e^{\beta S_v(\sigma)}}{e^{\beta S_v(\sigma)} + e^{-\beta S_v(\sigma)}}$ is the probability of updating the spin at v to $+1$). Since $\frac{1+\tanh(\beta S_v(\sigma))}{2}$ is non-decreasing in σ, it follows that the coupling is monotone.

This means that we can apply Theorem 8.8 to the Ising model and if we know the mixing time of the chain we would also be able to bound the expected running time of the coupling from the past algorithm. Note that in this case it can easily be seen that ℓ from Theorem 8.8 is the number of vertices in V.

9 Hit-mix

Lecture 17

In this section, we will see how one can use the hitting time to bound the mixing time. We will work with reversible Markov chains, even though similar but weaker results can be shown in the case of a non-reversible chain. Loosely speaking, we will start by defining a *ε-hit* times, which are going to measure the amount of time it takes for the Markov chain to hit (visit) all large enough sets with at least ε probability. We will get some results about these hit times first. We will then present technical proof that, under the reversibility assumption, allows us to compare the hit and mixing times of a chain. In fact, the comparison will end up being so sharp that we will be able to show that, as long as the relaxation time is smaller order than the mixing/hit time, the cutoff property is equivalent for the mixing times and the hit times (we define the hit cutoff in the same way as mixing times cutoff just using hit times in place of the mixing times). We will then give some examples where this technique can be used.

9.1 Hit times

Definition 9.1 (Hit times)**.** Let P be a finite, irreducible chain with invariant distribution π and state space S. For any $\alpha, \varepsilon \in (0,1)$ and $t \geq 0$ we define the maximal probability of not hitting a set of size at least α by time t starting from x as

$$p_x(\alpha, t) = \max_{A \subset S, \pi(A) \geq \alpha} \mathbb{P}_x(T_A > t)$$

where T_A is the hitting time of the set A defined formally as $T_A = \inf\{t : X_t \in A\}$. We also set $p(\alpha, t) = \max_{x \in S} p_x(\alpha, t)$. Finally, we define a hit time (more precisely $\varepsilon - \mathrm{hit}_\alpha$ time) starting from x and from the worst starting point, respectively, as

$$\mathrm{hit}_{\alpha,x}(\varepsilon) = \min\{t : p_x(\alpha, t) \leq \varepsilon\} \quad \text{and} \quad \mathrm{hit}_\alpha(\varepsilon) = \min\{t : p(\alpha, t) \leq \varepsilon\}.$$

So the $\varepsilon - \mathrm{hit}_\alpha$ time is just the smallest time t such that regardless of the starting point, for any set of π measure at least α, the Markov chain will by time t hit that set with probability at least $1 - \varepsilon$. //

Intuitively it is not so surprising that the hit times and mixing times are somehow connected quantities. This is because if the chain is ε mixed by time t, then even at this fixed time t it is distributed close (in total variation distance) to π which means that the probability that the chain at that time is not at some set of π measure a lot larger than ε can not be too close to 1. This simple argument already allows us to lower bound mixing times by hit times.

Claim 9.2 (Lower bound on mixing time by hit time)**.** *For any chain (even non-reversible) we have that for all $\varepsilon \in (0,1), \delta \in (0, 1-\varepsilon)$*

$$t_{\mathrm{mix}}(\varepsilon) \geq \mathrm{hit}_{\varepsilon+\delta}(1-\delta).$$

Lower bound TV distance by $\pi(A) - P^t(x, A)$ for x, A with $\pi(A) \geq \varepsilon + \delta$ and $\mathbb{P}_x(T_A \leq s) > \delta$ which exists for $t < \mathrm{hit}_\alpha(1-\delta)$.

Proof. If $s < \mathrm{hit}_{\varepsilon+\delta}(1-\delta)$ then by the definition of hit times there is a set A with $\pi(A) \geq \varepsilon + \delta$ and $x \in S$ such that $\mathbb{P}_x(T_A \leq s) < \delta$. But then, as $P^s(x, A) = \mathbb{P}_x(X_s \in A) \leq \mathbb{P}_x(T_A \leq s)$

$$\|\pi - P^s(x, \cdot)\|_{\mathrm{TV}} \geq \pi(A) - P^s(x, A) \geq \pi(A) - \mathbb{P}_x(T_A \leq s) > \varepsilon$$

and so $s < t_{\mathrm{mix}}(\varepsilon)$. □

Proposition 9.3 (Hit times submultiplicativity)**.** *For any chain (even non-reversible) for any* $\alpha, \varepsilon, \delta \in (0,1)$ *we have*

$$\mathrm{hit}_\alpha(\varepsilon\delta) \le \mathrm{hit}_\alpha(\varepsilon) + \mathrm{hit}_\alpha(\delta).$$

Notice that $\mathbb{P}_x(T_A > t+s) \le \mathbb{P}_x(T_A > t) \max_{z\in S} \mathbb{P}_x(T_A > s)$ shows $p(\alpha, t+s) \le p(\alpha,t)p(\alpha,s)$.

Proof. Let $A \subset S$ such that $\pi(A) \ge \alpha$, $x \in S$ and $t, s \ge 0$. We have, using Markov property and conditioning on the location of the chain at time t and maximising over it, that

$$\mathbb{P}_x(T_A > t+s) \le \mathbb{P}_x(T_A > t) \max_{z\in S} \mathbb{P}_x(T_A > s) \le p(\alpha,t)p(\alpha,s)$$

so $p(\alpha, t+s) \le p(\alpha,t)p(\alpha,s)$. The proof follows by taking $s = \mathrm{hit}_\alpha(\delta)$ and $t = \mathrm{hit}_\alpha(\varepsilon)$. □

Claim 9.4 (Hit times are decreasing in α)**.** *For any* $\delta \in (0,1)$ *and* $0 < \alpha \le \beta \le 1$ *we have*

$$\mathrm{hit}_\beta(\delta) \le \mathrm{hit}_\alpha(\delta).$$

Direct from the definition.

Proof. From the definition we can see that $p(\alpha,t) \ge p(\beta,t)$ and so when $p(\alpha,t)$ drops below δ we must also have $p(\beta,t) \le \delta$ and the claim follows. □

We will now conclude this subsection by showing that for $\alpha < \beta$, when the chain is reversible, one can also upper bound $\mathrm{hit}_\alpha(\delta)$ by $\mathrm{hit}_\beta(\delta - \varepsilon)$ and the relaxation time. The result will tell us that we can increase the parameter α at the cost of (decreasing) parameter δ. We will first need the following technical lemma, which we will prove using an example sheet result.

Lemma 9.5 (Hitting times bound by the relaxation time, reversible chains)**.** *For a reversible transition matrix* P *on* S*, with invariant distribution* π *and relaxation time* t_{rel} *we have for a non-empty* $A \subset S$ *that for all* $t \ge 0$

$$\mathbb{P}_\pi(T_A > t) \le \pi(A^c)\left(1 - \frac{\pi(A)}{t_{\mathrm{rel}}}\right)^t \le \pi(A^c)\exp\left(-\frac{t\pi(A)}{t_{\mathrm{rel}}}\right).$$

In particular, for $c, w > 0$ *let* $B(A,w,c) = \left\{y \in S : \mathbb{P}_y\left(T_A > \left\lceil \frac{t_{\mathrm{rel}} w}{\pi(A)} \right\rceil\right) \ge c\right\}$ *then*

$$\pi(B(A,w,c)) \le \frac{\pi(A^c)}{e^w c} \quad \text{and} \quad \pi(A)\mathbb{E}_\pi[T_A] \le t_{\mathrm{rel}}\pi(A^c). \tag{9.1}$$

Example Sheet 2, Question 10 tells you that this holds for connected A^c. Apply that to the connected components of A^c. Second part follows directly from the first.

Proof. To prove the first statement we use Example Sheet 2, Question 10 stating that if $B = A^c$ is connected set then we have

$$\mathbb{P}_{\pi_B}(T_A > t) \le \left(1 - \frac{\pi(A)}{t_{\mathrm{rel}}}\right)^t \le \exp\left(-\frac{t\pi(A)}{t_{\mathrm{rel}}}\right)$$

where for a set X, π_X is just π conditioned on being in X. This is almost the result we need to prove, except that we also need to deal with the case when the set B is not connected. To do this

we work with the connected components $C_1, \ldots, C_k$ of A^c. Example sheet result (applied to C_i^c which is valid as its complement, C_i, is connected) now gives

$$\begin{aligned}\mathbb{P}_\pi(T_A > t) &= \sum_{i=1}^k \pi(C_i)\mathbb{P}_{\pi_{C_i}}(T_A > t) = \sum_{i=1}^k \pi(C_i)\mathbb{P}_{\pi_{C_i}}\left(T_{C_i^c} > t\right) \\ &\leq \sum_{i=1}^k \pi(C_i)\left(1 - \frac{\pi(C_i^c)}{t_{\text{rel}}}\right)^t \leq \sum_{i=1}^k \pi(C_i)\left(1 - \frac{\pi(A)}{t_{\text{rel}}}\right)^t \\ &= \pi(A^c)\left(1 - \frac{\pi(A)}{t_{\text{rel}}}\right)^t \leq \pi(A^c)\exp\left(-\frac{t\pi(A)}{t_{\text{rel}}}\right).\end{aligned}$$

We now show how this gives (9.1). If $B = B(A, w, c)$ and $t = t(A, w) = \left\lceil \frac{t_{\text{rel}} w}{\pi(A)} \right\rceil$ by the above and the definition of set B, we have

$$\pi(B)c \leq \pi(B)\mathbb{P}_{\pi_B}(T_A > t) \leq \mathbb{P}_\pi(T_A > t) \leq \pi(A^c)\exp\left(-\frac{t\pi(A)}{t_{\text{rel}}}\right) \leq \pi(A^c)e^{-w}$$

proving the first inequality of (9.1). For the second inequality, we write

$$\mathbb{E}_\pi[T_A] = \sum_{t=0}^\infty \mathbb{P}_\pi(T_A > t) \leq \sum_{t=0}^\infty \pi(A^c)\left(1 - \frac{\pi(A)}{t_{\text{rel}}}\right)^t = \pi(A^c)\frac{t_{\text{rel}}}{\pi(A)},$$

which completes the proof. □

Proposition 9.6 (Comparing hit_α and hit_β, for reversible chains)**.** *For a reversible chain, with relaxation time t_{rel}, any $0 < \varepsilon < \delta < 1$ and $0 < \alpha \leq \beta < 1$ we have the following comparison of hit times*

$$\text{hit}_\alpha(\delta) \leq \text{hit}_\beta(\delta - \varepsilon) + \left\lceil \alpha^{-1} t_{\text{rel}} \log\left(\frac{1-\alpha}{(1-\beta)\varepsilon}\right)\right\rceil. \tag{9.2}$$

Using Lemma 9.5 $\widehat{A} = \{x \in S : \mathbb{P}_x(T_a > s)\}$ for $s \left\lceil \alpha^{-1} t_{\text{rel}} \log\left(\frac{1-\alpha}{(1-\beta)\varepsilon}\right)\right\rceil$ has π measure at least β for $\pi(A) \geq \alpha$ and we use the bound $\mathbb{P}_x(T_A > s + t) \leq \mathbb{P}_+\left(T_{\widehat{A}} > t\right) \max_{y \in \widehat{A}} \mathbb{P}_y(T_A > s)$.

Proof. The key of this proof will be the first inequality of (9.1). Fix A with $\pi(A) \geq \alpha$. This result gives an upper bound on the π measure of a set of points from which the probability of not reaching A in a certain time, which we will now label with s, is at least some constant c. By considering the complement, this also means that we can lower bound the π measure of a set from which we do not reach A by time s with probability at most $1 - c$. Lets define

$$\widehat{A} = \{x \in S : \mathbb{P}_x(T_A > s) \leq \varepsilon\}.$$

We notice that $\widehat{A}$ contains A and possibly some other points and, as we just explained, we can lower bound $\pi(\widehat{A})$. In fact, by increasing s we can see that (9.1) would give a lower bound on $\pi(\widehat{A})$ which can be made arbitrarily close to 1. We also notice that starting from a vertex in $\widehat{A}$ we know that we hit A in time s with probability at least $1 - \varepsilon$. Our goal is to compare $\text{hit}_\alpha(\delta)$ and $\text{hit}_\beta(\delta - \varepsilon)$ where $\beta \geq \alpha$. If we let $t = \text{hit}_\beta(\delta - \varepsilon)$ and if we pick s large enough such that $\pi(\widehat{A}) \geq \beta$ we have, using the union bound, that

$$\begin{aligned}\max_{x \in S} \mathbb{P}_x(T_A > s + t) &\leq \max_{x \in S} \mathbb{P}_x\left(T_{\widehat{A}} > t \cup T_A^{X_{T_{\widehat{A}}}} > s\right) \leq \max_{x \in S}\left(\mathbb{P}_x\left(T_{\widehat{A}} > t\right) + \mathbb{P}_x\left(T_A^{X_{T_{\widehat{A}}}} > s\right)\right) \\ &\leq \max_{x \in S} \mathbb{P}_x\left(T_{\widehat{A}} > t\right) + \max_{y \in \widehat{A}} \mathbb{P}_y(T_A > s) \leq (\delta - \varepsilon) + \varepsilon = \delta\end{aligned}$$

where $T_A^{X_{T_{\widehat{A}}}}$ is the time it takes for the chain to hit set A starting from $X_{T_{\widehat{A}}}$ [7]. We note that the first inequality above holds as in the case $T_A > s + t$ then either it takes time greater than t to hit $\widehat{A}$ or it takes more than s to hit set A starting from the location in which the chain hits set $\widehat{A}$. Final inequality is true by the definition of t and as $\pi(\widehat{A}) > \beta$ and as from the definition of $\widehat{A}$ for any $y \in \widehat{A}$ we have $\mathbb{P}_y(T_A > s) \leq \varepsilon$. From the definition of $\text{hit}_\alpha(\delta)$ this gives that

$$\text{hit}_\alpha(\delta) \leq t + s = \text{hit}_\beta(\delta - \varepsilon) + s$$

and so we are only required to carefully use equation (9.1) to get a bound on s.

We just need to take minimal possible s such that for any A with $\pi(A) \geq \alpha$, we have $\pi(\widehat{A}) \geq \beta$. Using notation from (9.1) we see that $\widehat{A} = B(A, w, \varepsilon)^c$ for $w = \frac{s\pi(A)}{t_{\text{rel}}}$ (as then $s = \left\lceil \frac{t_{\text{rel}} w}{\pi(A)} \right\rceil$) and so (9.1) gives

$$\pi(\widehat{A}) \geq 1 - \frac{\pi(A^c)}{e^w \varepsilon} \geq 1 - \varepsilon^{-1}(1 - \pi(A)) \exp\left(-\frac{s\pi(A)}{t_{\text{rel}}}\right) \geq 1 - \varepsilon^{-1}(1 - \alpha) \exp\left(-\frac{s\alpha}{t_{\text{rel}}}\right)$$

where for the third inequality we have used that $\alpha \leq \pi(A)$. To be able to lower bound the above by β we need to pick s such that $\varepsilon^{-1}(1 - \alpha) \exp\left(-\frac{s\alpha}{t_{\text{rel}}}\right) \leq 1 - \beta$ and so we see that taking $s = \left\lceil \alpha^{-1} t_{\text{rel}} \log\left(\frac{1-\alpha}{(1-\beta)\varepsilon}\right) \right\rceil$ works and this completes the proof. □

9.2 Hit-mix bound

We have already seen a lower bound on the mixing time by hit time in Claim 9.2. We will now want to establish an upper bound on the mixing time in terms of some hit time.

We start this section by stating two results we need for relating hit and mixing times. The first one is Example Sheet 2, Question 5a and the second one is a result which is proved using martingales, so a proof of it will not be examinable but for the sake of completeness, it will be presented at the end of this subsection. We will then give some motivation for the hit-mix comparison, prove a proposition which will be key to establishing hit-mix comparison and finally, we will establish the wanted comparison.

Recall that the Example Sheet 2 Question 5a states the following.

Lemma 9.7 (Generalised Poincaré inequality)**.** *For a reversible transition matrix P with invariant distribution π and any $f : S \to \mathbb{R}$ we have for all $t \geq 0$ that*

$$\text{Var}_\pi\left(P^t f\right) \leq e^{-2t/t_{\text{rel}}} \text{Var}_\pi(f).$$

We will also need the following, which we prove at the end of the section (and the proof is nonexaminable).

