

# Mixing times of Markov Chains

Jan Olucha Fuentes

May 22, 2025

Dear Reader,

This is a set of lecture notes typed for the course *Mixing Times of Markov Chains* taught at the University of Cambridge during the academic year 2024-2025.

Yours falsely,

JOF.

# Contents

<b>1</b>	<b>Markov Chains</b>	<b>7</b>
1.1	Elementary concepts . . . . .	7
1.2	Class Structure . . . . .	8
1.3	Transience and recurrence . . . . .	8
1.4	Invariant Distributions . . . . .	9
1.5	Time Reversal . . . . .	10
1.6	The Ergodic Theorem . . . . .	11
<b>2</b>	<b>The Total Variation and Coupling</b>	<b>13</b>
2.1	Examples . . . . .	20
<b>3</b>	<b>Mixing Times</b>	<b>21</b>
3.1	Examples . . . . .	28
<b>4</b>	<b>Markovian Couplings</b>	<b>33</b>
4.1	Examples . . . . .	36
<b>5</b>	<b>Strong Stationary Times</b>	<b>43</b>
5.1	Examples . . . . .	47
<b>6</b>	<b>Cutoff</b>	<b>57</b>
<b>7</b>	<b><math>\mathcal{L}^p</math> distance</b>	<b>61</b>
<b>8</b>	<b>Spectral Decomposition</b>	<b>65</b>
8.1	Examples . . . . .	68
<b>9</b>	<b>The Relaxation Time</b>	<b>71</b>
9.1	A necessary condition for cutoff . . . . .	77
9.2	Examples . . . . .	80

<b>10 Transitive chains</b>	<b>81</b>
10.1 Examples . . . . .	83
10.2 Wilson's Method . . . . .	84
<b>11 Dirichlet Forms and Bottleneck Ratio</b>	<b>89</b>
11.1 Canonical Paths . . . . .	98
11.2 Comparison technique . . . . .	102
11.3 Bottleneck ratio . . . . .	110
11.4 Expander graphs . . . . .	118
<b>12 Spectral profile and isoperimetric profile</b>	<b>125</b>
12.1 Spectral Profile . . . . .	125
12.2 Isoperimetric Profile . . . . .	138
<b>13 Geometric Techniques</b>	<b>143</b>
13.1 Varopoulos-Carne bound . . . . .	143
13.2 Path Coupling . . . . .	149
<b>14 Hit-Mix</b>	<b>161</b>
14.1 Properties of hit-time . . . . .	162
14.2 An upper bound on mixing time . . . . .	168
14.3 Hit cutoff . . . . .	172
14.4 Trees . . . . .	176
<b>15 Electrical Networks</b>	<b>187</b>
15.1 Effective Resistance . . . . .	193
15.2 Lower Bounds on Effective Resistance . . . . .	201
15.3 Cover times . . . . .	208
<b>A Product Chains</b>	<b>215</b>
<b>B Induced Chains</b>	<b>217</b>
<b>C Classic Markov Chains</b>	<b>219</b>
<b>D Background Results</b>	<b>221</b>

# Notation

☞: denotes an idea or proof that is hard, and/or requires unmotivated tools.

♪: denotes an idea that requires a tool seemingly out of thin air.

♪: denotes an idea that follows relatively simple knowing some prior ideas.

♪: denotes an easy idea or proof that can be reproduced with no problem.

$\Omega$ : a state space.

$\mathbf{P}$ : a probability measure.

$\mathbf{E}(X)$ : the integral of some  $X$  with respect to the measure  $\mathbf{P}$ . Also denoted by

$$\int_{\Omega} X \, d\mathbf{P} \quad \text{or} \quad \int_{\Omega} X(\omega) \mathbf{P}(d\omega) \quad \text{or} \quad \int_{\Omega} X(\omega) d\mathbf{P}(\omega)$$

$\mathbf{1}(A)$ : the indicator function of a set  $A$ .

$n \wedge m$ :  $\min(n, m)$ .

$n \vee m$ :  $\max(n, m)$ .

$\mathbf{R}^{\infty}$ : the space of  $\mathbf{R}$ -valued sequences.

$\mathcal{P}(\Omega)$ : the set of probability measures on  $\Omega$ .

$\mathcal{L}(\mu)$ : the law of a probability measure  $\mu$ .

$\mathbf{P}(A | \mathcal{F})$ : the conditional probability of an event  $A$  given a  $\sigma$ -algebra  $\mathcal{F}$ , is defined as  $\mathbf{E}[\mathbf{1}_A | \mathcal{F}]$



# Chapter 1

## Markov Chains

Let us begin by speed-running basic ideas of Markov Chain Theory.

### 1.1 Elementary concepts

**Definition 1.1** Let  $\Omega$  be a finite set,  $\mu$  a distribution on  $S$ , and  $P$  a stochastic matrix of size  $|\Omega| \times |\Omega|$ . A Markov Chain (MC) is a stochastic process  $(X_n : n \geq 0)$  with

1.  $X_0 \sim \mu$ .
2. For  $n \geq 0$ , conditional on  $X_n = i$ ,  $X_{n+1}$  has distribution  $(P(i, j) : j \in \Omega)$  and is independent of  $X_0, \dots, X_{n-1}$ .

Explicitly, these conditions mean

$$\mathbf{P}(X_0 = x_0, X_1 = x_1, \dots, X_t = x_t) = \mu(x_0)P(x_0, x_1)P(x_1, x_2) \cdots P(x_{t-1}, x_t)$$

The fundamental property of a Markov chain is its lack of memory. This is clear from the definition, but a strong version of this statement comes in form of the

**Theorem 1.2 (Markov Property)** Let  $(X_n : n \geq 0)$  be a Markov Chain with initial distribution  $\mu$  and transition matrix  $P$ . Then given any  $t \geq 0$ , we have that conditional on  $X_t = i$ ,  $(X_{n+t} : n \geq 0)$  is a Markov Chain with initial distribution  $\delta_i$ , with transition matrix  $P$ , and it is independent of  $X_0, \dots, X_t$ .