Theorem 9.8 (Starr's maximal inequality)**.** *For a reversible, irreducible, finite transition matrix P with invariant distribution π, $p \in (1, \infty)$ and $p^* = \frac{p}{p-1}$ we have that for any $f : S \to \mathbb{R}$*

$$\|f^*\|_p \leq p^* \|f\|_p,$$

where $f^* : S \to \mathbb{R}$ is the corresponding maximal function at even times, defined as

$$f^*(x) = \sup_{0 \leq t < \infty} |P^{2t}(f)(x)| = \sup_{0 \leq t < \infty} |\mathbb{E}_x[f(X_{2t})]|.$$

[7] of course here we are using the Strong Markov property i.e. that conditional on where we hit $\widehat{A}$ we can just start a new chain with transition P from there – this works as hitting times as stopping times so $T_{\widehat{A}}$ is a stopping time

The main idea in comparing mixing times and hit times is to show that after reaching all large sets with good probability the additional time it takes for the chain to mix is somehow controlled by the relaxation time. To do this we define for any fixed set A a good set for A, which satisfies that if the chain starts from the good set, the probability of it being in A after some time s is close to $\pi(A)$. So for any A if we can show that the good set for A has π measure at least $1-\delta$ then at time $t = \text{hit}_{1-\delta} + s$ we can bound $|P^t(x,A) - \pi(A)|$ and as the supremum of this quantity over all A is the total variation distance, this means that we would be able to bound it and upper bound a certain mixing time by t. We should think of a good set for A as a large set from which the chain "mixes in A" (by which we mean that in ends in A at that time with probability which is not much smaller than $\pi(A)$) in time which has the order of the relaxation time. We will now do this formally. We define the good sets first.

Definition 9.9 (Good set for A)**.** Let P be irreducible, reversible chain on S with invariant distribution π. Let $A \subset S$, $s, m > 0$. Denote $\sigma_s = e^{-s/t_{\text{rel}}}\sqrt{\text{Var}_\pi(\mathbb{1}_A)} = e^{-s/t_{\text{rel}}}\sqrt{\pi(A)(1-\pi(A))}$. We define a *good set* for A for time s with m standard deviations as

$$G_s(A,m) = \left\{y \in S : \left|P^t(y,A) - \pi(A)\right| < m\sigma_s \text{ for all } t \geq s\right\}.$$

//

Lecture 18

We will now state the main proposition from this subsection which is key for establishing the wanted hit-mix bound.

Proposition 9.10 (Good sets are large)**.** *Consider the setup from the Definition 9.9. We have for all $A \subset S, s \geq 0$ and $m > 0$ that*

$$\pi(G_s(A,m)) \geq 1 - 8m^{-2}.$$

Use the Starr's maximal inequality and generalised Poincaré to upper bound $\|f_s^*\|$ and $\|(Pf_s)^*\|$ by σ_s^2 where $f_s(x) = P^s(x,A) - \pi(A)$. Notice that $G_s(A,m)$ contains x such that $f_s(x) \leq \sigma_s m, (Pf_s)^* \leq \sigma_s m$ and so it is enough to upper bound $\pi((Pf_s)^* \geq m\sigma_s) + \pi((f_s)^* \geq m\sigma_s)$ for which one uses the previous bound and Markov's inequality (applied to squared expressions).

Proof. Let $f_s(x) = P^s(x,A) - \pi(A) = P^s(\mathbb{1}_A - \pi(A))(x)$. In the notation from Theorem 9.8

$$f_s^*(x) = \sup_{t\geq 0}|P^{2t}f_s(x)| = \sup_{t\geq 0}|P^{2t+s}(x,A) - \pi(A)|,$$

and

$$(Pf_s)^*(x) = \sup_{t\geq 0}|P^{2t+1+s}(x,A) - \pi(A)|.$$

Hence $\{x \in S : f_s^*(x) < m\sigma_s, (Pf_s)^*(x) < m\sigma_s\} = G_s(A,m)$ so

$$1 - \pi(G_s(A,m)) \leq \pi(\{f_s^*(x) \geq m\sigma_s\}) + \pi(\{(Pf_s)^*(x) \geq m\sigma_s\}) \tag{9.3}$$

We have, as $\pi P^s = \pi$ that $\mathbb{E}_\pi[f_s] = \mathbb{E}_\pi[f_0] = \mathbb{E}_\pi[\mathbb{1}_A - \pi(A)] = 0$. Now from the generalised Poincaré inequality (Lemma 9.7) we have

$$\|Pf_s\|_2^2 \leq \|f_s\|_2^2 = \text{Var}_\pi(P^s f_0) \leq e^{-2s/t_{\text{rel}}}\text{Var}_\pi(f_0) = e^{-2s/t_{\text{rel}}}\pi(A)(1-\pi(A)) = \sigma_s^2.$$

Hence, Markov's inequality and Starr's maximal inequality (Theorem 9.8), with $p = 2$ give

$$\begin{aligned}\pi(\{x : f_s^*(x) \geq m\sigma_s\}) = \pi(\{x : (f_s^*(x))^2 \geq m^2\sigma_s^2\}) &\leq \frac{\mathbb{E}_\pi\left[(f_s^*)^2\right]}{m^2\sigma_s^2} \\ &= \frac{\|f_s^*\|_2^2}{m^2\sigma_s^2} \leq \frac{(2\|f_s\|_2)^2}{m^2\sigma_s^2} \leq \frac{4}{m^2},\end{aligned}$$

and similarly $\pi(\{x : (Pf_s)^*(x) \geq m\sigma_s\}) \leq 4/m^2$. Plugging this into (9.3) completes the proof. □

We now prove an upper bound on the mixing time in terms of some hit time and relaxation time.

Theorem 9.11 (Upper bound on mixing times by the hit times, reversible chains)**.** *For reversible irreducible transition matrix P we have*

$$t_{\text{mix}}((\varepsilon+\delta)\wedge 1) \le \text{hit}_{1-\alpha}(\varepsilon) + \left\lceil \frac{t_{\text{rel}}}{2}\log^+\left(\frac{2(1-\varepsilon)^2}{\alpha\varepsilon\delta}\right)\right\rceil$$

where $\log^+ x = \max\{\log(x), 0\}$.

Show by conditioning on the time and place where the walk hits a good set of any set $A \subset S$ that $\mathbb{P}_x\Big(X_{t+s}\in A \;\Big|\; T_{G_s(A,\sqrt{8/\alpha})}\le t\Big) \ge \pi(A) - \sqrt{\frac{8}{\alpha}}\sigma_s$. For $t = \text{hit}_{1-\alpha}(\varepsilon)$ this gives that $P^{t+s}(x,A) \ge (1-\varepsilon)(\pi(A) - \sqrt{8/\alpha}\sigma_s)$ which gives for $s = \left\lceil \frac{t_{\text{rel}}}{2}\log^+\left(\frac{2(1-\varepsilon)^2}{\alpha\varepsilon\delta}\right)\right\rceil$ that $\pi(A) - P^{t+s}(x,A) \le \varepsilon+\delta$.

Proof. Let $x\in S$ and consider $A\subset S$. Let $G = G_s(A,\sqrt{8/\alpha})$. From Proposition 9.10 we have that $\pi(G)\ge 1-\alpha$. Conditioning on T_G and X_{T_G} and using the Markov property we have

$$\begin{aligned}
\mathbb{P}_x(X_{t+s}\in A \mid T_G\le t) &= \sum_{u=0}^{t}\sum_{y\in G}\mathbb{P}_x(X_{t+s}\in A, T_G = u, X_u = y \mid T_G\le t)\\
&= \sum_{u=0}^{t}\sum_{y\in G}\mathbb{P}_x(X_{t+s}\in A \mid T_G = u, X_u = y, T_G\le t)\,\mathbb{P}_x(T_G = u, X_u = y \mid T_G\le t)\\
&= \sum_{u=0}^{t}\sum_{y\in G}\mathbb{P}_y(X_{t+s-u}\in A)\,\mathbb{P}_x(T_G = u, X_u = y \mid T_G\le t)\\
&\ge \sum_{u=0}^{t}\sum_{y\in G}\left(\pi(A) - \sqrt{\frac{8}{\alpha}}\sigma_s\right)\mathbb{P}_x(T_G = u, X_u = y \mid T_G\le t) = \pi(A) - \sqrt{\frac{8}{\alpha}}\sigma_s,
\end{aligned}$$

where for the inequality we have used the definition of G and that $y\in G$. If we set $t = \text{hit}_{1-\alpha}(\varepsilon)$ then as $\pi(G)\ge 1-\alpha$ we have $\mathbb{P}_x(T_G\le t)\ge 1-\varepsilon$ and so the above gives

$$P^{t+s}(x,A) \ge \mathbb{P}_x(T_G\le t)\,\mathbb{P}_x(X_{t+s}\in A \mid T_G\le t) \ge (1-\varepsilon)\left(\pi(A) - \sqrt{\frac{8}{\alpha}}\sigma_s\right). \tag{9.4}$$

If we now choose $s = \left\lceil \frac{1}{2}t_{\text{rel}}\log^+\left(\frac{2(1-\varepsilon)^2}{\alpha\varepsilon\delta}\right)\right\rceil$ and x such that $d(t+s) = \|P(x,\cdot)-\pi\|_{\text{TV}}$ then if we upper bound $\pi(A) - P^{t+s}(x,A)$ for all A by $\varepsilon+\delta$ we would be done. We do this, using equation (9.4), as follows

$$\begin{aligned}
\pi(A) - P^{t+s}(x,A) &\le \pi(A) - (1-\varepsilon)\left(\pi(A) - \sqrt{\frac{8}{\alpha}}\sigma_s\right)\\
&= \varepsilon\pi(A) + (1-\varepsilon)\sqrt{\frac{8}{\alpha}}\sigma_s\\
&= \varepsilon\pi(A) + (1-\varepsilon)\sqrt{\frac{8}{\alpha}}e^{-s/t_{\text{rel}}}\sqrt{\pi(A)(1-\pi(A))}\\
&\le \varepsilon\left(\pi(A) + 2\sqrt{\frac{\delta}{\varepsilon}}\sqrt{\pi(A)(1-\pi(A))}\right)\\
&\le \varepsilon(1 + (2\sqrt{\delta/\varepsilon})^2/4) = \varepsilon+\delta
\end{aligned}$$

where in the last inequality we have used that for any $c > 0$, $x \in [0,1]$ we have $x + c\sqrt{x(1-x)} \leq 1 + c^2/4$. Indeed, this is true as $1 - x + c^2/4 \geq 2(c/2)\sqrt{1-x} \geq c\sqrt{x(1-x)}$, as $x \in [0,1]$. This completes the proof. □

We conclude this subsection with the proof of Starr's maximal inequality. Notice that it is enough to consider $f \geq 0$ and consider $R_t = \mathbb{E}[f(X_{2t}) \mid X_t] = \mathbb{E}[f(X_0) \mid X_t, X_{t+1}, \ldots]$ and show that R_{T-t} is a martingale with respect to its natural filtration. Apply Doob's $\mathcal{L}^p$ inequality to it and take the limit as $T \to \infty$.

Proof of Theorem 9.8 (Starr's maximal inequality).* [8] Let $p \in (1, \infty)$ and $f : S \to \mathbb{R}$. Let $p^* = \frac{p}{p-1}$ be the conjugate exponent of p. We first argue that it suffices to prove the theorem for $f \geq 0$. For general f if we denote $h := |f|$, then $|f^*| \leq h^*$ and so $\|f^*\|_p \leq \|h^*\|_p \leq p^*\|h\|_p = p^*\|f\|_p$ [this is as using triangle inequality $|(P^{2t}f)(x)| = |\sum_{y \in S} P^{2t}(x,y)f(y)| \leq \sum_{y \in S} P^{2t}(x,y)|f(y)| = P^{2t}h(x)$].

We now assume $f \geq 0$ and consider a Markov chain $(X_t)_{t \geq 0}$ with transition P, with $X_0 \sim \pi$. The tower property of conditional expectation (see Advanced Probability notes, Proposition 1.20)

$$P^{2t}f(X_0) = \mathbb{E}[f(X_{2t}) \mid X_0] = \mathbb{E}[\mathbb{E}[f(X_{2t}) \mid X_t] \mid X_0]\,. \tag{9.5}$$

We set $R_t = \mathbb{E}[f(X_{2t}) \mid X_t]$ and notice that since $X_0 \sim \pi$, by reversibility $(X_t, X_{t+1}, \ldots, X_{2t})$ has the same law as $(X_t, X_{t-1}, \ldots, X_0)$ so

$$R_t = \mathbb{E}[f(X_{2t}) \mid X_t] = \mathbb{E}[f(X_0) \mid X_t] = \mathbb{E}[f(X_0) \mid X_t, X_{t+1}, X_{t+2}, \ldots]$$

where the final equality above follows from the Markov property. Fix $T > 0$ and notice that the process $(R_{T-t})_{t=0}^T$ is a martingale with respect to its natural filtration $\mathcal{F}_t = \sigma(X_T, X_{T-1}, \ldots, X_{T-t})$ (see Advanced Probability notes, Definition 2.1). Indeed, using the tower property, we have for $t > u$

$$\begin{aligned}\mathbb{E}[R_{T-t} \mid \mathcal{F}_u] = \mathbb{E}[R_{T-t} \mid X_T, X_{T-1}, \ldots, X_{T-u}] &= \mathbb{E}[\mathbb{E}[f(X_0) \mid X_{T-t}, X_{T-t+1}, \ldots] \mid X_T, \ldots, X_{T-u}] \\ &= \mathbb{E}[f(X_0) \mid X_T, X_{T-1}, \ldots, X_{T-u}] = R_{T-u}.\end{aligned}$$

We note that R_{T-t} is clearly integrable, as given that S is finite it has to be bounded. Doob's $\mathcal{L}^p$ maximal inequality (see Advanced Probability notes, Theorem 2.18) we have

$$\|\max_{t \leq T} R_t\|_p \leq q\|R_0\|_p = p^*\|f(X_0)\|_p.$$

Denote $h_T = \max_{t \leq T} P^{2t}f$ and notice that by (9.5)

$$h_T(X_0) = \max_{t \leq T} \mathbb{E}[R_t \mid X_0] \leq \mathbb{E}\left[\max_{t \leq T} R_t \;\middle|\; X_0\right].$$

By conditional Jensen's inequality (see Advanced Probability notes, Proposition 1.19 (4)) taking $\mathcal{L}^p$ norms above gives

$$\|h_T(X_0)\|_p \leq \left\|\mathbb{E}\left[\max_{t \leq T} R_t \;\middle|\; X_0\right]\right\|_p \leq \|\max_{t \leq T} R_t\|_p \leq p^*\|f(X_0)\|_p.$$

The proof now follows by applying the monotone convergence theorem. □

[8] non-examinable

9.3 Hit cutoff

We recall that we have seen in Proposition 4.16 that we can have no cutoff without the product condition (i.e. if we do not have that $t_{\text{rel}} \ll t_{\text{mix}}$), so every chain which exhibits cutoff also satisfies the product condition. In the previous two subsections, we have found a way of comparing hit times and mixing times up to a difference which is some multiple of the relaxation time. This suggests that if there is a cutoff for mixing times there should be one for hit times. We now formally show that this is indeed true. We start by formally defining the hit cutoff.

Definition 9.12 (Hit cutoff)**.** For $\alpha \in (0,1)$ say that a sequence of chains exhibits hit_α *cutoff* if for all $\varepsilon \in (0,1/4)$ we have

$$\text{hit}_\alpha(\varepsilon) - \text{hit}_\alpha(1-\varepsilon) \ll \text{hit}_\alpha(1/4).$$

If the chain exhibits hit_α cutoff for all $\alpha \in (0,1)$ we say that the chain exhibits *hit cutoff*. //

We note that when we talk about the cutoff and say chain, we really mean a sequence of chains and that by $\text{hit}_\alpha(\varepsilon)$ we mean $\text{hit}_\alpha^n(\varepsilon)$ which corresponds to the nth chain, but we drop n from the notation.

We first show that hit_α cutoff and hit cutoff are equivalent for reversible chains which satisfy the product condition.

Proposition 9.13 (hit_α cutoff gives hit cutoff, reversible chain)**.** *For a sequence of reversible Markov chains which satisfy the product condition ($t_{\text{rel}} \ll t_{\text{mix}}$) we have that the following are equivalent:*

1. *There exists $\alpha \in (0,1)$ for which the sequence of chains exhibits hit_α cutoff*
2. *The sequence exhibits a hit cutoff (i.e. hit_α cutoff for all $\alpha \in (0,1)$)*

Moreover,

$$\text{hit}_\alpha(1/4) \asymp t_{\text{mix}}$$

and if there is a hit cutoff we also have that

$$\lim_{n\to\infty} \frac{\text{hit}_\alpha^n(1/4)}{\text{hit}_{1/2}^n(1/4)} = 1$$

for any $\alpha \in (0,1)$.

This follows by carefully combining the bounds we have on t_{mix} in terms of hit_α (Theorem 9.11 and Claim 9.2) as well as using submultiplicativity for both hit times and mixing times and Proposition 9.6.

Proof. We assume that the product condition holds. We first prove that $\text{hit}_\alpha(1/4) \asymp t_{\text{mix}}$. Fix $\alpha \in (0,1)$. Note that $(1-3\alpha/4)^{4\alpha^{-1}-1} \le 4e^{-3} \le 1/4$ so by the sub-multiplicativity of hit times (Proposition 9.3), Claim 9.2 with $\varepsilon = \alpha/4$ and $\delta = 3\alpha/4$ and Proposition 1.15 we have

$$\text{hit}_\alpha(1/4) \le 4\alpha^{-1}\text{hit}_\alpha(1-3\alpha/4) \le 4\alpha^{-1}t_{\text{mix}}(\alpha/4) \le 4\alpha^{-1}(2+\lceil\log_2(1/\alpha)\rceil)t_{\text{mix}}.$$

On the other hand, Theorem 9.11 for $(1-\alpha, 1/8, 1/8)$ in place of $(\alpha,\varepsilon,\delta)$ and sub-multiplicativity of hit times (Proposition 9.3) gives

$$\frac{t_{\text{mix}}}{2} - \left\lceil \frac{t_{\text{rel}}}{4}\log(2\cdot 49/(1-\alpha))\right\rceil \le \frac{\text{hit}_\alpha(1/8)}{2} \le \frac{\text{hit}_\alpha(1/16)}{2} \le \text{hit}_\alpha(1/4)$$

so we have $\mathrm{hit}_\alpha(1/4) \asymp t_{\mathrm{mix}}$. We now show that 1. implies 2. as the other implication is clearly true. We fix $0 < \alpha < \beta < 1$ and show that hit_α cutoff occurs iff hit_β cutoff occurs. Fix $\varepsilon \in (0, 1/8)$ and let $s = s(\alpha, \beta, \varepsilon) = \left\lceil t_{\mathrm{rel}} \alpha^{-1} \log\left(\frac{1-\alpha}{(1-\beta)\varepsilon}\right)\right\rceil$. By the Proposition 9.6 we have Lecture 19

$$\mathrm{hit}_\alpha(1-\varepsilon) \le \mathrm{hit}_\beta(1-2\varepsilon) + s \quad \text{and} \quad \mathrm{hit}_\alpha(2\varepsilon) \le \mathrm{hit}_\beta(\varepsilon) + s.$$

Moreover, as hit_α is decreasing in α we have

$$\mathrm{hit}_\beta(2\varepsilon) \le \mathrm{hit}_\alpha(2\varepsilon) \le \mathrm{hit}_\alpha(\varepsilon) \quad \text{and} \quad \mathrm{hit}_\beta(1-\varepsilon) \le \mathrm{hit}_\beta(1-2\varepsilon) \le \mathrm{hit}_\alpha(1-2\varepsilon).$$

Combining these inequalities gives

$$\begin{aligned}\mathrm{hit}_\beta(2\varepsilon) - \mathrm{hit}_\beta(1-2\varepsilon) &\le \mathrm{hit}_\alpha(\varepsilon) - \mathrm{hit}_\alpha(1-\varepsilon) + s,\\ \mathrm{hit}_\alpha(2\varepsilon) - \mathrm{hit}_\alpha(1-2\varepsilon) &\le \mathrm{hit}_\beta(\varepsilon) - \mathrm{hit}_\beta(1-\varepsilon) + s.\end{aligned}$$

Assume that cutoff holds for hit_β, we know that $\mathrm{hit}_\beta(1/4) \asymp \mathrm{hit}_\alpha(1/4) \asymp t_{\mathrm{mix}}$ and that the product condition holds, so $s \ll t_{\mathrm{mix}}$, so in the second inequality above the right-hand side is $o(\mathrm{hit}_\alpha(1/4))$ so $\mathrm{hit}_\alpha(2\varepsilon) - \mathrm{hit}_\alpha(1-2\varepsilon) \ll \mathrm{hit}_\alpha(1/4)$ and there is a hit_α cutoff. Analogously if there is hit_α cutoff we also have hit_β cutoff. Finally, if $\alpha < 1/2$ we have from Proposition 9.6 that

$$\mathrm{hit}_\alpha(1/4+\varepsilon) - s \le \mathrm{hit}_{\frac{1}{2}}(1/4) \le \mathrm{hit}_\alpha(1/4).$$

As we know the product condition holds and there is a hit cutoff then, as $s \asymp t_{\mathrm{rel}} \ll \mathrm{hit}_\alpha(1/4)$ and $\mathrm{hit}_\alpha(1/4+\varepsilon) = \mathrm{hit}_\alpha(1/4)(1-o(1))$[9], the left-hand side of the above is lower bounded by $(1-o(1))\mathrm{hit}_\alpha(1/4)$ and so the final assertion of the proposition holds for $\alpha < 1/2$. Proof for $\alpha > 1/2$ is similar. □

We now state the following result from is proved in Example Sheet 4 and use it to prove that hit cutoff is, under the product condition, equivalent to cutoff for reversible chains.

Proposition 9.14 (t_{mix} and $\mathrm{hit}_{1/2}$ comparison, reversible chain)**.** *For a reversible Markov chain and for all $\varepsilon \in (0, 1/4]$ the following inequalities hold*

$$\mathrm{hit}_{1/2}\left(\frac{3\varepsilon}{2}\right) - \lceil 2t_{\mathrm{rel}} |\log(\varepsilon)| \rceil \le t_{\mathrm{mix}}(\varepsilon) \le \mathrm{hit}_{1/2}\left(\frac{\varepsilon}{2}\right) + \lceil t_{\mathrm{rel}} |\log(\varepsilon/4)| \rceil \tag{9.6}$$

$$\mathrm{hit}_{1/2}\left(1-\frac{\varepsilon}{2}\right) - \lceil 2t_{\mathrm{rel}} |\log(\varepsilon)| \rceil \le t_{\mathrm{mix}}(1-\varepsilon) \le \mathrm{hit}_{1/2}(1-2\varepsilon) + \lceil t_{\mathrm{rel}} \log(8)/2 \rceil \mathbb{1}_{\{\varepsilon > 1/18\}} \tag{9.7}$$

Theorem 9.15 (Hit cutoff and cutoff are equivalent under the product condition, reversible chain)**.** *Consider a reversible chain for which the product condition holds. Then, for any α the chain exhibits* hit_α *cutoff if and only if it exhibits cutoff.*

This follows directly from Propositions 9.14 and 9.13 for $\alpha = \frac{1}{2}$.

Proof. From Proposition 9.13 we already know that in the case of reversible chain for which the product condition holds having hit_α cutoff for any α means that there is a hit cutoff, so it is enough to show that cutoff is equivalent to $\mathrm{hit}_{1/2}$ cutoff. From the same proposition we also know that $\mathrm{hit}_{1/2}(1/4) \asymp t_{\mathrm{mix}}$ and so the proof follows from the inequalities (9.6) and (9.7) (as $t_{\mathrm{rel}} \ll t_{\mathrm{mix}}$ and $t_{\mathrm{rel}} \ll \mathrm{hit}_{1/2}$). □

[9] Note that the hit_α-cutoff says that $\mathrm{hit}_\alpha(\varepsilon) - \mathrm{hit}_\alpha(1-\varepsilon) \ll \mathrm{hit}_\alpha\left(\frac{1}{4}\right)$ for all $\varepsilon \le 1/4$ but as $\mathrm{hit}_\alpha(\delta)$ is decreasing in δ (this follows directly from the definition) this means that $\mathrm{hit}_\alpha(\delta) = \mathrm{hit}_\alpha\left(\frac{1}{4}\right)(1+o(1))$ for all $\delta \in (0,1)$. Indeed, this is as if we pick $\varepsilon \le 1/4$ such that $\delta \in (\varepsilon, 1-\varepsilon)$ then we know that $\mathrm{hit}_\alpha(\delta) \in [\mathrm{hit}_\alpha(1-\varepsilon), \mathrm{hit}_\alpha(\varepsilon)]$ and $\mathrm{hit}_\alpha\left(\frac{1}{4}\right) \in [\mathrm{hit}_\alpha(1-\varepsilon), \mathrm{hit}_\alpha(\varepsilon)]$, so the conclusion follows as the length of this interval is $\ll \mathrm{hit}_\alpha\left(\frac{1}{4}\right)$.

9.4 Trees

We will now use the hit-mix results to establish that the cutoff holds for Markov chains on trees under the product condition. We have seen that there can be no cutoff if the product condition doesn't hold. However, we have not shown any general result saying that cutoff always occurs when the product condition holds. In fact, this general result doesn't hold. However, in the case of Markov chains on trees, this ends up being the case. We will show it using the hit-mix technique we have developed.

Definition 9.16 (Notation for Markov chains on trees)**.** Recall from Example 0.13 that a Markov chain transition matrix P on the tree $T = (V, E)$ is reversible with respect to its invariant measure π. A vertex v is called a *central vertex* if each connected component of $T \setminus \{v\}$ has π measure at most $1/2$. We fix a central vertex of o of T (such vertex always exists) and call it a *root*. We write $x \prec y$ if x is on the path from o to y. For any x we also label $\ell(x) = (x_0 = x, x_1, \ldots, x_k = o)$ the path from x to o. We call x_1 a *parent* of vertex x, and we label it $p_x = x_1$. We also define a *subtree* from u to be $\mathcal{T}_u = \{v \in T : u \in \ell(v)\}$[10]. Finally, we define the following ε *hitting time of* o as

$$\tau_o(\varepsilon) = \min\{t : \mathbb{P}_x(T_o > t) \leq \varepsilon, \forall x \in S\}.$$

//

The following claim is shown in Example Sheet 4.

Claim 9.17 (Number of central vertices)**.** *Every tree, with a distribution π on it, has at least one and at most two central vertices.*

Lemma 9.18 (Comparing τ_o and $\mathrm{hit}_{1/2}$)**.** *Denote $s_\delta = \lceil 4t_{\mathrm{rel}} \log(4\delta/9) \rceil$. Then for $0 < \delta < \varepsilon < 1$*

$$\tau_o(\varepsilon) \leq \mathrm{hit}_{1/2}(\varepsilon) \leq \tau_o(\varepsilon - \delta) + s_\delta.$$

If we look at any $x \neq o$ to hit a set $\{o\}$ union connected component of $T \setminus \{o\}$ not containing x, which has $\pi \geq 1/2$, we need to first hit $\{o\}$. This gives $\tau_o \leq \mathrm{hit}_{1/2}$. For the other bound it is enough to show that if $\pi(A) \geq 1/2$, then $\mathbb{P}_o(T_A \geq s_\delta) \leq \delta$. We split connected components of $T \setminus \{o\}$ into two parts, T_1 and T_2 both having $\pi \leq 2/3$ and if WLOG $A_1 = T_1 \cap A$, has $\pi \geq 1/4$ we have $\mathbb{P}_o(T_A \geq s_\delta) \leq T_{A_1} \geq s_\delta o \leq \mathbb{P}_{\pi_{T_2 \cup \{o\}}}(T_{A_1} > s_\delta)$ which we bound using Lemma 9.5.

Proof. Consider a central vertex o and let $x \in V, x \neq o$. If we label the union of {o} with all connected components of $T \setminus \{o\}$ which do not contain x by A (i.e. A is T without the component of $T \setminus \{o\}$ which contains x), then, as o is the central vertex, the component of $T \setminus \{o\}$ containing x has π measure at most $1/2$ and so $\pi(A) \geq 1/2$. We notice that to reach A starting from x the chain needs to reach o first. This implies that

$$\tau_o(\varepsilon) \leq \mathrm{hit}_{1/2}(\varepsilon).$$

For the other inequality, fix $x \in V$ and $A \subset V$ with $\pi(A) \geq 1/2$. Using the union bound and Markov property we have that

$$\mathbb{P}_x(T_A > \tau_o(\varepsilon - \delta) + s_\delta) \leq \mathbb{P}_x(T_o > \tau_o(\varepsilon - \delta)) + \mathbb{P}_o(T_A > s_\delta) \leq \varepsilon - \delta + \mathbb{P}_o(T_A > s_\delta).$$

Therefore, it is enough to bound $\mathbb{P}_o(T_A > s_\delta) < \delta$. This is trivially true if $o \in A$ so assume $o \notin A$. We note that we can partition $T \setminus \{o\}$ into two union of connected components T_1 and T_2 such that

[10] this is the tree containing u and all descendants from u i.e. vertices for which the unique path to o goes through u.

$\pi(T_1) \leq 2/3$ and $\pi(T_2) \leq 2/3$. (This is clearly true if there are only two components of $T \setminus \{o\}$. If there are at least three components the third largest one has to have π measure at most $1/3$, so if we start by any partition where the largest component is in T_1 and the second largest is in T_2, then if one of these is bigger than $2/3$ we can move all other components from larger of T_1 and T_2 to the smaller one, one by one, until the first time the price of the larger one becomes $\leq 2/3$. We claim that both components have a size at most $2/3$ at this point. Indeed, after the last move the size of the smaller component (before this move) increased by at most $1/3$ and before the move it was at most $1/3$ as the other component had a size of at least $2/3$.) Let $A_i = T_i \cap A$, for $i = 1, 2$ and assume WLOG that $\pi(A_1) \geq 1/4$. Now if the chain starts from $T_2 \cup \{o\}$ it must hit o before hitting A_1 so we have

$$\mathbb{P}_o(T_A > s_\delta) \leq \mathbb{P}_o(T_{A_1} > s_\delta) \leq \mathbb{P}_{\pi_{T_2 \cup \{o\}}}(T_{A_1} > s_\delta) \leq \pi(T_2 \cup \{o\})^{-1} \mathbb{P}_\pi(T_{A_1} > s_\delta)$$

(where we have used that $T_A > s_\delta$ implies that $T_{A_1} > s_\delta$ as if we have not hit A we couldn't have hit A_1). As $\pi(A_1) \geq 1/4$ and $\pi(T_2 \cup \{o\})) \geq 1/3$ we are done by upper bounding above using Lemma 9.5. □

As mentioned, the main result of this subsection will be the following

Theorem 9.19 (Cutoff for Markov chains on trees)**.** *For a Markov chain on a tree $T = (V, E)$ with $|V| \geq 3$, for all $\varepsilon \leq 1/4$ we have*

$$t_{\text{mix}}(\varepsilon) - t_{\text{mix}}(1 - \varepsilon) \lesssim \sqrt{\varepsilon^{-1} t_{\text{mix}} t_{\text{rel}}}.$$

This means that a sequence of Markov chains on trees for which the product condition holds exhibits cutoff with a cutoff window of order $\sqrt{t_{\text{mix}} t_{\text{rel}}}$.

To prove the above we can see from Lemma 9.18 (using that $t_{\text{rel}} \lesssim t_{\text{mix}}$ always) that it is enough to show that $\tau_o(\varepsilon) - \tau_o(1 - \varepsilon) \lesssim \sqrt{\varepsilon^{-1} t_{\text{mix}} t_{\text{rel}}}$. So we can now work with hitting times of o and try to show that they are concentrated around their expectation. This would mean that the difference of times when starting from x to hit o with at least a small probability and with at least some large probability will be smaller order than the expected hitting time of o, which would be enough to conclude that there is cutoff (under the product condition). Formally, we first start with the following result which we use to bound the variance of hitting times (which will then allow us to show concentration).

Proposition 9.20 (Hitting time and conductance of a set, reversible chain)**.** *Consider a reversible irreducible transition matrix P on S with invariant distribution π and let A be a non-empty proper subset of S. Let ψ_{A^c} be a distribution on A^c defined by $\psi_{A^c}(y) = \mathbb{P}_{\pi_A}(X_1 = y \mid X_1 \in A^c)$. Then for all $t \geq 1$*

$$\frac{\mathbb{P}_{\pi_{A^c}}(T_A = t)}{\Phi(A^c)} = \mathbb{P}_{\psi_{A^c}}(T_A \geq t)\,,$$

where we recall that $\Phi(A)$ is a conductance of a set from Definition 5.18. This implies that

$$\mathbb{E}_{\psi_{A^c}}[T_A] = \frac{1}{\Phi(A^c)} \quad \textit{and} \quad \mathbb{E}_{\psi_{A^c}}\left[T_A^2\right] = \mathbb{E}_{\psi_{A^c}}[T_A]\,(2\mathbb{E}_{\pi_{A^c}}[T_A] - 1) \leq \frac{2\mathbb{E}_{\psi_{A^c}}[T_A]\, t_{\text{rel}}}{\pi(A)}.$$

Final inequality comes from Lemma 9.5 and the two equalities in the second part come from suitably summing over the first part. For the first part $\pi(A^c)\mathbb{P}_{\pi_{A^c}}(T_A = t) = \mathbb{P}_\pi(X_1 \notin A, \dots X_t \notin A) - \mathbb{P}_\pi(X_1 \notin A, \dots X_t \notin A, X_{t+1} \notin A)$ and after shifting the time by -1 in the second term in RHS this is $\pi(A)\Phi(A)\mathbb{P}_{\psi_{A^c}}(T_A \geq t)$.

Proof. We note that the final inequality of the proposition follows from Lemma 9.5 (and this is where we require reversibility). Summing $\frac{\mathbb{P}_{\pi_{A^c}}(T_A=t)}{\Phi(A^c)} = \mathbb{P}_{\psi_{A^c}}(T_A \geq t)$ over t and multiplying it by $2t-1$ and summing over t gives the final two equalities of the proposition [i.e. $\mathbb{E}_{\psi_{A^c}}[T_A] = \sum_{t\geq 1}\mathbb{P}_{\psi_{A^c}}(T_A \geq t) = \sum_{t\geq 1}\frac{\mathbb{P}_{\pi_{A^c}}(T_A=t)}{\Phi(A^c)} = \frac{1}{\Phi(A^c)}$ and $\mathbb{E}_{\psi_{A^c}}\left[T_A^2\right] = \sum_{t\geq 1}(\sum_{s=1}^{t}(2s-1))\mathbb{P}_{\psi_{A^c}}(T_A = t) = \sum_{s\geq 1}(2s-1)\mathbb{P}_{\psi_{A^c}}(T_A \geq s) = \sum_{s\geq 1}(2s-1)\frac{\mathbb{P}_{\pi_{A^c}}(T_A=s)}{\Phi(A^c)} = \frac{2\mathbb{E}_{\pi_{A^c}}[T_A]-1}{\Phi(A^c)}$ as $\sum_{s=1}^{t}(2s-1) = t^2$], so we only need to prove the first part of the proposition.