We are also interested in knowing the probability of the Markov Chain being in some state at a given point in time. Noting that our distribution  $\mu$  can simply be thought of as a vector  $(\mu(i) : i \in \Omega)$  gives

a clear way of computing the distribution of the Chain after one unit of time,  $\mu'$ , namely

$$\mu'(i) = \sum_{j \in \Omega} \mu(j)P(j, i) = (\mu P)_i$$

It turns out that there is also a simple way of computing the probability of a given state after  $n$  time units, this is just a simple consequence of the Markov Property:

**Theorem 1.3** Let  $(X_n : n \geq 0)$  be a Markov Chain with initial distribution  $\mu$ . Then

$$\mathbf{P}(X_t = j \mid X_0 = i) \equiv \mathbf{P}_i(X_t = j) = P^n(i, j)$$

*Proof.*

$$\mathbf{P}_i(X_t = j) = \sum_{x_1, \dots, x_{t-1}} P(i, x_1)P(x_1, x_2) \cdots P(x_{t-1}, j) = P^n(i, j)$$



## 1.2 Class Structure

Sometimes we may break down a Markov Chain into smaller pieces, each of which is easier to understand. This is done through the identification of *communicating classes*.

**Definition 1.4** Let  $(X_n : n \geq 0)$  be a Markov Chain. We say a state  $i$  leads to a state  $j$ , written  $i \rightarrow j$  if  $\mathbf{P}_i(X_n = j \text{ for some } n \geq 0) > 0$ . If  $i \rightarrow j$  and  $j \rightarrow i$  we say  $i$  and  $j$  communicate, written as  $i \leftrightarrow j$ .

It is clear that  $\leftrightarrow$  defines an equivalence class on  $\Omega$ , and as such it partitions  $\Omega$  into *communicating classes*.

**Definition 1.5** A communicating class  $C$  is closed if  $i \in C$  and  $i \rightarrow j$  implies  $j \in C$ . A state  $i$  is absorbing if  $\{i\}$  is a closed class. If  $\Omega$  is a single class under the Markov Chain, then we refer to the chain as irreducible.

## 1.3 Transience and recurrence

**Definition 1.6** Given a Markov Chain  $(X_n : n \geq 0)$  and an event  $A$ , we can define the *hitting times*



of  $A$ :

$$T_A = \inf\{t \geq 0 : X_t \in A\} \quad T_A^+ = \inf\{t > 0 : X_t \in A\}$$

**Definition 1.7** A Markov Chain for which  $\mathbf{P}_i(T_i^+ < \infty) = 1$  is called recurrent. Otherwise, it is called transient. A Markov Chain for which  $\mathbf{E}_i[T_i^+] \equiv \mathbf{E}[T_i^+ | X_0 = i] < \infty$  is called positive recurrent.

**Theorem 1.8** An irreducible Markov Chain on a finite state space is positive recurrent.

*Proof.* Appendix.



**Theorem 1.9**  $(X_n)$  is transient if and only if  $\sum_{n \geq 0} (P^n)(i, i) < \infty$ .

See the appendix for some more results on this area.

## 1.4 Invariant Distributions

Recall that a distribution  $\mu$  on the state space  $\Omega$  may be thought of as a vector  $(\mu(i) : i \in \Omega)$ , and as such we had a notion of multiplication by the transition matrix  $P$ . The long-term properties of a Markov Chain are connected with the notion of an invariant distribution.

**Definition 1.10** (Invariant distribution ) A distribution  $\pi$  on  $\Omega$  is said to be invariant if

$$\pi P = \pi$$

For notational expediency, we shall reserve the use of  $\pi$  for an invariant measure. It turns out that invariant measures exist for certain kinds of Markov Chains

**Theorem 1.11** (Existence of  $\pi$ ) Let  $(X_n)$  be an irreducible, positive recurrent Markov Chain, then  $(X_n)$  admits an invariant distribution  $\pi$ , and moreover

$$\pi(x) = \frac{1}{\mathbf{E}_x[T_x^+]}$$

*Proof.* See Appendix.



**Remark 1.12** In this course, only irreducible finite state chains will be considered, therefore every chain we consider will have an invariant distribution.

## 1.5 Time Reversal

For Markov Chains, the past and the future are independent given the present. This symmetry in time suggests looking at a Markov Chain running backwards. However, convergence to equilibrium suggests a behavior that is asymmetrical in time: one may start from a highly organised state such as a point mass, and see it decay into chaos, the invariant distribution. This suggests that if we wish complete time-symmetry we must begin in equilibrium. This is indeed the case, however, the transition matrix will be different.

**Theorem 1.13** Let  $P$  be an irreducible Markov Chain with invariant distribution  $\pi$ . Suppose that  $(X_n : 0 \leq n \leq N)$  is Markov( $\pi, P$ ) and set  $(Y_n : 0 \leq n \leq N)$  by  $Y_n = X_{N-n}$ . Then  $Y_n$  is Markov( $\pi, P^*$ ) where

$$(P^*)(i, j) = P(j, i) \frac{\pi(j)}{\pi(i)}$$

Moreover  $Y_n$  is irreducible with invariant distribution  $\pi$ .

If a Markov Chain has that  $P^* = P$ , i.e:

$$\pi(i)P(i, j) = \pi(j)P(j, i)$$

we say that the Markov Chain is time-reversible. That is to say, if it starts running from the equilibrium distribution, one cannot distinguish between the original chain and the time-reversed one. The equation above is called the Detail Balanced equation. It is a simple result that if a distribution  $\lambda$  on  $\Omega$  is in Detail Balance, then  $\lambda$  is invariant. It turns out that  $P^*$  makes another appearance:

**Proposition 1.14** Let  $f, g : \Omega \rightarrow \mathbf{R}$  be two functions. Then

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

where

$$\langle f, g \rangle_\mu = \sum_{x \in \Omega} \mu(x) f(x) g(x)$$

*Proof.*

$$\begin{aligned}
 \langle Pf, g \rangle_\pi &= \sum_{x \in \Omega} \pi(x) \left( \sum_{y \in \Omega} P(x, y) f(y) \right) g(x) \\
 &= \sum_{x, y \in \Omega} \pi(y) P^*(y, x) f(y) g(x) \\
 &= \sum_{y \in \Omega} \pi(y) \left( \sum_{x \in \Omega} P^*(y, x) g(x) \right) f(y) \\
 &= \langle f, P^* \rangle_\pi
 \end{aligned}$$



## 1.6 The Ergodic Theorem


**Definition 1.15** (Aperiodicity) A Markov Chain  $(X_n : n \geq 0)$  is aperiodic if for all states  $x \in \Omega$ ,  $\gcd(t > 0 : P^t(x, x) > 0) = 1$ . I.e: the times at which the chain can return to any given starting point are co-prime.