Notice that $\{T_A = t\} = \{X_0 \notin A, \dots X_{t-1} \notin A, X_t \in A\}$ and $\{T_A^+ = t+1\} = \{X_1 \notin A, \dots X_t \notin A, X_{t+1} \in A\}$ so stationarity gives $\mathbb{P}_\pi(T_A = t) = \mathbb{P}_\pi\big(T_A^+ = t+1\big)$ and so for $t \geq 1$

$$\begin{aligned}
\pi(A^c)\mathbb{P}_{\pi_{A^c}}(T_A = t) &= \mathbb{P}_\pi(T_A = t) = \mathbb{P}_\pi\big(T_A^+ = t+1\big) = \mathbb{P}_\pi(X_1 \notin A, \dots X_t \notin A, X_{t+1} \in A)\\
&= \mathbb{P}_\pi(X_1 \notin A, \dots X_t \notin A) - \mathbb{P}_\pi(X_1 \notin A, \dots X_t \notin A, X_{t+1} \notin A)\\
&= \mathbb{P}_\pi(X_1 \notin A, \dots X_t \notin A) - \mathbb{P}_\pi(X_0 \notin A, \dots X_t \notin A)\\
&= \mathbb{P}_\pi(X_0 \in A, X_1 \notin A, \dots X_t \notin A) = \pi(A)\Phi(A)\mathbb{P}_{\psi_{A^c}}(X_0 \notin A, \dots X_{t-1} \notin A)\\
&= \pi(A)\Phi(A)\mathbb{P}_{\psi_{A^c}}(T_A \geq t)
\end{aligned}$$

which completes the proof using the fact that $\pi(A)\Phi(A) = \pi(A^c)\Phi(A^c)$. □

As a consequence, noticing that for $u \neq o$ $\psi_{\mathcal{T}_u}$ is exactly u (recall that $\mathcal{T}_u$ is set of all descendants of u), as starting from $T \setminus \mathcal{T}_u$ we can only jump to $\mathcal{T}_u$ through u, gives the following lemma (which follows directly from the above proposition).

Lemma 9.21 (Expected hitting time of the parent)**.** *For $u \neq 0$ (recall that p_u is a parent of u) we have*

$$\mathbb{E}_u[T_{p_u}] = \frac{\pi(\mathcal{T}_u)}{\pi(u)P(u,p_u)} \quad \textit{and} \quad \mathbb{E}_u\left[T_{p_u}^2\right] = 2\mathbb{E}_u[T_{p_u}]\,\mathbb{E}_{\pi_{\mathcal{T}_u}}[T_{p_u}] - \mathbb{E}_u[T_{p_u}] \leq 4\mathbb{E}_u[T_{p_u}]\,t_{\text{rel}}.$$

This gives the following corollary. Lecture 20

Corollary 9.22 (Concentration of hitting time of an ancestor)**.** *Let $x, y \in V$ such that $y \prec x$ and $c > 0$. Let $\sigma_{x,y} = \sqrt{4\mathbb{E}_x[T_y]\,t_{\text{rel}}}$. Then*

$$\text{Var}_x\,(T_y) \leq \sigma_{x,y}^2,$$

$$\mathbb{P}_x(T_y \geq \mathbb{E}_x[T_y] + c\sigma_{x,y}) \leq \frac{1}{1+c^2} \quad \textit{and} \quad \mathbb{P}_x(T_y \leq \mathbb{E}_x[T_y] - c\sigma_{x,y}) \leq \frac{1}{1+c^2}.$$

Bound the variance by the sum of variances of times to hit a parent from each vertex along the path x to y and bound these by their squared expectations. For the next bound apply Markov's to $\mathbb{P}_x\Big(\big(T_y - \mathbb{E}_x[T_y] + \frac{\sigma_{x,y}}{c}\big)^2 \geq \big(c\sigma_{x,y} + \frac{\sigma_{x,y}}{c}\big)^2\Big)$

Proof. Let $v_0 = x, v_1, \dots, v_k = y$ be a path from x to y. Using Lemma 9.21 we have

$$\text{Var}_x\,(T_y) = \sum_{i=1}^{k}\text{Var}_{v_{i-1}}\,(T_{v_i}) \leq \sum_{i=1}^{k}\mathbb{E}_{v_{i-1}}\left[T_{v_i}^2\right] \leq 4t_{\text{rel}}\sum_{i=1}^{k}\mathbb{E}_{v_{i-1}}[T_{v_i}] = \sigma_{x,y}^2.$$

The second part of the proposition now follows by suitably applying Markov's inequality as follows

$$\mathbb{P}_x(T_y - \mathbb{E}_x[T_y] \geq c\sigma_{x,y}) \leq \mathbb{P}_x\left(\left(T_y - \mathbb{E}_x[T_y] + \frac{\sigma_{x,y}}{c}\right)^2 \geq \left(c\sigma_{x,y} + \frac{\sigma_{x,y}}{c}\right)^2\right)$$
$$\leq \mathbb{E}_x\left[\frac{\left(T_y - \mathbb{E}_x[T_y] + \frac{\sigma_{x,y}}{c}\right)^2}{\left(c\sigma_{x,y} + \frac{\sigma_{x,y}}{c}\right)^2}\right] = \frac{\mathrm{Var}_x(T_y) + \frac{\sigma_{x,y}^2}{c^2}}{\frac{(c^2+1)^2}{c^2}\sigma_{x,y}^2} \leq \frac{\sigma_{x,y}^2 + \frac{\sigma_{x,y}^2}{c^2}}{\frac{(c^2+1)^2}{c^2}\sigma_{x,y}^2} = \frac{1}{1+c^2}.$$

The other inequality follows in the same way. □

We now prove the main theorem of this subsection. Corollary 9.22 gives the concentration of T_o upon showing that $\mathbb{E}_x[T_o] \leq 4t_{\mathrm{mix}}$ and this bounds $\tau_o(\varepsilon) - \tau_o(1-\varepsilon)$ and Proposition 9.14 and Lemma 9.18 give the proof. To get $\mathbb{E}_x[T_o] \leq 4t_{\mathrm{mix}}$ we see that the set of A defined to be T without the connected component of x in $T \setminus \{o\}$ has $\pi \geq 1/2$ so the probability that the chain is in it at time t_{mix} is at least $1/4$ regardless of the starting point. Therefore we can run repeated trials of time length t_{mix} until we find the chain in A and notice that the expected number of trials until success is at most 4 and that the chain had to hit o before it reached A.

Proof of Theorem 9.19. For a chain on the tree T we first show that $\mathbb{E}_x[T_o] \leq 4t_{\mathrm{mix}}$. For vertex $x \neq o$ let C_x be the component of $T \setminus \{o\}$ containing x and let $A = T \setminus C_x$. Then let $\tau_A = \inf\{k \in \mathbb{N} : X_{kt_{\mathrm{mix}}} \in A\}$. Then $T_o \leq \tau_A t_{\mathrm{mix}}$, as in order to get to A from x the chain needs to pass through o. By definition of t_{mix} for any starting vertex $y \in V$ we have that $|P^{t_{\mathrm{mix}}}(y, A) - \pi(A)| \leq 1/4$ and so as $\pi(A) \geq 1/2$ this means that $P^{t_{\mathrm{mix}}}(y, A) \geq 1/4$. This and the Markov property imply that τ_A is stochastically dominated by a Geometric with parameter $1/4$ so $\mathbb{E}_x[T_o] \leq \mathbb{E}_x[\tau_A]\, t_{\mathrm{mix}} \leq 4t_{\mathrm{mix}}$. We now fix $\varepsilon \in (0, 1/4]$ and label

$$\tau = \max_{x \in V} \mathbb{E}_x[T_o] \quad \text{and} \quad \kappa_\varepsilon = \sqrt{4\varepsilon^{-1}\tau t_{\mathrm{rel}}}.$$

Notice that $\mathbb{E}_x[T_o] \leq 4t_{\mathrm{mix}}$ implies that $\tau \leq 4t_{\mathrm{mix}}$. Notice also that from the Corollary 9.22 we have that for all $x \in V$, $0 \leq \mathrm{Var}_x(T_o) \leq \sqrt{4\tau t_{\mathrm{rel}}}$. Let x be the vertex for which $\mathbb{E}_x[T_o]$ is maximal, then from the last part of the same Corollary we also get that for $c = \sqrt{\varepsilon^{-1} - 1}$ and all $y \in V$, as $\mathbb{E}_y[T_o] \leq \mathbb{E}_x[T_o] = \tau$ that

$$\mathbb{P}_y\big(T_o \geq \tau + c\sqrt{4\tau t_{\mathrm{rel}}}\big) \leq \frac{1}{1+c^2} = \varepsilon \quad \text{and} \quad \mathbb{P}_x\big(T_o \leq \tau - c\sqrt{4\tau t_{\mathrm{rel}}}\big) \leq \frac{1}{1+c^2} = \varepsilon.$$

From the first inequality, as $\kappa_\varepsilon \geq c\sqrt{4\tau t_{\mathrm{rel}}}$ it follows that

$$\tau_o(\varepsilon) \leq \tau + \kappa_\varepsilon$$

and from the second inequality, we have that

$$\max_{y \in V} \mathbb{P}_y\big(T_o > \tau - c\sqrt{4\tau t_{\mathrm{rel}}}\big) \geq \mathbb{P}_x\big(T_o > \tau - c\sqrt{4\tau t_{\mathrm{rel}}}\big) \geq 1 - \varepsilon$$

which implies that

$$\tau_o(1-\varepsilon) \geq \tau - \kappa_\varepsilon.$$

Now that we have established an analogue of cutoff property of τ_o the proof follows by applying hit-mix as follows.

From Proposition 9.14 (by adding +2 to the upper bound and dropping the two integer parts)

$$t_{\mathrm{mix}}(\varepsilon) - t_{\mathrm{mix}}(1-\varepsilon) \leq \mathrm{hit}_{1/2}(\varepsilon/2) - \mathrm{hit}_{1/2}(1-\varepsilon/2) + 3t_{\mathrm{rel}}|\log(\varepsilon)| + t_{\mathrm{rel}}\log(4) + 2.$$

From Lemma 9.18 taking $(\varepsilon/2, \varepsilon/4)$ for (ε, δ) parameters from the lemma

$$\mathrm{hit}_{1/2}(\varepsilon/2) - \mathrm{hit}_{1/2}(1-\varepsilon/2) \leq \tau_o(\varepsilon/4) - \tau_o(1-\varepsilon/2) + s_{\varepsilon/4}.$$

The above two equations and bound we just established on τ_o now give

$$t_{\mathrm{mix}}(\varepsilon) - t_{\mathrm{mix}}(1-\varepsilon) \leq \kappa_{\varepsilon/4} + \kappa_{\varepsilon/2} + t_{\mathrm{rel}}(7|\log(\varepsilon)| + 4\log(9) - 3\log(4)) + 3.$$

We have already seen that $\kappa_\varepsilon \lesssim 4\sqrt{\varepsilon^{-1} t_{\mathrm{mix}} t_{\mathrm{rel}}}$ and as $t_{\mathrm{rel}} \lesssim t_{\mathrm{mix}}$ (from Theorem 4.12) the result of the lemma follows using the fact that $|\log(\varepsilon)| \leq \frac{2}{\sqrt{\varepsilon}}$. □

10 Electrical networks (reversible chains)

In this section, we work with reversible Markov chains only. We will study an incredibly useful way of thinking about Markov chains, known as electrical networks. This works for weighted random walks on graphs, but all reversible Markov chains can be taught of as weighted random walks. The theory we develop here can be applied to infinite chains and has its use in determining if the chain is recurrent or transient. However, this is beyond the scope of this course and we will focus here on finite chains and see how this theory gives bounds on expected hitting times. These can be used to bound the mixing times and we will also see an application of this theory in bounding mixing times in the case of one particular class of chains known as the random walk on the Lamplighter.

Definition 10.1 (Conductance and resistance)**.** Consider a undirected graph $G = (V, E)$ and a sequence of non-negative numbers $(c(e))_{e \in E}$, which we call *conductances*. For an edge $e = \{x, y\}$ write $c(x, y) = c(e) = c(y, x)$. Call the reciprocal $r(e) = 1/c(e)$ the *resistance* of the edge e. A Markov chain defined by G and the conductances $(c(e))_{e \in E}$ is called the weighted random walk with eights $(c(e))_{e \in E}$ and has transition probabilities

$$P(x, y) = \frac{c(x, y)}{c(x)}$$

where $c(x) = \sum_{x \sim y} c(x, y)$. It is easy to check that for $c_G = \sum_{x \in V} c(x)$ the invariant distribution $\pi(x) = \frac{c(x)}{c_G}$ is in detailed balance with P and so this chain is reversible. Moreover, any reversible chain can be represented as a weighted random walk on a graph by setting $c(x, y) = \pi(x)P(x, y)$. //

Definition 10.2 (Harmonic function)**.** Function $h : V \to \mathbb{R}$ is called *harmonic* for P at vertex $x \in V$ if

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

We call a function harmovic for P on a set A if it is harmonic at all $x \in A$. //

We recall that we denoted a hitting time of a set B by T_B.

Proposition 10.3 (Unique extension of harmonic functions)**.** *Let X be an irreducible Markov chain on S with transition matrix P and let $B \subset S$. Let $f : B \to \mathbb{R}$ be a function defined on B. Then the function defined by*

$$h(x) = \mathbb{E}_x[f(X_{T_B})]$$

is the unique extension $h : S \to \mathbb{R}$ of f (i.e. $h(x) = f(x)$ for $x \in B$) which is harmonic at all vertices in $S \setminus B$.

Check from the definition that the above function is harmonic outside of B. Check that if we have two harmonic functions on B^c which agree on B then their difference g is harmonic and this gives that it is 0 as it is 0 on B [Assume that the maximum is at a and is > 0 and get that all vertices x which communicate with a when edges in B are deleted have $g(x) = g(a)$ and as graph is connected $g > 0$ for some point in B giving contradiction].

Proof. The property $h(x) = f(x)$ for $x \in B$ clearly holds for above defined h. We now show that it is harmonic outside B. Indeed, for $x \notin B$ conditioning on the first step from x and using the Markov property we have

$$h(x) = \sum_{y \in S} P(x,y)\mathbb{E}_x[f(X_{T_B}) \mid X_1 = y] = \sum_{y \in S} P(x,y)\mathbb{E}_y[f(X_{T_B})] = \sum_{y \in S} P(x,y)h(y).$$

The proof of uniqueness follows by assuming that two different functions are satisfying these properties and considering the difference of these and showing it has to be 0. The key is that the required property implies that the two functions have a difference of 0 for vertices in B by showing that the maximum (and in the same way minimum) have to be 0. Here are the details.

Suppose that h' is another function satisfying these properties and let $g = h - h'$. It is easy to check that g also has to be harmonic on $S \setminus B$ (i.e. that the difference of two harmonic functions on a given set is still harmonic on that set) and clearly $g = 0$ on B. We will show that $g \leq 0$. Assume not and let

$$a \in \{x \in S : g(x) = \max_{y \in S} g(y)\}$$

as the maximum is positive we know that $a \notin B$. Consider $x \in S$ with $P(a,x) > 0$. If $g(x) < g(a)$ then we know that as the function is harmonic at a

$$g(a) = \sum_{y \in S} g(y)P(a,y) = g(x)P(a,x) + \sum_{y \in S, y \neq x} P(a,y)g(y) < g(a),$$

which is clearly a contradiction and so $g(y) = g(a)$ for all y neighbours of a. This now gives that $g(x) = g(a)$ for all x which communicates with A when the edges between the vertices in B are deleted. But as the graph is connected this implies that there is an x in B with $g(a) = g(x)$ so the maximum of g is 0. In the same way, we show that the minimum of g has to be 0 and this implies that $g \equiv 0$. □

Definition 10.4 (Flow)**.** A *flow* θ on $G = (V, E)$ is a function defined on directed edges $((x,y))_{x \sim y}$ for which

$$\theta(x,y) = -\theta(y,x).$$

The *divergence* of the flow θ is defined to be

$$\operatorname{div}\theta(x) = \sum_{y \sim x} \theta(x,y).$$

For a vertex $a \in V$ which we call a *source* and $b \in V$ which we call a *sink* we define a *flow* θ *from* a *to* b to be any flow such that

$$\operatorname{div}\theta(x) = 0 \forall x \notin \{a,b\} \quad \text{and} \quad \operatorname{div}\theta(a) \geq 0.$$

We define a *strength* of a flow from a to b to be $\|\theta\| = \operatorname{div}\theta(a)$ and we call flow a *unit flow* if $\|\theta\| = 1$. //

Remark 10.5. *We notice that from the asymmetry of flow, we have*

$$\sum_{x\in V} \operatorname{div}\theta(x) = \sum_{\{x,y\}\in E} (\theta(x,y)+\theta(y,x)) = 0$$

so $\operatorname{div}\theta(b) = -\operatorname{div}\theta(a)$ *for a flow from* a *to* b*. We remark that we also call the condition* $\operatorname{div}\theta(x) = 0$ *the Kirchoff's node low and we think of it as saying that "flow in equals flow out".*

Definition 10.6 (Voltage and current)**.** A *voltage* W is a harmonic function on $V\setminus\{a,b\}$. By Proposition 10.3 a voltage always exists and is unique for any fixed boundary values of $W(a)$ and $W(b)$. Given a voltage W we define a *current flow* I on oriented edges via

$$I(x,y) = \frac{W(x)-W(y)}{r(x,y)} = c(x,y)(W(x)-W(y)).$$

//

It is not hard to check that current is indeed a flow either from a to z or z to a[11] [we can think of flows from z to a as negative flows from a to z] and also that a unit current flow is unique.

Claim 10.7 (Current satisfies Ohm's and cycle law)**.** *Ohm's law says that for a current* I *corresponding to* W *we have* $r(x,y)I(x,y) = W(x)-W(y)$ *and it is satisfied directly from the definition of current flow. It is also not hard to check that the current flow satisfies a cycle law which states that if we have a sequence of directed edges* $(x_1,x_2),(x_2,x_3),\dots(x_k,x_{k+1}=x_1)$ *which form a cycle then*

$$\sum_{i=1}^{k} r(x_i,x_{i+1})I(x_i,x_{i+1}) = 0.$$

Lecture 21

Proposition 10.8 (Flow satisfying cycle law is current flow)**.** *Let* I *be a current and* θ *a flow from* a *to* z *satisfying the cycle law for any cycle. If* $\|\theta\| = \|I\|$*, then* $\theta = I$*.*

Look at $I-\theta$ and as it satisfies cycle law and Kirchoff's node law for all vertices, see that it has to be 0 [if it is positive on one edge it has to be on another one which shares a vertex and eventually this gives a cycle along which the function is positive].

Proof. Consider the function $f = \theta - I$. Then since θ and I are flows with the same strength, it follows that f satisfies Kirchoff's node law at all nodes[12] and the cycle law. Suppose that $\theta \neq I$. Then, without loss of generality, there must exist an edge (x_1,x_2) such that $f((x_1,x_2)) > 0$. Since $\sum_{y\sim x} f(x,y) = 0$, we get that there must exist an edge (x_2,x_3) to which (x_1,x_2) leads such that $f((x_2,x_3)) > 0$. Continuing in this way we get a sequence of oriented edges with a positive value of f. Since the graph is finite, at some point this sequence of edges should revisit a point. This then violates the cycle law. Hence $\theta = I$. □

[11] Just check that the current has divergence 0 at all $x \neq a,z$ which holds as $\sum_{y\sim x} I(x,y) = \sum_{y\sim x} c(x,y)(W(x)-W(y)) = c(x)\sum_{y\sim x} P(x,y)(W(x)-W(y)) = c(x)W(x) - c(x)\sum_{y\sim x} P(x,y)W(y) = c(x)W(x) - c(x)W(x) = 0$ which follows as W is harmonic on $V\setminus\{a,z\}$.

[12] it's satisfied at a and z as they have the same strength and at all other nodes they both have a divergence 0 as we know that currents are also flows from a to z (which could, in general, have a negative strength)

10.1 Effective resistance

Definition 10.9 (Effective resistance and conductance)**.** For a graph $G = (V, E)$ and $a, z \in V$ we define an *effective resistance* as

$$R_{\text{eff}}(a, z) = \frac{W(a) - W(z)}{\|I\|}$$

where W is any voltage any voltage from a to z and I is the current corresponding to it. We define the *effective conductance* as $C_{\text{eff}} = 1/R_{\text{eff}}$. //

Claim 10.10 (Effective resistance is well defined)**.** *The effective resistance is well defined, i.e. it is a property of a graph and does not depend on the choice of voltage.*

If W_0 is a voltage with $W_0(a) = 1$ and $W_0(z) = 0$ then just use that any voltage can be expressed using it (as harmonic functions are unique) as $W(x) = (W(a) - W(z))W_0(x) + W(z)$.

Proof. We need to show that

$$\frac{W(a) - W(z)}{\|I\|}$$

is constant for any voltage and current corresponding to it. If we set W_0 to be the voltage with $W_0(a) = 1$ and $W_0(z) = 0$ then by the uniqueness of harmonic functions, as the below function is easily checked to be harmonic with boundary conditions same as W, we have that for any voltage W and any $x \in V$

$$W(x) = (W(a) - W(z))W_0(x) + W(z).$$

Let I_0 be the current of W_0 and note that by the definition of current flow

$$\|I\| = \sum_{x \sim a} \frac{W(a) - W(x)}{r(a, x)} = \sum_{x \sim a} \frac{(W(a) - W(z))(W_0(a) - W_0(x))}{r(a, x)} = (W(a) - W(z))\|I_0\|.$$

This means that

$$\frac{W(a) - W(z)}{\|I\|} = \frac{1}{\|I_0\|}$$

and completes the proof. □

The reason above is called effective resistance is that if we would replace the whole network by a single edge between a and z and place resistance R_{eff} on it the current which would correspond to some voltage W from a to z in this new network would have the same strength as the current which would flow in the original network. So if we only care about the current strength we can replace the network with one edge having resistance R_{eff}.

We now give the first result which shows the connection between the electrical networks and random walks.

Proposition 10.11 (Effective resistance and escape probability)**.** *Let X be a reversible chain on $G = (V, E)$ and let $(c(e))_{e \in E}, (r(e))_{e \in E}$ be corresponding conductances and resistances. Then for any $x, a, z \in V$ we have*

$$\mathbb{P}_a\big(T_z < T_a^+\big) = \frac{1}{c(a)R_{\text{eff}}(a, z)}.$$

We call the quantity $\mathbb{P}_a(T_z < T_a^+)$ escape probability (from a to z).

As $h(x) = \mathbb{P}_x(T_z < T_a)$ is harmonic then by uniqueness $h(x) = \frac{W(a)-W(x)}{W(a)-W(z)}$ and we write $\mathbb{P}_a(T_z < T_a^+) = \sum_{x\sim a} P(a,x)\mathbb{P}_x(T_z < T_a)$.

Proof. The function $f(x) = \mathbb{P}_x(T_z < T_a)$ is a harmonic function on $V \setminus \{a, z\}$ and $f(a) = 0$ and $f(z) = 1$. So from Proposition 10.3 we get that f has to be equal to the function

$$h(x) = \frac{W(a) - W(x)}{W(a) - W(z)}$$

where W is a voltage, as they are both harmonic with the same boundary values. Therefore, we obtain

$$\mathbb{P}_a\big(T_z < T_a^+\big) = \sum_{x\sim a} P(a,x)\mathbb{P}_x(T_z < T_a) = \sum_{x\sim a} \frac{c(a,x)}{c(a)} \cdot \frac{W(a) - W(x)}{W(a) - W(z)}.$$

By the definition of the current flow, the above sum is equal to

$$\frac{\sum_{x\sim a} I(a,x)}{c(a)(W(a) - W(z))} = \frac{1}{c(a)R_{\text{eff}}(a,z)}$$

and this proves the proposition. □

We now define a Green's function which counts an expected amount of time spent at some fixed state until the stopping time. We will then relate it to the effective resistance.

Definition 10.12 (Green's function)**.** For any stopping time T we define a *Greens function* for the random walk stopped at T as

$$G_T(a,x) = \mathbb{E}_a\left[\sum_{t=0}^{\infty} \mathbb{1}_{\{X_t=x, T>t\}}\right].$$

//

Lemma 10.13 (Green's function and effective resistance)**.** *For a reversible Markov chain on $G = (V, E)$ and all $a, z \in V$ we have*

$$G_{T_z}(a,a) = c(a)R_{\text{eff}}(a,z).$$

Number of returns to a before hitting z is geometric with parameter $\mathbb{P}_a(T_z < T_a^+)$.

Proof. Due to Markov property the number of visits to a by the hitting time of z has a geometric distribution with $\mathbb{P}_a(T_z < T_a^+)$ so the proof follows from Proposition 10.11. □

The reason electrical networks are very useful tools is that we can use some simple rules to simplify the network which would not change the strength of the current flow for a given voltage. We have the following rules.

Proposition 10.14 (Simplifying the network)**.** *The following three rules hold*

1. **Conductances in parallel add:** *For two edges sharing both endpoints (i.e. a multigraph) then we can replace the two edges with a single edge which has conductance equal to the sum of conductances of the two edges and we would have that the same current would flow through and the same voltage difference will be applied.*

2. **Resistances in series add:** *For a vertex $v \in V$ of degree 2 which is connected to v_1 and v_2 we can replace the two edges at v (and remove v) by an edge between v_1 and v_2 which has the resistance which is the sum of resistances of $\{v, v_1\}$ and $\{v, v_2\}$ and this will not change the voltage or current.*

3. **Gluing:** *If two vertices have the same voltage we can glue them to a single vertex while keeping all the edges and this will not change the voltage or current (this possibly creates a multigraph but we can add the conductances in parallel to delete multi-edges).*

Check Kirchoff's and Ohm's laws for natural choice of currents and same voltages as in the initial network.

Proof. It is not hard to check that in all of the above cases Kirchoff's and Ohm's laws still hold for the same voltages and natural choices of current in a new network, as follows. For the conductances in parallel, the new current is the sum of the currents through the two edges we replaced and for the resistances in series the current on (v_1, v_2) is the same as one on (v, v_2) which, as current is a flow so $\operatorname{div} I(v) = 0$, is the same as current on (v_1, v). For the final statement notice that there is no current between the vertices with the same voltage so all the currents and voltages remain the same. □

Example 10.15 (Effective resistance on a tree)**.** *Consider two vertices a and b on a finite connected tree T which has capacity 1 on each edge. Then we can see that $R_{\text{eff}}(a, b)$ is the graph distance of a and b. This is because if x is a vertex which is not on the path a, b, if we look at the unique path from x to a and call y place where it first meets the unique path between a and b then the voltage at y is same as the one at x. Indeed, we know that voltage is a harmonic function so as $\mathbb{E}_x\Big[W(X_{T_{\{a,b\}}})\Big] = \mathbb{E}_y\Big[W(X_{T_{\{a,b\}}})\Big]$ using Proposition 10.3 $W(x) = W(y) = \mathbb{E}_y\Big[W(X_{T_{\{a,b\}}})\Big]$. Therefore there is no current between x and y so we can delete all parts of the graph which are not on a path between a and b (i.e. use gluing and delete self-loops). Finally add the resistances in parallel on a path from a to b to get wanted.*

Definition 10.16 (Energy)**.** Let θ be a flow on a finite connected graph G. We define its *energy* by

$$\mathcal{E}(\theta) = \sum_{e \in E} (\theta(e))^2 r(e),$$

where the sum is taken over all unoriented edges E (and as θ is antisymmetric we do not need to specify the orientation of e). //

The energy should not be confused with the Dirichlet form, we are just using the same letter to denote it but they are not the same (and one involves a flow which is a function on edges while the second one is for the function on vertices and they are multiplying the square of the function on the edge/difference of the function on vertices with different quantities).

The following theorem gives an equivalent definition of effective resistance as the minimal energy over flows of unit strength from a to z.

Theorem 10.17 (Thomson's principle)**.** *For all a and z we have*

$$R_{\text{eff}}(a, z) = \inf\{\mathcal{E}(\theta) : \theta \text{ is a unit flow from } a \text{ to } z\}.$$

Moreover, there is a unique minimiser for the above and it is the unit current flow from a to z.

For unit current flow i with voltage ϕ check that $\mathcal{E}(i) = R_{\text{eff}}(a,z)$ and for any flow j check that $\mathcal{E}(j) > \mathcal{E}(i)$ if $i \neq j$, by writing $\mathcal{E}(j)$ using $k = j - i$ and noticing that $\sum_{e\in E} i(e)k(e)r(e) = 0$.

Proof. Let i be the unit current flow from a to z associated with the voltage ϕ. Using Ohm's law and then that i is a unit flow from a to z so $\operatorname{div} i(x) = 0$ for $x \notin \{a, z\}$ and $\operatorname{div} i(a) = 1 = -\operatorname{div} i(z)$ and i is antisymmetric we have

$$\begin{aligned}\mathcal{E}(i) &= \frac{1}{2}\sum_{\substack{x,y\in V\\ x\sim y}} i(x,y)^2 r(x,y) = \frac{1}{2}\sum_{\substack{x,y\in V\\ x\sim y}} i(x,y)(\phi(x)-\phi(y))\\ &= \sum_{\substack{x,y\in V\\ x\sim y}} i(x,y)\phi(x) = \sum_{x\in V}\phi(x)\operatorname{div} i(x) = \phi(a)-\phi(z) = R_{\text{eff}}(a,z).\end{aligned}$$

We now need to show that for any other unit flow j we have $\mathcal{E}(j) \geq \mathcal{E}(i)$ with equality if and only if $i = j$. We define $k = j - i$ and notice that it is also a flow and has strength 0. We get

$$\begin{aligned}\mathcal{E}(j) &= \sum_{e\in E}(j(e))^2 r(e) = \sum_{e\in E}(i(e)+k(e))^2 r(e) = \sum_{e\in E} i(e)^2 r(e) + \sum_{e\in E} k(e)^2 r(e) + 2\sum_{e\in E} i(e)k(e)r(e)\\ &= \mathcal{E}(i) + \mathcal{E}(k) + 2\sum_{e\in E} i(e)k(e)r(e).\end{aligned}$$

We will now show that $\sum_{e\in E} i(e)k(e)r(e) = 0$. We know that as i is the unit current flow associated with ϕ we have that for $e = (x, y)$

$$i(x,y) = \frac{\phi(x)-\phi(y)}{r(x,y)}.$$

Therefore,

$$\sum_{e\in E} i(e)k(e)r(e) = \frac{1}{2}\sum_{\substack{x,y\in V\\ x\sim y}}(\phi(x)-\phi(y))k(x,y) = \sum_{\substack{x,y\in V\\ x\sim y}}\phi(x)k(x,y) = \sum_{x\in V}\phi(x)\operatorname{div} k(x) = 0$$

where the last equality holds as k is a zero strength flow so $\operatorname{div} k(x) = 0$ for all x. As $\mathcal{E}(k) \geq 0$ this proves that $\mathcal{E}(j) \geq \mathcal{E}(i)$. Moreover, the equality holds only if $\mathcal{E}(k) = 0$ but as all terms in the sum for energy are non-negative that can only happen if they are all 0 but that means that $k \equiv 0$ so $i = j$. □

Lecture 22

Corollary 10.18 (Gluing does not increase effective resistance)**.** *The procedure of gluing from Proposition 10.14 can not increase effective resistance between any vertices we have not glued together.*

Thomson's principle implies this, as all flows in the initial network are flows in the network after gluing.

Proof. The infimum in Thomson's principle is taken over a higher number of flows after gluing, as all unit flows in the previous network are still unit flows as if we keep the multi-edges and loops initially the flows as function on edges will not change at all and divergence at the two vertices we glued would just be a sum of their divergences [we can see that after gluing we possibly have some

other flows which have divergence at the "glued vertex" which is zero (or $\operatorname{div}\theta(a)$ or $\operatorname{div}\theta(z)$ if we glued a or respectively z to some other vertex) but before gluing the divergence at the two vertices we have glued is not such that the function is a flow]. As the unit current flow going through the glued network will not go through the loops we can delete them, as from Thomson's principle we know that the energy is minimised by the unit current flow. □

Theorem 10.19 (Rayleigh's monotonicity principle)**.** *The effective resistance is a monotone increasing function as a function of the component resistances, i.e. if $(r(e))_{e\in E}$ and $(r'(e))_{e\in E}$ are such that $r(e) \le r'(e)$ for all $e \in E$ then*

$$R_{\text{eff}}(a, z; r) \le R_{\text{eff}}(a, z; r').$$

This follows from Thomson's principle.

Proof. Let i and i' be the unit current flows associates to the resistance $r(e)$ and $r'(e)$, respectively. Then by Thomson's principle, we get

$$R_{\text{eff}}(a, z; r) = \sum_{e\in E}(i(e))^2 r(e) \le \sum_{e\in E}(i'(e))^2 r(e)$$

where the inequality follows since the energy is minimised by the unit current flow. From the assumption on the resistance we now have

$$\sum_{e\in E}(i'(e))^2 r(e) \le \sum_{e\in E}(i'(e))^2 r'(e) = R_{\text{eff}}(a, z; r')$$

which completes the proof. □

Corollary 10.20 (Adding an edge increases escape probability)**.** *Let $G = (V, E)$ be a finite connected graph and let $a \in V$. Suppose we add an edge which is not connected to a. Then this would increase the escape probability to z, $\mathbb{P}_a(T_z < T_a^+)$.*

Proposition 10.11 and Rayleigh's monotonicity principle directly imply this as adding an edge decreases its resistance from infinity to a finite value.