**Theorem 1.16** (Aperiodicity is a class property) Let  $(X_n)$  be an irreducible Markov Chain, then if some state  $x_0$  is aperiodic, then all states are aperiodic.

*Proof.* We show in fact something better, that if two states  $x$  and  $y$  communicate, then the periods  $d(x) \equiv \gcd(t > 0 : P^t(x, x) > 0)$  and  $d(y)$  coincide. The claim then follows immediately.

Since  $x$  and  $y$  communicate, there exists integers  $n$  and  $m$  such that  $P^n(x, y) > 0$  and  $P^m(y, x) > 0$ . Then for any  $s \in \{t > 0 : P^t(x, x) > 0\}$

$$P^{n+m}(y, y) \geq P^m(y, x)P^n(x, y) > 0 \quad P^{n+s+m}(y, y) \geq P^m(y, x)P^s(x, x)P^n(x, y) > 0$$

Which means that  $d(y) \mid m + n$  and  $d(y) \mid m + n + s$  hence  $d(y) \mid s$ . Therefore  $d(y) \mid d(x)$ . By symmetric arguments  $d(x) \mid d(y)$  and as such the periods coincide. 

**Definition 1.17** (Ergodicity) A Markov Chain is said to be ergodic if it is irreducible, aperiodic and positive recurrent.

**Theorem 1.18 (The Ergodic Theorem)** Let  $(X_n)$  be an Ergodic Markov Chain with invariant distribution  $\pi$ . Then for all  $i, j \in \Omega$  we have that

$$(P^t)(x, y) \rightarrow \pi(y) \quad t \rightarrow \infty$$

# Chapter 2

## The Total Variation and Coupling

In this section we learn:

- The definition of the Total Variation (TV) and different ways to characterise TV. How the TV turns the space of probability measures into a metric space with a convex-like property where the transition matrix is a non-expansion.
- The definition of a coupling and how couplings provide an upper bound for TV.

The goal of this course is to examine in depth the Ergodic Theorem. In particular, this theorem has the defect that it offers no information about the speed of convergence. To study this, we first need to define a metric on the space of probability measures, in order for us to determine this rate of convergence. This will incidentally prove the Ergodic Theorem. This metric is given by the total variation.

**Definition 2.1** Let  $\Omega$  be a finite space and let  $\mu, \nu$  be two measures on  $\Omega$ . Their total variation distance is defined by

$$d_{\text{TV}}(\mu, \nu) \equiv \|\mu - \nu\|_{\text{TV}} := \sup_{A \subseteq \Omega} |\mu(A) - \nu(A)| = \sup_{A \subseteq \Omega} \mu(A) - \nu(A)$$

**Remark 2.2** The reason why we can remove the absolute value signs is that

$$\mu(A) - \nu(A) = -(\mu(A^c) - \nu(A^c))$$

We wish to now formalise the fact that this is indeed a distance between measures. It is quite clear that this quantity is symmetric, let us show the other two requirements. To do so some results that provide alternative formulations for total variation distance will be useful.

**Lemma 2.3** (Alternatives for TV ♪) Let  $\mu$  and  $\nu$  be two probability measures. Then

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

**Main idea:** One defines the set  $B$  of states where  $\mu$  is at least  $\nu$ , and shows that  $\|\mu - \nu\|_{\text{TV}} = \mu(B) - \nu(B)$ . Then one can expand this in two ways, as a sum over states in  $B$ , or by splitting into an antisymmetric expression involving  $B^c$ . For the supremum one does a standard argument to show an upper bound and one chooses the signum function for the lower bound.

*Proof.* The clever idea is to consider the set  $B = \{x \in \Omega : \mu(x) \geq \nu(x)\}$ , then we see that given any  $A \subseteq \Omega$

$$\begin{aligned} \mu(A) - \nu(A) &= \mu(A \cap B) - \nu(A \cap B) + \underbrace{\mu(A \cap B^c) - \nu(A \cap B^c)}_{\leq 0} \\ &\leq \mu(A \cap B) - \nu(A \cap B) + \underbrace{\mu(A^c \cap B) - \nu(A^c \cap B)}_{\geq 0} \end{aligned}$$

Where the inequality came due to the fact that if for some set  $D$  we have that  $\mu(D) - \nu(D) \leq 0$ , then  $\mu(D^c) - \nu(D^c) \geq 0$ . Moreover,

$$\mu(A) - \nu(A) = \mu(B) - \nu(B) - \underbrace{(\mu(A^c \cap B) - \nu(A^c \cap B))}_{\geq 0},$$

and so  $\mu(A) - \nu(A) \leq \mu(B) - \nu(B)$ . It follows that  $\|\mu - \nu\|_{\text{TV}} = \mu(B) - \nu(B)$ . Everything will now follow relatively simply:

$$\begin{aligned} \|\mu - \nu\|_{\text{TV}} &= \mu(B) - \nu(B) = \sum_{x \in B} \mu(x) - \nu(x) \\ &= \sum_{x \in B} (\mu(x) - \nu(x))^+ + \underbrace{\sum_{x \in B^c} (\mu(x) - \nu(x))^+}_{=0} \\ &= \sum_{x \in \Omega} (\mu(x) - \nu(x))^+ \quad (\text{First Goal}) \\ &= \sum_{x \in \Omega} \mu(x) - \mu(x) \wedge \nu(x) = 1 - \sum_{x \in \Omega} \mu(x) \wedge \nu(x) \quad (\text{Second Goal}) \end{aligned}$$