Proof. Recall that from Proposition 10.11 $\mathbb{P}_a(T_z < T_a^+) = \frac{1}{c(a)R_{\text{eff}}(a,z)}$. Adding an edge means that we decrease the resistance of the edge from ∞ to a finite number (as initially the capacity of that non-existent edge is 0, and so resistance is ∞). Using Rayleigh's monotonicity principle, this means the effective resistance will decrease. As the edge we added is not connected to a we know that $c(a)$ stays unchanged and the proof follows. □

We now present a technique to obtain lower bounds on effective resistance.

Definition 10.21 (Cutset)**.** We call a set of edges Π an edge-cutset separating a and z if every path from a to z contains at least one edge from Π. //

Proposition 10.22 (Nash-Williams inequality)**.** *If $(\Pi_k)_{k=1}^m$ is a sequence of m disjoint edge-cutsets which separate a and z then*

$$R_{\text{eff}}(a, z) \ge \sum_{k=1}^{m}\left(\sum_{e\in\Pi_k} c(e)\right)^{-1}.$$

Using Cauchy-Schwarz get that $\left(\sum_{e\in\Pi_k} c(e)\right)\cdot\left(\sum_{e\in\Pi_k} r(e)(\theta(e))^2\right) \geq \left(\sum_{e\in\Pi_k} |\theta(e)|\right)^2$ which when we lower bound this by 1 completes the proof using Thomson's principle. Lower bound by 1 holds as θ is of unit strength and Π_k is a cutset (to see this sum the divergence over vertices which can be reached from a when Π_k is removed).

Proof. By Thomson's principle it suffices to prove that for any flow θ from a to z of unit strength we have

$$\sum_{e\in E}(\theta(e))^2 r(e) \geq \sum_{k=1}^{m}\left(\sum_{e\in\Pi_k} c(e)\right)^{-1}.$$

Since the sets Π_k are disjoint we get

$$\sum_{e\in E}(\theta(e))^2 r(e) \geq \sum_{k=1}^{m}\sum_{e\in\Pi_k}(\theta(e))^2 r(e).$$

By the Cauchy-Schwarz inequality, we get

$$\left(\sum_{e\in\Pi_k} c(e)\right)\cdot\left(\sum_{e\in\Pi_k} r(e)(\theta(e))^2\right) \geq \left(\sum_{e\in\Pi_k}\sqrt{c(e)}\sqrt{r(e)}|\theta(e)|\right)^2 = \left(\sum_{e\in\Pi_k}|\theta(e)|\right)^2 \geq 1$$

which completes the proof after rearranging. We note that the last inequality holds since the sets Π_k are cutsets separating a and z and θ has unit strength. Indeed, to see this notice that for any cutset Π and $A = \{v \in V : \text{ there is a path } a \text{ to } v \text{ in } G\setminus\Pi\}\cup\{a\}$ we have

$$1 = \|\theta\| = \sum_{x\in A}\operatorname{div}\theta(x) = \sum_{x\in A, y\in V}\theta((x,y)) = \sum_{x\in A, y\notin A}\theta((x,y)) \leq \sum_{e\in\Pi}|\theta(e)|$$

where the last equality holds as for $y \in A$ we have $\theta((x,y))+\theta((y,x)) = 0$ while the second equality holds as $z \notin A$ by the definition of a cutset. □

Proposition 10.23 (Effective resistance of a grid)**.** *Let $a = (1,1)$ be the lower left-hand corner of $B_n = \mathbb{Z}^2 \cap [1,n]^2$ and let $z = (n,n)$ be the upper right-hand corner of B_n. Suppose each edge of B_n has unit conductance. The effective resistance satisfies*

$$\frac{\log(n-1)}{2} \leq R_{\text{eff}}(a,z) \leq 2\log n.$$

For lower bound take the cutsets of edges which are at a distance k from one corner, and for the upper bound construct the flow from the bottom right corner to the diagonal such that sum of incoming flows to vertices of coordinate sum k is $1/(k-1)$ and then reflect it above the diagonal [to do this use Pólya's urn process which starts from one black and one white ball in an urn and at every time step picks a ball at uniformly at random from the urn and adds a ball of the same colour].

Proof. We first show the lower bound. Let $\Pi_k = \{(v,u) \in B_n : \|v\|_\infty = k \text{ and } \|u\|_\infty = k-1\}$. Then (Π_k) are disjoint edge-cutsets separating a from z. By counting the number of edges we get $|\Pi_k| = 2(k-1)$ and since $c(e) = 1$ for all edges e, we get

$$R_{\text{eff}}(a,z) \geq \sum_{k=2}^{n}\frac{1}{2(k-1)} \geq \frac{\log(n-1)}{2}$$

and this completes the proof. For the upper bound, we will construct a unit flow which we will show has energy bounded by $2\log n$ and so we will be done using Thomson's principle (Theorem 10.17). Consider Pólya's urn process, which is defined as follows. We have two balls in the urn initially, one black, one white and at every step we do the following. We take a uniformly random ball out of the urn, look at its colour, return it to the urn and add one more ball of the same colour. The sequence of ordered pairs listing the numbers of black and white balls forms a Markov chain with state space which is subset $\{1,2,\dots\}^2$ and the non-zero transitions are $P((i,j),(i+1,j)) = \frac{i}{i+j} = 1 - P((i,j),(i,j+1))$.

Run this process on the square grid, starting at vertex $a = (1,1)$, and stop when you reach the main diagonal $x + y = n + 1$. Direct all edges of the square from the bottom left to the top right, and assign each edge e on the bottom left half of the square the flow

$$f(e) = \mathbb{P}(\text{the process went through } e)\,.$$

To complete the construction, assign the upper right half of the square symmetrical flow values.

It is easy to check and prove using induction [see that if the probability to pass through $(i,j-1)$ and $(i-1,j)$ are $\frac{1}{i+j-2}$ then $\frac{1}{i+j-2}P((i-1,j),(i,j)) + \frac{1}{i+j-2}P((i,j-1),(i,j)) = \frac{1}{i+j-1}\frac{i-1+j-1}{i+j} = \frac{1}{i+j-1}$ is the probability to pass through (i,j)] that for any $k \geq 0$, Pólya's urn process starting from $(1,1)$ is equally likely to pass through each of the $k+1$ pairs (i,j) for which $i + j = k + 2$.

Consequently, when (i,j) is a vertex in the square for which $i + j = k + 2$, the sum of the flows on its incoming edges is $\frac{1}{k+1}$. Thus, the energy of the flow f can be bounded as follows [note that energy is sum over undirected edges and squares the flow so we can sum over the direction on the edges where the flow is positive]:

$$\mathcal{E}(f) \leq \sum_{k=1}^{n-1} 2\left(\frac{1}{k+1}\right)^2 (k+1) = \sum_{k=1}^{n-1} \frac{2}{k+1} \leq 2\log n$$

[where we have used that all edges have capacity 1 and that $(a+b)^2 \geq a^2 + b^2$ for a and b being positive incoming flows to some vertex and that there are $k+1$ vertices with the sum of coordinates $k+2$ and the factor 2 comes from adding the sum for the edges above the diagonal] completing the proof. □

Lecture 23

Recall that Example Sheet 1, Question 12a gives the following result.

Lemma 10.24 (Green's function and expected hitting time)**.** *et X be an irreducible Markov chain on a finite state space. Let τ be a stopping time satisfying $\mathbb{E}[\tau] < \infty$ and $\mathbb{P}_a(X_\tau = a) = 1$ for some a in the state space. Then for all x we have*

$$G_\tau(a,x) = \mathbb{E}_a[\tau] \cdot \pi(x).$$

We now also show the following.

Proposition 10.25 (Commute time identity)**.** *Let X be a reversible Markov chain on a finite state space. Then for all a, b we have*

$$\mathbb{E}_a[T_{a,b}] = \mathbb{E}_a[T_b] + \mathbb{E}_b[T_a] = c(G) \cdot R_{\text{eff}}(a,b),$$

where $c(G) = 2\sum_{e\in E} c(e)$ and $T_{a,b} = \inf\{t > 0 : X_t = a\, \exists u \in \{1,2,\dots t-1\} : X_u = b\}$ is the first time the walk comes back to a after having visited b.

This holds from the previous lemma and Lemma 10.13 applied to stopping time T_b.

Proof. The stopping time $T_{a,b}$, satisfies $\mathbb{P}_a\big(X_{T_{a,b}} = a\big) = 1$. Also note that we only visit a before time T_b, i.e.,

$$G_{T_{a,b}}(a,a) = G_{T_b}(a,a) = c(a)R_{\text{eff}}(a,b),$$

where the last equality follows from Lemma 10.13. We can now apply the previous lemma to finish the proof. □

Remark 10.26. *The previous identity immediately gives us that the effective resistance satisfies the triangle inequality [note that* $\mathbb{E}_a[T_b] + \mathbb{E}_b[T_a] + \mathbb{E}_b[T_c] + \mathbb{E}_c[T_b] = \mathbb{E}_a[T_b] + \mathbb{E}_b[T_c] + \mathbb{E}_c[T_b] + \mathbb{E}_b[T_a] \geq \mathbb{E}_a[T_c] + \mathbb{E}_c[T_a]$ *where we used that* $\mathbb{E}_x[T_y] + \mathbb{E}_y[T_z] \geq \mathbb{E}_x[T_z]$*]. Hence the effective resistance defines a metric on the graph G (the other two properties are trivially satisfied).*

So we have seen some ways of bounding the effective resistance and we have also seen that we can use the effective resistance to find an expected commute time (for a fixed vertex a, b commute time is the time it takes to get from a to b and then back to a). We notice that $\mathbb{E}_a[T_b] \leq \mathbb{E}_a[T_{a,b}]$ so moreover the effective resistance can be used to upper bound the hitting times. We have seen that it can also be used to find Green's function for some stopping times, e.g. we can find the expected time spent in a starting vertex until a visit to another fixed vertex and we can also bound escape probabilities. These results have many interesting applications (e.g. for infinite chains it can be used to determine if the chain is recurrent or not) and they can also be useful for bounding mixing times. We will soon see a particular example of Markov chains where the expected hitting times of a single vertex set for random walks on graphs give particularly good bounds on the mixing time of that chain. Before that, we will bound something called the cover time, which is the expected time it takes the Markov chain to visit all states.

10.2 Cover time

Definition 10.27 (Cover time and maximal hitting time)**.** For a Markov chain X on state space S we define a *maximal hitting time* to be

$$t_{\text{hit}} = \max_{x,y \in S} \mathbb{E}_x[T_y]\,.$$

We define a random variable $T_{\text{cov}} = \inf\{t \geq 0 : \forall v \in S \exists s \leq t \text{ such that } X_s = v\}$ to be the random variable which is the first time by which the chain has visited all states in S (i.e. it has covered S by then). We define the *cover time* to be

$$t_{\text{cov}} = \max_{x \in S} \mathbb{E}_x[T_{\text{cov}}]$$

//

We first give a bound on the cover time in terms of the maximal hitting time.

Theorem 10.28 (Cover time bound)**.** *If the size of the state space is* $|S| = n$ *then*

$$t_{\text{hit}} \leq t_{\text{cov}} \leq t_{\text{hit}} \left(1 + \frac{1}{2} + \frac{1}{3} + \ldots + \frac{1}{n-1}\right).$$

Moreover, we also have a (possibly stronger) lower bound in terms of $t^A_{\min} = \min_{x,y \in A, x \neq y} \mathbb{E}_x[T_y]$ *as*

$$t_{\text{cov}} \geq \max_{A \subset S} t^A_{\min} \left(1 + \frac{1}{2} + \ldots + \frac{1}{|A|-1}\right).$$

Starting from a fixed state x the idea for both parts is to permute the states at random and add expected hitting times to the next state, making sure we add this time only in case we have not already visited the next state so far. As any order of visiting the first k new states is equally likely, as our ordering is uniformly random, we multiply the kth term we add with $1/k$ (as for any fixed order in which the states are visited, this is the probability that the k th element in the uniformly random permutation is not visited before the first $k-1$ elements from this permutation).

Proof. The inequality $t_{\text{hit}} \leq t_{\text{cov}}$ follows trivially as for any state $x \in S$ we have $T_x \leq T_{\text{cov}}$ as $T_{\text{cov}} = \max_{x \in S} T_x$.

For an upper bound, notice that to cover the graph we need to hit all states, so we can take some order of states, e.g. lets label them from 0 to $n-1 = |S|-1$, and just add the expected hitting time of each state starting from the previous state in that order (assume our starting state was labelled 0). An upper bound on the expectation of this hitting time of state i from $i-1$ is always t_{hit} as that is the maximal expected hitting time. This gives a trivial upper bound of $t_{\text{cov}} \leq (n-1)t_{\text{hit}}$. However, it could happen that we have already visited state i before visiting $i-1$, so we have added some t_{hit} which was not necessary. If we work with the fixed ordering the probability of this happening will depend on the graph and that ordering and so we would not be able to get any better general bound[13]. So the main idea in this proof is to take a uniformly random permutation of all states $1, 2, \ldots, n-1$ say $\sigma \in S_{n-1}$ and then notice that to upper bound the cover time we need to add $\mathbb{E}_{\sigma(i-1)}\left[T_{\sigma(i)}\right]$ only if $\sigma(i)$ is the state which is visited last among $\sigma(1), \sigma(2), \ldots, \sigma(i)$, but this has probability $\frac{1}{i}$. This basically completes the proof, but we now present it more formally.

Label the starting state with 0 and let $\sigma \in S_{n-1}$ be a uniformly random permutation of the other states which are labelled from 1 to $n-1$. Define τ_k to be the first time the walk has visited states $\sigma(1), \ldots \sigma(k)$ and let $L_k = X_{T_k}$. We have for any states $x, y \in S$

$$\mathbb{E}_0[\tau_k - \tau_{k-1} \mid \sigma(k) = L_k = y, L_{k-1} = x] = \mathbb{E}_x[T_y] \leq t_{\text{hit}}$$

and averaging over x and y gives

$$\mathbb{E}_0[\tau_k - \tau_{k-1} \mid \sigma(k) = L_k] \leq t_{\text{hit}}.$$

We also know that $\mathbb{E}_0[\tau_k - \tau_{k-1} \mid \sigma(k) \neq L_k] = 0$ and as σ is a uniform permutation, conditional on the set $\{\sigma(1), \ldots, \sigma(k)\}$ the probability that $\sigma(k)$ is the state which is last visited among these is $1/k$ so

$$\mathbb{P}_0(\sigma(k) = L_k) = \frac{1}{k}.$$

This means that

$$\mathbb{E}_0[\tau_k - \tau_{k-1}] \leq \frac{1}{k} t_{\text{hit}}$$

and summing from 1 to $n-1$ completes the proof of the upper bound as $\tau_{n-1} = T_{\text{cov}}$.

The final lower bound works similarly to the above. We wait for the first time we visit a state in A, we label that state by 0 and start from there. We will visit all other states in A by the cover time. We label them with $1, 2, \ldots, |A|-1$. As for the upper bound on the cover time we take a random permutation $\sigma \in S_{|A|-1}$, but this time only of the states $1, \ldots, |A|-1$ which are all states

[13] e.g. if the graph is a path, starting from the leftmost point the order in which we visit new vertices is fixed and we know that the rightmost vertex is the last visited one, while if it is a complete graph then any order is equally likely

in A except the one we have visited the first. We notice that if we keep the labels as before for all $x, y \in A$

$$\mathbb{E}_0[\tau_k - \tau_{k-1} \mid \sigma(k) = L_k = y, L_{k-1} = x] = \mathbb{E}_x[T_y] \geq t^A_{\min}$$

and averaging over x and y gives

$$\mathbb{E}_0[\tau_k - \tau_{k-1} \mid \sigma(k) = L_k] \geq t^A_{\min}.$$

As before $\mathbb{E}_0[\tau_k - \tau_{k-1} \mid \sigma(k) \neq L_k] = 0$ and $\mathbb{P}_0(\sigma(k) = L_k) = \frac{1}{k}$. This means that

$$\mathbb{E}_0[\tau_k - \tau_{k-1}] \geq \frac{1}{k} t^A_{\min}$$

and summing from 1 to $|A| - 1$ and taking maximum over subsets A completes the proof of the lower bound as $\tau_{|A|-1} \leq T_{\text{cov}}$. □

We now state the following proposition which holds easily using the commute time identity and then we use it with the previous theorem and the already established effective resistance to bound the cover time for the grid.

Proposition 10.29 (Maximal hitting time and effective resistance)**.** *For a reversible chain on state space V we have that*

$$\frac{1}{2} \max_{a,b \in V} c(G) R_{\text{eff}}(a, b) \leq t_{\text{hit}} \leq \max_{a,b \in V} c(G) R_{\text{eff}}(a, b).$$

This follows directly from the commute time identity.