Returning to the first lines

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

Finally:

$$\sup_{f: S \rightarrow [-1,1]} |\mu(f) - \nu(f)| = \sup_{f: S \rightarrow [-1,1]} \left| \sum_{x \in \Omega} (\mu(x) - \nu(x)) f(x) \right| \leq \sum_{x \in \Omega} |\mu(x) - \nu(x)| = 2 \|\mu - \nu\|_{\text{TV}}$$

And conversely, choosing  $g(x) = \text{sgn}(\mu(x) - \nu(x))$  we have that

$$\sup_f |\mu(f) - \nu(f)| \geq |\mu(g) - \nu(g)| = \left| \sum_{x \in \Omega} (\mu(x) - \nu(x)) g(x) \right| = \sum_{x \in \Omega} |\mu(x) - \nu(x)| = 2 \|\mu - \nu\|_{\text{TV}}$$



**Example 2.4 (Bernoulli distributions)** Let  $\mu \sim \text{Ber}(p)$  and  $\nu \sim \text{Ber}(q)$ . Then using the previous result we have that

$$\|\mu - \nu\|_{\text{TV}} = \frac{1}{2} \sum_{x \in \{0,1\}} |\mu(x) - \nu(x)| = \frac{1}{2} |1 - p - 1 + q| + \frac{1}{2} |q - p| = |p - q|$$

**Lemma 2.5 (Range of Total Variation 🎵)** Let  $\mu$  and  $\nu$  be two probability measures. Then

$$0 \leq \|\mu - \nu\|_{\text{TV}} \leq 1$$

With equality at zero if and only if the two measures are equal, and equality at one if and only if their supports are disjoint.

**Main idea:** The bounds themselves are obvious. To show the "equality at 1 iff" parts, use  $\mathcal{L}^1$  characterisation of TV as well as usual definition of TV.

*Proof.* Since  $\mu$  and  $\nu$  are probability measures, the most they can differ by is one, thus obtaining the upper bound. The lower bound for zero is clear. If  $\|\mu - \nu\|_{\text{TV}} = 0$ , it means that for all  $A \subseteq \Omega$ ,  $|\mu(A) - \nu(A)| \leq 0$ . This obviously implies that  $\mu(A) = \nu(A)$  for all  $A \subseteq \Omega$ . Conversely, if the two

measures agree, it is obvious that the total variation is zero. Suppose now that  $\text{supp}(\mu)$  and  $\text{supp}(\nu)$  are disjoint. Then we can split the sum

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

Conversely, if  $\|\mu - \nu\|_{\text{TV}} = 1$ , then due to finiteness of state space, there exists some set  $A \subseteq \Omega$  with  $\mu(A) = 1$  and  $\nu(A) = 0$ . This means that  $\text{supp}(\mu) \subseteq A$ , and consequently the two supports are disjoint.  $\heartsuit$

The remaining ingredient to show that the total variation distance is actually a metric is the triangle inequality, but this follows trivially from the fact that

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

by using the triangle inequality  $|\mu(x) - \nu(x)| \leq |\mu(x) - \rho(x)| + |\rho(x) - \nu(x)|$ . Thus we can finally say

**Theorem 2.6** The space  $(\mathcal{P}(\Omega), d_{\text{TV}})$  is a metric space.

We can now introduce the idea of couplings:

**Definition 2.7** Let  $\mu, \nu \in \mathcal{P}(\Omega)$ , and  $X, Y$  be two random variables on the same probability space such that  $\mathcal{L}(X) = \mu$  and  $\mathcal{L}(Y) = \nu$ . Then we refer to the random vector  $(X, Y)$  as a coupling. Sometimes we write  $\Pi(\mu, \nu)$  for the set of all couplings of  $\mu, \nu$ .

Coupling is a general technique in probability, but in our case, it will allow us to obtain one more way of bounding the total variation distance.

**Example 2.8 (An example of a coupling of Bernoulli distributions)** Let  $\mu$  and  $\nu$  be Bernoulli  $p$  and  $q$  distributions respectively. Suppose that  $p < q$ . To construct a coupling for  $\mu$  and  $\nu$  we could take a uniform distribution  $U \sim \mathcal{U}[0, 1]$  and have

$$X = \mathbf{1}_{\{U \leq p\}} \quad Y = \mathbf{1}_{\{U \leq q\}}$$

Notice that  $\mathbf{P}(X \neq Y) = q - p$  which is equal to the total variation distance between  $\mu$  and  $\nu$ . Coincidence? I don't think so.

It turns out that the total variation distance is upper bounded by the probability of disagreement of a coupling, and moreover, there always exists a coupling, called the optimal coupling, that attains this lower bound.



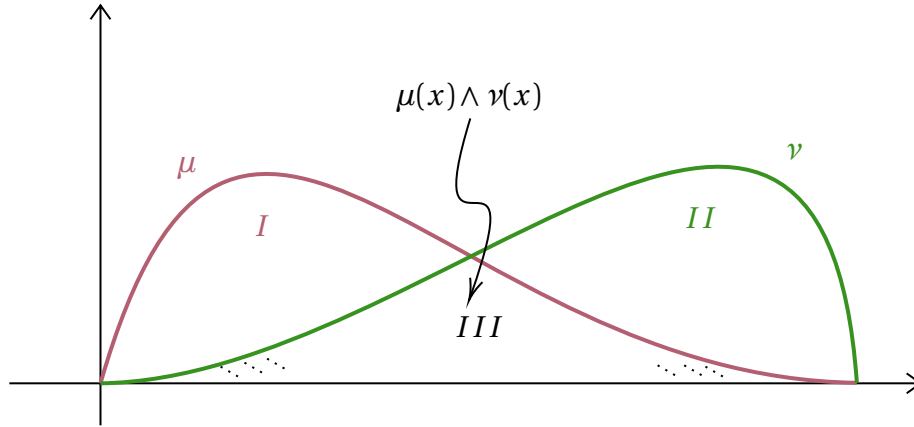


Figure 2.1: The diagram that says it all

**Theorem 2.9** (The coupling inequality ♪) Let  $\mu, \nu \in \mathcal{P}(\Omega)$ . Then

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

Moreover, this infimum is attained.