Proof. The upper bound directly holds from the commute time identity (Proposition 10.25) as $\mathbb{E}_a[T_{a,b}] \geq \mathbb{E}_b[T_a]$. The lower bound also holds from the commute time identity as

$$\max_{a,b} \{\mathbb{E}_b[T_a], \mathbb{E}_a[T_b]\} \geq \frac{1}{2} \mathbb{E}_a[T_{a,b}]$$

and so the proof is completed by taking maximums. □

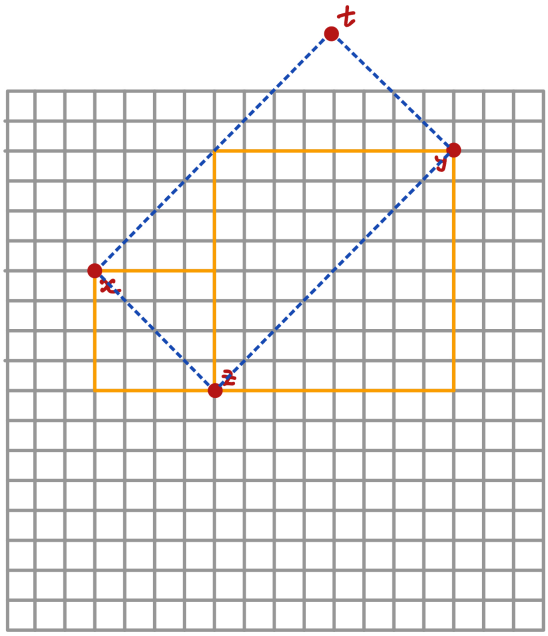
Claim 10.30 (Cover time for the grid)**.** *For a random walk on an $n \times n$ grid, i.e. box B_n from Proposition 10.23 we have*

$$t_{\text{hit}} \asymp n^2 \log n \quad \textit{and} \quad t_{\text{cov}} \asymp n^2 (\log n)^2.$$

We use the bounds from Proposition 10.23 and the above claim and Theorem 10.28 where for the lower bound we take $A = \{(k\lfloor\sqrt{n}\rfloor, k\lfloor\sqrt{n}\rfloor) : k = 0, 1, \ldots, \lfloor\sqrt{n}\rfloor\}$. To bound effective resistance for any two points use the triangle inequality for the upper bound and for lower bound use cutsets analogous to ones from the proof of Proposition 10.23.

Proof. The proof of result fo t_{hit} directly from the above claim and Proposition 10.23. We only need to justify why Proposition 10.23 gives an upper bound by $\log n$ on $R_{\text{eff}}((a, b), (c, d))$ when $(a, b) \neq (1, 1)$ or $(c, d) \neq (n, n)$. If $(a, b) = (1, 1)$ and $(c, d) = (k, k)$ we have an upper bound by $\asymp \log(k)$ by noticing that if we remove edges we increase their resistances to infinity so by Rayleigh monotonicity this means that the effective resistance increases, so we can upper bound $R_{\text{eff}}((1, 1), (k, k))$ by the same quantity but for the $[1, k]^2$ grid. The same holds for any two coordinates which form a square. If they do not form a square, we can use the triangle inequality, which we know holds for effective

resistances, and the bound we have for vertices forming a square. If we want to bound the effective resistance from (a,b) to (c,d) first assume that $a-c$ and $b-d$ have the same parity. Then for some (r,s) we have that (a,b) and (r,s) as well as (c,d) and (r,s) form squares and so the triangle inequality gives bound $2R_{\text{eff}}((1,1),(n,n))$. Such (r,s) exist because at least one of the pairs of lines through (a,b) of slope 1 and line through (c,d) of slope -1 or through (a,b) of slope -1 and line through (c,d) of slope 1 will intersect inside the box (this is because if WLOG the horizontal distance between $x=(a,b)$ and $y=(c,d)$ is larger than the vertical one, we can notice that the vertical distance between the two intersection points z and t is the same as the horizontal distance between x and y and so at least one the points is in the box, as is demonstrated in the picture below where $(r,s)=z$). If $a-c$ and $c-d$ are not the same parity first use the triangle inequality on resistances between (a,b) and $(a\pm 1,b)$ [which is from Thomson's principle bounded by 1 by considering a unit flow which has value one on this edge] and $(a\pm 1,b)$ and (c,d) [which we can bound as distances of first and second coordinates have the same parity] where we pick the sign to be + if c>a and − if $c\le a$.



For vertices x and y in the box for which differences of corresponding coordinates has the same parity the intersections of lines with slope ± 1 from them are vertices of the grid z and t and at least one of them has to be inside the box we are considering.

For the t_{cov} upper bound now follows from Theorem 10.28 and the lower bout also follows from the last part of the theorem by setting $A=\{(k\lfloor\sqrt{n}\rfloor,k\lfloor\sqrt{n}\rfloor):k=0,1,\dots,\lfloor\sqrt{n}\rfloor\}$, where we note that the lower bound on the effective resistance from $(k\lfloor\sqrt{n}\rfloor,k\lfloor\sqrt{n}\rfloor)$ to $((k+1)\lfloor\sqrt{n}\rfloor,(k+1)\lfloor\sqrt{n}\rfloor)$ follows as we can consider the cutsets of all edges in the box which connect vertices on $\|\cdot\|_\infty$ distance ℓ and $\ell+1$ from $(k\lfloor\sqrt{n}\rfloor,k\lfloor\sqrt{n}\rfloor)$ and we get $\frac{1}{4}\cdot$ the lower bound from Proposition 10.23 applied to side length $\sqrt{n}$ box, which is of order $\log n$ (we multiply by 1/4 as we have four quadrants around $(k\lfloor\sqrt{n}\rfloor,k\lfloor\sqrt{n}\rfloor)$ over which we count the number of edges of given distance while in the proposition we only count ones from the upper right one). □

In the next subsection, we show how to control the mixing times by some maximal hitting times and cover times for a class of chains called Lamplighter graphs.

10.3 Lamplighter graphs

Lecture 24

Proofs from this section are not examinable. However, you are encouraged to have a look at them as some ideas from them might be useful in general.

We define a Markov chain known as the random walk on a lamplighter graph as follows.

Definition 10.31 (Walk on a lamplighter graph)**.** For a connected graph $G = (V, E)$ we define a *lamplighter graph* to be a graph $G^\diamond = (V^\diamond, E^\diamond)$ where $V^\diamond = V \times \{0,1\}^V$ and $\{(v, f), (v', f')\} \subset V \times \{0,1\}^V$ is an edge in $E^\diamond$ if $\{v, v'\} \in E$ or $v = v'$ and $f(u) = f'(u)$ for all $u \notin \{v, v'\}$. We call v the *position of the lamplighter* and we call f the *configuration of lamps*. *Lazy random walk on lamplighter* is a Markov chain on $G^\diamond$ with the following non-zero transitions

$$P_\diamond((v, f), (v', f')) = \frac{1}{8\deg(v)} \text{ if } \{(v, f), (v', f')\} \in E^\diamond \text{ and } v \neq v'$$
$$P_\diamond((v, f), (v, f')) = \frac{1}{4} \text{ if } \{(v, f), (v, f')\} \in E^\diamond.$$

We think of $G^\diamond$ as G with lamps at each vertex which have states $\{0, 1\}$ (or off and on). We then run a lazy simple random walk on G which changes the states of the lamps as it moves according to the *refresh, move, refresh rule*, which means that for each move it refreshes the lamp at the state where it currently is (to a uniform state from $\{0, 1\}$) then it moves according to lazy walk on G and refreshes the lamp at the new location.

It is not hard to check that the invariant distribution of the (lazy) lamplighter walk is given by $\pi^\diamond = \pi \times \mathcal{L}(U\{0,1\}^V)$ where $\mathcal{L}(U\{0,1\}^V)$ is a distribution of random variable where each state in V has a value 0 or 1 assigned, each one equally likely and π is invariant of a simple random walk on G. We can also check that $P_\diamond$ is in detailed balance with $\pi_\diamond$ so it is reversible. //

Remark 10.32. *We could define a lamplighter walk corresponding to the simple random walk on G or any transition matrix P on G, as the walk which moves according to P and just refreshes the states of the lamps at the start and end of each random walk step. However, we consider the lazy walk only as that makes some of our future analysis less complicated (and in fact, in a very general case some of the following results would not necessarily be true). We also note that this could be generalised by having lamps with more than two states where states are changed according to some rule which is not necessarily just the "refreshed to uniform" rule, but here we will only work with two state lamps which are refreshed to a new uniform state on each switch.*

We note that while the above might look similar to the Glauber dynamics on a graph with colouring in the sense that we can think of state of the lamps as two colours the differences are that in the Glauber dynamics, we use the connections in the graph to decide the rule we use to update the colours of vertices once we have chosen the vertex we are updating, while we always pick a uniform vertex to update, while here we use the connections in the graph to see which state/colour we will update next and the updates are done according to some prescribed rule which is independent of the states of the neighbours in G of the vertex we are updating.

We are interested in the mixing and relaxation time of the lamplighter walk chain. We can notice that the chain is lazy so the relaxation time will be of the order of $(1-\lambda_2)^{-1} = \gamma^{-1}$ i.e. of the inverse of the spectral gap, as for the lazy chain all eigenvalues are bounded away from -1 (the laziness parameter is smaller than $1/2$ so they do not have to be positive, but they are still bounded away by a constant from -1) so the absolute spectral gap is the same order as the spectral gap, i.e. $\gamma_* \gtrsim \gamma$. We can right away notice that once we have visited a vertex in G the lamp at it was refreshed, so is from now on going to be in the uniform state. This means that at the random time T_{cov} all lamps

will have been refreshed so the state of all lamps will be uniform. So if we can find a large enough time which is bigger than the mixing time of lazy walk on G and with a large probability bigger than T_{cov} that would be an upper bound on some mixing time, as positions of all lamps will likely be uniform at that point and the random walk on G would also be at a location close to π. We will now show this kind of result more formally.

Throughout the following, we will add a $\diamond$ when we talk about the quantities related to the lamplighter walk while the quantities without the $\diamond$ will correspond to the lazy simple random walk on G.

Lemma 10.33 (Relaxation time of the lamplighter)**.** *Let G be a connected graph with at least two vertices and $G^\diamond$ a corresponding lamplighter. We have*

$$\frac{1}{4\log 4} t_{\text{hit}} \leq t_{\text{rel}}^\diamond \leq 6 t_{\text{hit}}.$$

Lower bound: use the variational characterisation to find an upper bound on the lamplighter chain after $t_{\text{hit}}/4$ steps for the test function which is $\phi(v,f) = f(w)$ where w is a vertex maximising $\mathbb{E}_\pi[T_w]$ [get $\mathcal{E}_t^\diamond(\varphi) = \frac{1}{4}\mathbb{P}_\pi(T_w \leq t)$ and show $\mathbb{P}_\pi(T_w \leq t) \leq \frac{3}{4}$ for $t = t_{\text{hit}}/4$]. Upper bound: consider for eigenfunction ϕ of eigenvalue $\lambda_{2,\diamond}$ the quantity $M = \max_{x\in V f,g\in\{0,1\}^V} \frac{|\varphi(x,f)-\varphi(x,g)|}{\|f-g\|_1}$ and get $\lambda_{2,\diamond}^{2t_{\text{hit}}} M \leq M \sup_{x\in V f,g\in\{0,1\}^V} \mathbb{E}\left[\frac{\|F_t - G_t\|_1}{\|f-g\|_1}\right]$ where $(X_t,F_t),(X_t,G_t)$ are lamplighter couplings using the same random walk and updating the lamps in the same way. The proof follows from this if $M > 0$ and if not we see that ϕ is eigenfunction for P with eigenvalue $\lambda_{2,\diamond}$ which also gives the wanted upper bound.

Proof. * We first show the *lower bound*. The main idea is to use the variational characterisation of the spectral gap for the chain with the transition probabilities as the lamplighter after $t_{\text{hit}}/4$ steps, rather than just looking at one step of the chain. This is not an unusual idea because, in general, we know that the eigenvalues of a chain after t steps are just the initial eigenvalues raised to power t, so if we can bound the spectral gap of the chain after t steps by a constant smaller than 1 this would imply a lower bound on the relaxation time for the original chain of order t [exact algebra which allows us to conclude this is similar to the final steps presented here in the proof of this lower bound]. The motivation for applying it in this context is that if we wanted to directly upper bound the spectral gap (using variational characterisation) we would need to be able to upper bound the the ratio of Dirichlet form and the variance of some function after one step of the chain by $\lesssim \frac{1}{t_{\text{hit}}}$ which does not seem easy. However, after time which is small enough constant times t_{hit} it is not unreasonable to expect that the vertex in G which is least likely to be visited by that time by the walk starts from π, will not be visited by the walk starting from π with some probability lower bounded by a constant bigger than 0 (proof of this uses that we can compare expected hitting time from π for the worst vertex with t_{hit}). Then taking the function which just takes the values of the lamp at that vertex gives a constant (smaller than 1) upper bound on the ratio of the Dirichlet form and variance for the lamplighter and upper bound the spectral gap. The formal proof is as follows:

Fix a vertex $w \in G$ that maximizes $\mathbb{E}_\pi[T_w]$, and define $\varphi : V^\diamond \to \{0,1\}$ by $\varphi(v,f) = f(w)$. Then $\text{Var}_{\pi^\diamond}(\varphi) = \frac{1}{4}$. Let (Y_t) be the Markov chain on $G^\diamond$ with initial distribution $\pi^\diamond$, and so Y_t has distribution $\pi^\diamond$ for all $t \geq 0$. We write $Y_t = (X_t, F_t)$, where X_t is the position of the walk at time t, and F_t is the configuration of lamps at time t. If we label by $\mathcal{E}_t^\diamond$ the Dirichlet form corresponding to matrix $P_\diamond^t$ (i.e. chain where one step is according to the distribution of lamplighter after t steps) and then condition on the walk's path up to time t we have that

$$\mathcal{E}_t^\diamond(\varphi) = \frac{1}{2}\mathbb{E}_{\pi^\diamond}\Big[(\varphi(Y_t) - \varphi(Y_0))^2\Big] = \frac{1}{2}\mathbb{E}_{\pi^\diamond}\Big[\mathbb{E}\Big[(\varphi(Y_t) - \varphi(Y_0))^2 \;\Big|\; X_0,\ldots,X_t\Big]\Big].$$

Observe that

$$\mathbb{E}\Big[(\varphi(Y_t)-\varphi(Y_0))^2 \;\Big|\; X_0,\dots,X_t\Big] = \mathbb{E}\Big[(F_t(w)-F_0(w))^2 \;\Big|\; X_0,\dots,X_t\Big] = \frac{1}{2}\mathbb{1}_{\{T_w\le t\}},$$

as $|F_t(w)-F_0(w)|=1$ if and only if the walk visits w by time t, and, at the walk's last visit to w before or at time t, the lamp at w is refreshed to a state different from its initial state. Combining the above equality two equalities shows that for any t,

$$\mathcal{E}^{\diamond}_t(\varphi) = \frac{1}{4}\mathbb{P}_\pi(T_w\le t)\,. \tag{10.1}$$

If a walk on G started at v has not hit w by time t, the expected additional time to arrive at w is bounded above by t_{hit} giving that

$$\mathbb{E}_v[T_w] \le t + t_{\text{hit}}\mathbb{P}_v(T_w>t)$$

and averaging this over π shows that

$$\mathbb{E}_\pi[T_w] \le t + t_{\text{hit}}\mathbb{P}_\pi(T_w>t)\,.$$

By Example Sheet 2, Question 12 and our choice of w, we have $t_{\text{hit}}\le 2\mathbb{E}_\pi[T_w]$, whence above implies that

$$t_{\text{hit}} \le 2t + 2t_{\text{hit}}\mathbb{P}_\pi(T_w>t) \implies 2t_{\text{hit}}\mathbb{P}_\pi(T_w\le t)\le 2t+t_{\text{hit}}.$$

Substituting $t=t_{\text{hit}}/4$ and rearranging yields

$$\mathbb{P}_\pi(T_w\le t_{\text{hit}}/4)\le\frac{3}{4}.$$

Let $\lambda_{2,\diamond}$ be the second largest eigenvalue of $P_\diamond$. By the variational characterisation of the spectral gap (Theorem 5.3), we thus have

$$1-\lambda_{2,\diamond}^{t_{\text{hit}}/4} \le \frac{\mathcal{E}^{\diamond}_{t_{\text{hit}}/4}(\varphi)}{\text{Var}_{\pi^\diamond}(\varphi)}\le\frac{3}{4},$$

as $\lambda_{2,\diamond}^{t_{\text{hit}}/4}$ is the second largest eigenvalue of $P_\diamond^{t_{\text{hit}}/4}$. Therefore, rearranging and using $1-\lambda_{2,\diamond}\le\log(\lambda_{2,\diamond}^{-1})$

$$\frac{1}{4}\le\lambda_{2,\diamond}^{t_{\text{hit}}/4} \implies \log(4)\ge\frac{t_{\text{hit}}\log(\lambda_{2,\diamond}^{-1})}{4}\ge\frac{t_{\text{hit}}(1-\lambda_{2,\diamond})}{4}$$

giving the claimed lower bound. (Note that as the walk is lazy with parameter $1/4$ by Example Sheet 2, Question 1 $\frac{1}{\gamma^\diamond}\ge\frac{4(|G^\diamond|-1)}{3|G^\diamond|}\ge 1$ for $|G^\diamond|\ge 4$ so $\lambda_{2,\diamond}>0$ which we used in the log bound.)