**Main idea:** For the upper bound, express  $\mu(A)$  as  $\mathbf{P}(X \in A)$  and  $\nu(A)$  in a respective manner, then work towards writing  $\mathbf{P}(X \neq Y)$ . To construct the coupling, look at the diagram that says it all: throw a coin at the diagram, and if it lands in region *III*, then sample  $X$  and  $Y$  to be the same. If the coin lands on the other regions, sample  $X$  and  $Y$  according to the "rescaled" curves in regions *I* and *II*. This construction has that  $X$  and  $Y$  agree when the coin falls on region *III*, which occurs with probability

$$p = \sum_{x \in \Omega} \mu(x) \wedge \nu(x) = 1 - \|\mu - \nu\|_{\text{TV}}$$

thus satisfying our request.

*Proof.* Let  $A \subseteq \Omega$ , and  $(X, Y) \in \Pi(\mu, \nu)$ . Then

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

Thus  $\|\nu - \mu\|_{\text{TV}} \leq \mathbf{P}(X \neq Y)$  for all couplings  $(X, Y)$ . Now we show there is some coupling that attains this bound. First of all, suppose that we can construct a coupling  $(X, Y)$  with

$$\mathbf{P}(\{X = Y\}) = \sum_{x \in \Omega} \mu(x) \wedge \nu(x) =: p$$

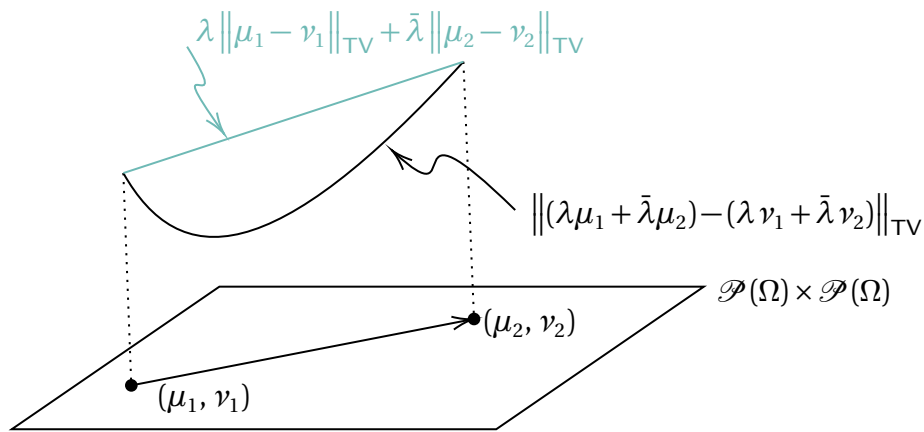


Figure 2.2: The map  $(\mu, \nu) \mapsto \|\mu - \nu\|_{\text{TV}}$  is convex

Then  $\mathbf{P}(\{X \neq Y\}) = \|\mu - \nu\|_{\text{TV}}$  by one of the equivalent characterisations. The coupling is constructed as follows: toss a coin with probability  $p$  of obtaining heads. If a head is obtained, then set  $X = Y = Z$  where  $Z$  is sampled from the distribution

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

If a tails is obtained, then sample

$$X \sim \frac{\mu(x) - \nu(x)}{1-p} \mathbf{1}_{\{\mu(x) > \nu(x)\}} \quad Y \sim \frac{\nu(x) - \mu(x)}{1-p} \mathbf{1}_{\{\mu(x) < \nu(x)\}}$$

In the case a tails is obtained, then the supports of the distributions of  $X$  and  $Y$  are disjoint so they cannot be equal. Therefore the probability that  $X = Y$  is  $p$ . Now we can easily check that  $(X, Y)$  is a coupling by using the law of total probability:

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

And similarly for  $Y$ . ♡

We finish this section with one more property of the total variation distance, which I illustrate in a diagram because I'm a stupid visual learner:

**Proposition 2.10 (Convexity and non-expansion ♪)** The function  $(\mu, \nu) \mapsto \|\mu - \nu\|_{\text{TV}}$  is convex. Moreover, given a transition matrix  $P$ ,  $P$  is a non-expansion, i.e:

$$\|\mu P - \nu P\|_{\text{TV}} \leq \|\mu - \nu\|_{\text{TV}} \quad \forall \mu, \nu \in \mathcal{P}(\Omega)$$

**Main idea:** Convexity is an easy calculation, non-expansivity uses the fact that the set of all functions of the form  $Pf$  where  $f : \Omega \rightarrow [-1, 1]$  is in fact contained in the set of all functions  $f : \Omega \rightarrow [-1, 1]$ , and

so the supremum over the integrals cannot be larger.

*Proof.* Checking convexity is simple: let  $\lambda \in [0, 1]$  and let  $\bar{\lambda}$  denote  $1 - \lambda$

$$\begin{aligned} \|(\lambda\mu_1 + \bar{\lambda}\mu_2) - (\lambda\nu_1 + \bar{\lambda}\nu_2)\|_{\text{TV}} &= \max_{A \subseteq \Omega} [(\lambda\mu_1 + \bar{\lambda}\mu_2)(A) - (\lambda\nu_1 + \bar{\lambda}\nu_2)(A)] \\ &= \max_{A \subseteq \Omega} [\lambda(\mu_1 - \nu_1) + \bar{\lambda}(\mu_2 - \nu_2)] \\ &\leq \max_{A \subseteq \Omega} [\lambda(\mu_1 - \nu_1)] + \max_{A \subseteq \Omega} [\bar{\lambda}(\mu_2 - \nu_2)] \\ &= \lambda \|\mu_1 - \nu_1\|_{\text{TV}} + \bar{\lambda} \|\mu_2 - \nu_2\|_{\text{TV}} \end{aligned}$$