We now move on to the *upper bound*. We again control the second largest eigenvalue of the chain after t steps rather than after 1 step directly. The idea is to consider the absolute value of the difference of eigenfunction corresponding to the second largest eigenvalue at the states which correspond to the walk being in the same vertex of a graph but to different configurations of lamps and look at the ratio of this quantity and the number of lamps which differ in the two configurations. If the maximum of this ratio $M=0$, then the eigenfunction does not depend on the configuration of lamps, and it is an eigenfunction for the transition of the walk with the same eigenvalue and the proof follows (as relaxation time is bounded by mixing time which is bounded by a constant multiple of t_{hit}). If $M>0$ we look at the chain after $2t_{\text{hit}}$ states and couple the transitions of it when started from the same vertex of the graph by running the same walks and updating lamps at every visited vertex in the same way and noticing that by the time $2t_{\text{hit}}$ each vertex has probability that it was

visited which is at least $1/2$ and once it was visited the two configurations have the same state of the lamp. This means that the expected ratio of the number of lamps where configurations differ is dropped by a factor of $1/2$ after time $2t_{\text{hit}}$ and we can use this to upper bound the second largest eigenvalue of the transition matrix at this time by considering $\lambda_{2,\diamond}^{2t_{\text{hit}}} M$ [see proof for details]. Proof is as follows:

If $\lambda_{2,\diamond} < \frac{1}{2}$, then $t_{\text{rel}}^{\diamond} \leq 2 \leq 2t_{\text{hit}}$. Without loss of generality, assume that $\lambda_{2,\diamond} \geq \frac{1}{2}$. Suppose that φ is an eigenfunction for $P_\diamond$ with eigenvalue $\lambda_{2,\diamond}$. Note that for lamp configurations f and g on G, the $\mathcal{L}^1$ norm $\|f-g\|_1$ is equal to the number of bits in which f and g differ. Let

$$M = \max_{x \in V f,g \in \{0,1\}^V} \frac{|\varphi(x,f) - \varphi(x,g)|}{\|f-g\|_1}.$$

If $M = 0$, then $\varphi(x,f)$ depends only on x and $\psi(x) = \varphi(x,f)$ is an eigenfunction for the transition matrix P with eigenvalue $\lambda_{2,\diamond}$. Applying $2d(t) \geq \lambda_{2,\diamond}^t$ from Theorem 4.12 with $t = 2t_{\text{hit}} + 1$ together with $t_{\text{mix}} \leq 2t_{\text{hit}} + 1$ from Example Sheet, Question 11 yields

$$\lambda_{2,\diamond}^{2t_{\text{hit}}+1} \leq 2d(2t_{\text{hit}} + 1) \leq 2 \cdot \frac{1}{4} = \frac{1}{2}.$$

Now we treat the case $M > 0$. Couple two lamplighter walks, one started at (x,f) and one at (x,g), by using the same lamplighter steps and updating the configurations so that they agree at each site visited by the lamplighter. Let (X_t, F_t) and (X_t, G_t) denote the positions of the coupled walks after $t = 2t_{\text{hit}}$ steps. Because φ is an eigenfunction for $P_\diamond$, so $P_\diamond^t \varphi = \lambda_{2,\diamond}^t \varphi$ we have

$$\begin{aligned}
\lambda_{2,\diamond}^{2t_{\text{hit}}} M &= \sup_{x \in V f,g \in \{0,1\}^V} \frac{\left| P_\diamond^{2t_{\text{hit}}} \varphi(x,f) - P_\diamond^{2t_{\text{hit}}} \varphi(x,g) \right|}{\|f-g\|_1} \\
&\leq \sup_{x \in V f,g \in \{0,1\}^V} \mathbb{E}\left[\frac{|\varphi(X_t,F_t) - \varphi(X_t,G_t)|}{\|F_t - G_t\|_1} \frac{\|F_t - G_t\|_1}{\|f-g\|_1} \right] \leq M \sup_{x \in V f,g \in \{0,1\}^V} \mathbb{E}\left[\frac{\|F_t - G_t\|_1}{\|f-g\|_1} \right].
\end{aligned}$$

At time $t = 2t_{\text{hit}}$, each lamp that contributes to $\|f-g\|_1$ has a probability of at least $1/2$ of having been visited (by applying Markov's inequality), so $\mathbb{E}[\|F_t - G_t\|_1] \leq \|f-g\|_1/2$. Dividing by M gives the bound of $\lambda_{2,\diamond}^{t_{\text{hit}}} \leq \frac{1}{2}$.

Thus in both cases, $\lambda_{2,\diamond}^{2t_{\text{hit}}+1} \leq \frac{1}{2}$. Let $\gamma_\diamond = 1 - \lambda_{2,\diamond}$. Then

$$\frac{1}{2} \geq (1-\gamma_\diamond)^{2t_{\text{hit}}+1} \geq 1 - (2t_{\text{hit}} + 1)\gamma_\diamond \geq 1 - 3t_{\text{hit}}\gamma_\diamond.$$

We conclude that $t_{\text{rel}}^{\diamond} \leq 6t_{\text{hit}}$. □

We now give a mixing time bound.

Theorem 10.34 (Mixing time of the lampligher walk)**.** *Let G be a graph and $G^\diamond$ a corresponding lamplighter. We have*

$$\frac{1}{8} t_{\text{cov}} \leq t_{\text{mix}}^{\diamond} \leq 17 t_{\text{cov}}.$$

Upper bound: At time $u = 8t_{\text{cov}} + t_{\text{mix}}(1/8) \leq 17t_{\text{cov}}$ there is a probability $\geq 1/8$ that the cover time happened and chain had additional $t_{\text{mix}}(1/8)$ steps after T_{cov}. This implies the upper bound using the convexity of total variation and the fact that after the cover time, all lamps are uniform. Lower bound: Get that $s^\diamond(t) \geq \mathbb{P}_w(T_{\text{cov}} > t)$ where $s^\diamond$ is the separation distance for the lamplighter. This gives that $\mathbb{P}_w(T_{\text{cov}} > 4t_{\text{mix}}^{\diamond}) \leq 1/2$ for all $w \in V$, which by doing repeated trials of time length $4t_{\text{mix}}$, which attempt to each cover the graph and succeed with probability $\geq 1/2$ implies that t_{cov} is at most $4t_{\text{mix}}$ times and expectation of geometric $1/2$.

Proof. * We first prove the *upper bound*.

Let (X_t, F_t) denote the state of the lamplighter chain at time t. We will run the lamplighter chain long enough that, with high probability, every lamp has been visited and enough additional steps have been taken to randomize the position of the lamplighter.

Set $u = 8t_{\text{cov}} + t_{\text{mix}}(1/8)$ and fix an initial state $(v, 0)$. Define the probability distribution μ_s on $G^\diamond$ by

$$\mu_s = \mathbb{P}_{(v,0)}((X_u, F_u) \in \cdot \mid T_{\text{cov}} = s)\,.$$

Then

$$P^u_\diamond((v,0),\cdot) = \sum_{s\in V} \mathbb{P}_{(v,0)}(T_{\text{cov}} = s)\,\mu_s.$$

By the convexity of total variation distance,

$$\|P^u_\diamond((v,0),\cdot) - \pi^\diamond\|_{\text{TV}} \le \sum_s \mathbb{P}_{(v,0)}(T_{\text{cov}} = s)\,\|\mu_s - \pi^\diamond\|_{\text{TV}}.$$

Since by Markov's inequality $\mathbb{P}(T_{\text{cov}} > 8t_{\text{cov}}) < \frac{1}{8}$ and the total variation distance between distributions is bounded by 1, we can bound

$$\|P^u_\diamond((v,0),\cdot) - \pi^\diamond\|_{\text{TV}} \le \frac{1}{8} + \sum_{s\le 8t_{\text{cov}}} \mathbb{P}_{(v,0)}(T_{\text{cov}} = s)\,\|\mu_s - \pi^\diamond\|_{\text{TV}}. \tag{10.2}$$

Recall that G has vertex set V. Let ν denote the uniform distribution on $\{0,1\}^V$. For $s \le u$, conditional on $T_{\text{cov}} = s$ and $X_s = x$, the distribution of F_u equals ν, the distribution of X_u is $P^{u-s}(x,\cdot)$, and F_u and X_u are independent. Thus,

$$\mu_s = \sum_{x\in V} [P^{u-s}(x,\cdot)\times\nu]\mathbb{P}_{(v,0)}(X_s = x \mid T_{\text{cov}} = s)\,.$$

By the convexity of total variation distance and Example Sheet 1, Question 1 (giving an upper bound on the total variation distance for product of distributions), since $\pi^\diamond = \pi\times\nu$,

$$\|\mu_s - \pi^\diamond\|_{\text{TV}} \le \sum_{x\in V} \|P^{u-s}(x,\cdot)\times\nu - \pi^\diamond\|_{\text{TV}}\mathbb{P}_{(v,0)}(X_s = x \mid T_{\text{cov}} = s) \le \max_{x\in V}\|P^{u-s}(x,\cdot) - \pi\|_{\text{TV}}.$$

For $s \le 8t_{\text{cov}}$, we have $u - s \ge t_{\text{mix}}(1/8)$, by definition of u. Consequently, for $s \le 8t_{\text{cov}}$,

$$\|\mu_s - \pi^\diamond\|_{\text{TV}} \le \frac{1}{8}.$$

Using this with (10.2) shows that

$$\|P^u_\diamond((v,0),\cdot) - \pi^\diamond\|_{\text{TV}} \le \frac{1}{8} + (1)\left(\frac{1}{8}\right) = \frac{1}{4}.$$

To complete the proof of the upper bound, by Proposition 1.15 and Example Sheet 2 Question 11 saying $t_{\text{mix}} \le 2t_{\text{hit}} + 1$

$$t_{\text{mix}}(1/8) \le 3t_{\text{mix}} \le 9t_{\text{cov}}.$$

Since $u = 8t_{\text{cov}} + t_{\text{mix}}(1/8)$, it follows that $t_{\text{mix}} \le 17t_{\text{cov}}$.

Lower bound. We will first prove

$$s^\diamond(t) \ge \mathbb{P}_w(T_{\text{cov}} > t) \tag{10.3}$$

for any $w \in V, t > 0$ where $s^\diamond$ is the separation distance (recall Definition 2.7) for the lamplighter walk. Let $w_t \in V$ be the vertex minimizing $\mathbb{P}_w(X_t = w_t \mid T_{\text{cov}} \le t)/\pi(w_t)$. Since $\mathbb{P}_w(X_t = \cdot \mid T_{\text{cov}} \le t)$ and π are both probability distributions on V, we have $\mathbb{P}_w(X_t = w_t \mid T_{\text{cov}} \le t) \le \pi(w_t)$. Suppose $|V| = n$. Since the only way to go from all lamps off to all lamps on is to visit every vertex and turn every lamp on when you visit it the last time before time t, we have

$$\frac{P^t_\diamond((w,0),(w_t,1))}{\pi^\diamond(w_t,1)} = \frac{\mathbb{P}_w(T_{\text{cov}} \le t)\, 2^{-n}\mathbb{P}_w(X_t = w_t \mid T_{\text{cov}} \le t)}{2^{-n}\pi(w_t)} \le \mathbb{P}_w(T_{\text{cov}} \le t)\,.$$

where 1 and 0 above represent the identically 1 and identically 0 functions respectively. Subtracting from 1 and using the definition of the separation distance yields $s^\diamond(t) \ge \mathbb{P}_w(T_{\text{cov}} > t)$.

Lemma 2.10 yields

$$s^\diamond(4t^\diamond_{\text{mix}}) \le 1 - (1 - (2d^\diamond(2t^\diamond_{\text{mix}})) \wedge 1)^2 \le 1 - \left(\frac{3}{4}\right)^2 < \frac{1}{2}.$$

By (10.3) with $t = 4t^\diamond_{\text{mix}}$, we have $\mathbb{P}_w(T_{\text{cov}} > 4t^\diamond_{\text{mix}}) < \frac{1}{2}$ for any $w \in V$. We will now show that this implies that $t_{\text{cov}} \le 8t^\diamond_{\text{mix}}$. In fact, we show that for any t if $\mathbb{P}_w(T_{\text{cov}} > t) < \frac{1}{2}$ then $t_{\text{cov}} \le 2t$ and the proof will follow by using this with $t = 4t^\diamond_{\text{mix}}$. If we run a walk from any vertex for time t the probability that we have covered all vertices by then is at least $1/2$, then if we run a random walk repeatedly k times for time t we have $\mathbb{P}_x(T_{\text{cov}} > kt) \le \frac{1}{2^k}$. So we conclude that T_{cov} is dominated by t times a geometric random variable with success probability $1/2$ and so t_{cov} is at most $2t$. □

Corollary 10.35 (Relaxation and mixing time of a lamplighter on a two dimensional box)**.** *Applying Lemma 10.33 and Theorem 10.34 with the Claim 10.30 gives that for a lamplighter walk on the $n \times n$ grid we have*

$$t^\diamond_{\text{rel}} \asymp n^2 \log n \quad \textit{and} \quad t_{\text{mix}} \asymp n^2 (\log n)^2.$$

A Non-Markovian coupling counterexample to Theorem 2.3

We give here an example of a Markov chain for which there is a non-Markovian coupling from two states x and y such that $\|P^t(x,\cdot) - P^t(y,\cdot)\|_{\text{TV}} > \mathbb{P}_{x,y}(\tau_{\text{couple}} > t)$ for some $t \in \mathbb{N}$. In fact, the constructed chain will satisfy that $\lim_{t\to\infty} \|P^t(x,\cdot) - P^t(y,\cdot)\|_{\text{TV}} > 0$ (we know the limit exists as this total variation distance is a decreasing function in t), but $\mathbb{P}_{x,y}(\tau_{\text{couple}} < \infty) = 1$, and so there must be $t \in \mathbb{N}$ such that $\|P^t(x,\cdot) - P^t(y,\cdot)\|_{\text{TV}} > \mathbb{P}_{x,y}(\tau_{\text{couple}} > t)$ (as the right hand side converges to 0 as $t \to \infty$). The chain, on state space, $\{1,2,3,4,5,6\}$ is defined as follows:

$$\begin{aligned} &P(1,3) = P(2,4) = P(3,5) = P(4,6) = 1 - p \\ &P(1,4) = P(2,3) = P(3,6) = P(4,5) = p \\ &P(5,5) = P(6,6) = 1 \end{aligned}$$

while all other transition probabilities are 0. The chain has two possible absorbing states, and it will reach one of them after 3 steps. The probability of reaching either of them is positive from both states, which directly implies that $\|P^3(1,\cdot) - P^3(2,\cdot)\|_{\text{TV}} > 0$ and as the chain does not move after time 3 it means that $\lim_{t\to\infty} \|P^t(1,\cdot) - P^t(2,\cdot)\|_{\text{TV}} = \|P^3(1,\cdot) - P^3(2,\cdot)\|_{\text{TV}} > 0$. Take $p = 1 - \frac{\sqrt{2}}{2}$ and consider the following (non-Markovian) coupling (X, Y) of the Markov chain started from states 1 and 2, respectively.

$$\big((X_1, X_2), (Y_1, Y_2)\big) = \begin{cases} ((3,5),(3,5)) & \text{with probability } \frac{\sqrt{2}-1}{2}, \\ ((3,5),(3,6)) & \text{with probability } \frac{3-2\sqrt{2}}{2}, \\ ((3,5),(4,5)) & \text{with probability } \frac{\sqrt{2}-1}{2}, \\ ((3,6),(4,6)) & \text{with probability } \frac{\sqrt{2}-1}{2}, \\ ((4,5),(4,6)) & \text{with probability } \frac{3-2\sqrt{2}}{2}, \\ ((4,6),(4,6)) & \text{with probability } \frac{\sqrt{2}-1}{2}. \end{cases}$$

We can check that the marginals are

$$(X_1, X_2) = \begin{cases} (3,5) & \text{with probability } \frac{1}{2}, \\ (3,6) & \text{with probability } \frac{\sqrt{2}-1}{2}, \\ (4,5) & \text{with probability } \frac{3-2\sqrt{2}}{2}, \\ (4,6) & \text{with probability } \frac{\sqrt{2}-1}{2}. \end{cases}$$

and,

$$(Y_1, Y_2) = \begin{cases} (3,5) & \text{with probability } \frac{\sqrt{2}-1}{2}, \\ (3,6) & \text{with probability } \frac{3-2\sqrt{2}}{2}, \\ (4,5) & \text{with probability } \frac{\sqrt{2}-1}{2}, \\ (4,6) & \text{with probability } \frac{1}{2}. \end{cases}$$

which agrees with the transitions of our chain. Note that in each of the six cases in the coupling, we have either $X_1 = Y_1$ or $X_2 = Y_2$. We therefore have $\mathbb{P}_{x,y}(\tau_{\text{couple}} < \infty) = 1$, as desired.