For non-expansivity, we have that

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

Since  $P$  is stochastic, we have that

$$|Pf(x)| = \left| \sum_{y \in \Omega} P(x, y) f(y) \right| \leq 1$$

and as such

$$\{Pf \mid f: S \rightarrow [-1, 1]\} \subseteq \{f: S \rightarrow [-1, 1]\}$$

so

$$\frac{1}{2} \sup_{f: S \rightarrow [-1, 1]} |\mu Pf - \nu Pf| \leq \frac{1}{2} \sup_{f: S \rightarrow [-1, 1]} |\mu f - \nu f| = \|\mu - \nu\|_{\text{TV}}$$



## 2.1 Examples

Here are some examples. I could have included many others, but these are the ones I found hard once:

**Example 2.11** Let  $Y$  be an  $\mathbf{N}$ -valued random variable with the property that  $\mathbf{P}[Y = j] \leq c$  and  $\mathbf{P}[Y = j]$  is decreasing in  $j$ . Let  $Z$  be an independent  $\mathbf{N}$ -valued random variable. Show that

$$\|\mathcal{L}(Y + Z) - \mathcal{L}(Y)\|_{\text{TV}} \leq c \mathbf{E}[Z].$$

*Proof.* This is an interesting example as we have some sort of two sources of randomness to take care of, so it is a good way to see how conditioning can save the day. First let's tackle the case with  $Y + k$  instead of  $Y + Z$ . Note:

$$\|\mathcal{L}(Y + k) - \mathcal{L}(Y)\|_{\text{TV}} = \frac{1}{2} \sum_{n \geq 1} |\mathbf{P}[Y = n - k] - \mathbf{P}[Y = n]| \quad (2.1)$$

$$= \frac{1}{2} \left( \sum_{n=1}^k \mathbf{P}[Y = n] + \sum_{n=k+1}^{\infty} \mathbf{P}[Y = n - k] - \mathbf{P}[Y = n] \right) \quad (2.2)$$

$$= \sum_{n=1}^k \mathbf{P}[Y = n] \leq c k \quad (2.3)$$

Where 2.2 comes from the fact that since  $Y$  is  $\mathbf{N}$ -valued,  $\mathbf{P}[Y \leq 0] = 0$ , and so we just split the sum. Moreover, in this same step we also used the decreasing property to get rid of the absolute value signs on the second sum. Then in 2.3 we used the fact that the second sum telescopes, and is equal to the first sum. Now we are ready to use this to finish the exercise, all we need to do is condition on  $Z$  and be careful about it. For each  $k$ , let  $V_k$  and  $W_k$  be the TV-optimal couplings of the laws of  $Y$  and  $Y + k$ , and let them be independent of  $Z$ . Then

$$\begin{aligned} \|\mathcal{L}(Y + Z) - \mathcal{L}(Y)\|_{\text{TV}} &\leq \mathbf{P}[V_Z \neq W_Z] \\ &= \sum_{n \geq 1} \mathbf{P}[Z = n] \mathbf{P}[V_n \neq W_n] \\ &\leq k \sum_{n \geq 1} n \mathbf{P}[Z = n]. \end{aligned}$$



# Chapter 3

## Mixing Times

**Definition 3.1** Let  $(X_n)$  be a Markov Chain with invariant distribution  $\pi$  and transition matrix  $P$ . The distance to stationarity at a time  $t$  is given by

$$d_P(t) \equiv d(t) := \max_{x \in \Omega} \|P^t(x, \cdot) - \pi(\cdot)\|_{\text{TV}}$$

In other words, the distance to stationarity is the worst TV distance from the the distribution of the Markov Chain at the  $t^{\text{th}}$  time, initiated at any starting point to its invariant distribution. We also write

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

For the worst two starting point distance. We have some preliminary facts:

**Proposition 3.2** (Properties of distance to stationarity ♪) Let  $(X_n)$  be as above, then  $d(t)$  and  $\bar{d}(t)$  are non-increasing. Moreover, given any distribution  $\nu$ , we have that  $d(t) \geq \|\nu P^t - \pi\|_{\text{TV}}$ .

**Main idea:** For non-increasingness, we use the fact that  $P$  is a non-expansion of TV. For the second inequality, we express  $\nu$  as a convex combination of point masses.

*Proof.* By Proposition 2.10, we have that if  $t_1 \leq t_2$ , then

$$d(t_1) = \max_{x \in \Omega} \|P^{t_1}(x, \cdot) - \pi(\cdot)\|_{\text{TV}} \geq \max_{x \in \Omega} \left\| P^{t_1}(x, \cdot) P - \underbrace{\pi P}_{\pi} \right\|_{\text{TV}} \geq \dots \geq d(t_2)$$

To show the second part, we recall again from Proposition 2.10 that the map  $(\mu, \nu) \mapsto \|\mu - \nu\|_{\text{TV}}$  is convex. In particular, for a collection  $(p_i : i \in I)$  of real numbers such that  $\sum_i p_i = 1$ , and a

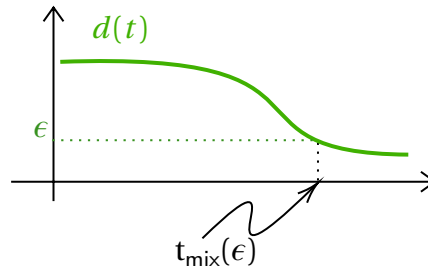


Figure 3.1: Mixing time

collection  $(\mu_i : i \in I)$  of distributions, we have that

$$\left\| \left( \sum_{i \in I} a_i \mu_i \right) - \nu \right\|_{\text{TV}} \leq \sum_{i \in I} a_i \|\mu_i - \nu\|_{\text{TV}}$$

Therefore,

$$\begin{aligned} \left\| \left( \sum_{y \in \Omega} \nu(y) \delta_y(y) P^t(y, \cdot) \right) - \pi \right\|_{\text{TV}} &\leq \sum_{y \in \Omega} \nu(y) \|P^t(y, \cdot) - \pi(\cdot)\|_{\text{TV}} \\ &\leq \max_{x \in \Omega} \|P^t(x, \cdot) - \pi(\cdot)\|_{\text{TV}} \sum_{y \in \Omega} \nu(y) = \max_{x \in \Omega} \|P^t(x, \cdot) - \pi(\cdot)\|_{\text{TV}}. \end{aligned}$$

♡

We are now ready to define a mixing time:

**Definition 3.3** Let  $\epsilon > 0$  be given, the  $\epsilon$ -mixing time is given by

$$t_{\text{mix}}(\epsilon) = \inf\{t \geq 0 : d(t) \leq \epsilon\}$$

i.e: the least time for which the distance to stationarity is no more than  $\epsilon$ . Thanks to Proposition 3.2 we know this definition is not nonsense, because if we reach the mixing time, then waiting any further will not make us "unmix". Obscure historical reasons made it so that when written by itself,  $t_{\text{mix}}$  denotes  $t_{\text{mix}}(0.25)$ .

We now move on by using this framework to prove the Ergodic Theorem, and in fact we will do it by showing something stronger, which actually gives us an understanding of the nature of the decay of the distance to equilibrium.

**Theorem 3.4** (Geometric decay of distance to equilibrium, 🍷) Let  $(X_n)$  be an irreducible, aperiodic Markov Chain on a finite state space with invariant distribution  $\pi$ . There exists some  $\alpha \in (0, 1)$  and

positive constant  $C$  so that for all  $t$  we have

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

**Main idea:** By aperiodicity and irreducibility, there is some  $r \geq 0$  such that  $P^r$  has positive entries. The final goal now is to express  $P^t$  as a convex combination of  $\Pi$ , a matrix whose rows are all  $\pi$ , and some other matrix. Then we use the convexity of the "TV map". A good place to start is by expressing  $P^r$  as such a convex combination, then extending for  $P^{rk}$  and finally for  $P^{rk+l}$ . To see what a good choice of the "convexity parameter" would be, note that we want to write

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

and since we want  $Q$  to be non-negative, we need an  $\alpha$  such that

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

We also need our  $\alpha$  to be at most 1.

*Proof.* We begin by noting that since the chain is irreducible and aperiodic, by Theorem D.7 we have the existence of some  $r \geq 0$  so that  $P^r$  has positive entries. From this we can define

$$\alpha = \min_{x, y \in \Omega} \frac{P^r(x, y)}{\pi(y)}$$

Which is clearly positive, and moreover, since for any  $x \in \Omega$

$$\sum_y P^r(x, y) = 1 = \sum_y \pi(y)$$

we must have that for some  $y^*$ ,  $P^r(x, y^*) \leq \pi(y^*)$ , indeed, if all  $P^r(x, y)$  were strictly greater than  $\pi(y)$  the sums over  $y$  would not coincide. Therefore we see that

$$\alpha \leq \frac{P^r(x, y^*)}{\pi(y^*)} \leq 1$$

Now we have concluded that  $\alpha \in (0, 1]$ . If  $\alpha = 1$ , then this means that given any  $x, y \in \Omega$ , then

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

But due to the requirement that they both sum to one, it must be that we have strict equality and hence  $d(r) = 0$  so the inequality holds trivially. Suppose otherwise that  $\alpha < 1$ . Then since for

any pair it holds that  $P^r(x, y) \geq \alpha\pi(y)$ , we may write

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

Where  $Q(x, y)$  is trivially checked to be a stochastic matrix. The interpretation of this is that to sample from  $P^r$ , with probability  $\alpha$  you sample from  $\pi$  and with probability  $(1 - \alpha)$  you sample from  $Q(x, y)$ . For reasons that will become apparent now, it is better to construct a matrix  $\Pi$  whose rows are the row vector  $\pi$ , and simply write the matrix equation

$$P^r = \alpha\Pi + (1 - \alpha)Q$$

Now we claim that  $P^{rk} = (1 - (1 - \alpha)^k)\Pi + (1 - \alpha)^k Q^k$ . The argument goes by induction. The base case we have just shown, so suppose it holds for  $k = n$ , then

$$\begin{aligned} P^{r(k+1)} &= P^{rk} P^r \\ &= ((1 - (1 - \alpha)^k)\Pi + (1 - \alpha)^k Q^k) P^r \\ &= (1 - (1 - \alpha)^k)\Pi P^r + (1 - \alpha)^k Q^k P^r \\ &\stackrel{(1)}{=} (1 - (1 - \alpha)^k)\Pi + (1 - \alpha)^k (\alpha Q^k \Pi + (1 - \alpha) Q^{k+1}) \\ &\stackrel{(2)}{=} (1 - (1 - \alpha)^{k+1})\Pi + (1 - \alpha)^{k+1} Q^{k+1} \end{aligned}$$

Where step (1) came from the fact that  $\Pi P^r = \Pi$  by expressing this matrix multiplication as a sum and then using the fact that  $\pi = \pi P^r$ . Step (2) comes using Lemma D.8 and a bit of algebra.

Now we may multiply by any  $P^l$  and as such

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

Finally, since any  $t \geq 0$  may be expressed as some  $rk + l$ , we have that:

$$\begin{aligned} \|P^t(x, \cdot) - \pi(\cdot)\|_{TV} &= \|\{(1 - (1 - \alpha)^k)\Pi(x, \cdot) + (1 - \alpha)^k (Q^k P^l)(x, \cdot)\} - \pi(\cdot)\|_{TV} \\ &\stackrel{(!)}{\leq} (1 - (1 - \alpha)^k) \|\Pi(x, \cdot) - \pi(\cdot)\|_{TV} + (1 - \alpha)^k \|(Q^k P^l)(x, \cdot) - \pi(\cdot)\|_{TV} \\ &= (1 - \alpha)^k \|Q^k P^l - \pi\|_{TV} \\ &\leq (1 - \alpha)^k \end{aligned}$$

Where step (!) comes from the convexity of the map  $(\mu, \nu) \mapsto \|\mu - \nu\|_{TV}$  and where the last inequality comes from the fact that TV is bounded above by 1. Now to finish off, since  $l \leq r$ , we see that

$$(1 - \alpha)^k = (1 - \alpha)^{\frac{t-l}{r}} \leq (1 - \alpha)^{\frac{t-r}{r}} = \frac{1}{1 - \alpha} (1 - \alpha)^{\frac{t}{r}}$$



so all in all

$$\|P^t(x, \cdot) - \pi(\cdot)\|_{TV} \leq C a^t$$

where  $C = (1 - \alpha)^{-1}$  and  $a = (1 - \alpha)^{-r}$

♡

We now set off to compute bounds for mixing times. To do so we will need some useful inequalities relating distances to stationarity

**Lemma 3.5 (Comparison of  $d$  and  $\bar{d}$ )** For all  $t$  we have that

$$d(t) \leq \bar{d}(t) \leq 2d(t)$$

**Main idea:** The upper bound follows from the triangle inequality. The lower bound follows after noting that if you want to extract a  $P^t(y, A)$  (for the  $\bar{d}$  term) out of thin air, you may do so by writing  $\pi(A) = \sum_y \pi(y) P^t(y, A)$ .

*Proof.* The inequality  $\bar{d}(t) \leq 2d(t)$  comes from the triangle inequality. Indeed, fix  $x$  and  $y$ , then

$$\|P^t(x, \cdot) - P^t(y, \cdot)\|_{TV} \leq \|P^t(x, \cdot) - \pi(\cdot)\|_{TV} + \|P^t(y, \cdot) - \pi(\cdot)\|_{TV}$$

Now maximise both sides of the inequality with respect to  $x$  and  $y$ . For the other inequality, we note that

$$\begin{aligned} d(t) &= \max_x \max_A |P^t(x, A) - \pi(A)| \\ &\stackrel{(!)}{=} \max_x \max_A \left| \sum_{y \in \Omega} \pi(y) P^t(x, A) - \pi(y) P^t(y, A) \right| \\ &\leq \max_x \max_A \max_y |P^t(x, A) - P^t(y, A)| \sum_y \pi(y) \\ &= \max_x \max_y \|P^t(x, \cdot) - P^t(y, \cdot)\|_{TV} \\ &= \bar{d}(t) \end{aligned}$$

Where the only non-trivial step was step (!), where we used  $\pi P^t = \pi$ .

♡

**Lemma 3.6 (Submultiplicativity of  $d(t)$ )** Let  $s, t \geq 0$ . Then  $\bar{d}(s + t) \leq \bar{d}(s) \bar{d}(t)$ . We also have that  $d(s + t) \leq d(s) \bar{d}(t)$ .

**Main idea:** The heart of the proof lies in taking a TV-optimal coupling  $(X, Y)$  of  $P^s(x, \cdot)$  and  $P^s(y, \cdot)$ , and then noticing that

$$P^{t+s}(x, z) = \mathbf{E}[P^t(X, z)] \quad P^{t+s}(y, z) = \mathbf{E}[P^t(Y, z)]$$

with this one can cleverly push towards the goal by noting that  $\mathbf{E}[\mathbf{1}(X \neq Y)] = \|P^s(x, \cdot) - P^s(y, \cdot)\|_{\text{TV}}$ .

*Proof.* Fix  $x, y \in \Omega$ . From Theorem 2.9 we know that there exists a coupling  $(X, Y)$  such that  $X \sim P^s(x, \cdot)$  and  $Y \sim P^s(y, \cdot)$  and

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

Notice that by the Markov Property

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

Similarly  $P^{t+s}(y, z) = \mathbf{E}[P^t(Y, z)]$  Therefore, using one of the characterisations of total variation, we see that

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

Where step (!) comes from the fact that almost surely

$$\frac{1}{2} \sum_z |P^t(X, z) - P^t(Y, z)| = \|P^t(X, \cdot) - P^t(Y, \cdot)\|_{\text{TV}} \leq \max_{x, y} \|P^t(x, \cdot) - P^t(y, \cdot)\|_{\text{TV}} = \bar{d}(t)$$

Reading off the inequality:

$$\|P^{s+t}(x, \cdot) - P^{s+t}(y, \cdot)\|_{\text{TV}} \leq \|P^s(x, \cdot) - P^s(y, \cdot)\|_{\text{TV}} \bar{d}(t)$$

Hence maximising on both sides over  $x$  and  $y$  gives the result that  $\bar{d}(t+s) \leq \bar{d}(t) \bar{d}(s)$ . To show the rest of the claim, we effectively replicate the argument but choose a different coupling.

Let  $(X, Y)$  be an optimal coupling for  $P^s(x, \cdot)$  and  $\pi(\cdot)$ , we start by making the following preliminary observation:

$$\mathbf{E}[P^t(Y, z)] = \sum_y P^t(y, z) \pi(y) = \pi(z)$$

Therefore we may repeat the argument above now:

$$\begin{aligned}
 \|P^{s+t}(x, \cdot) - \pi(\cdot)\|_{\text{TV}} &= \frac{1}{2} \sum_{z \in \Omega} |\mathbf{E}[P^t(X, z)] - \pi(z)| \\
 &= \frac{1}{2} \sum_{z \in \Omega} |\mathbf{E}[P^t(X, z)] - \mathbf{E}[P^t(Y, z)]| \\
 &\leq \frac{1}{2} \mathbf{E} \left[ \mathbf{1}(X \neq Y) \sum_z |P^t(X, z) - P^t(Y, z)| \right] \\
 &\leq \mathbf{E}[\mathbf{1}(X \neq Y) \bar{d}(t)] = \mathbf{P}(X \neq Y) \bar{d}(t) = \|P^s(x, \cdot) - \pi(\cdot)\|_{\text{TV}} \bar{d}(t)
 \end{aligned}$$

Now maximising over  $x$  gives the desired claim. ♡

**Remark 3.7** As immediate consequences we can note that

- $d(s+t) \leq 2d(s)d(t)$
- $d(kt) \leq \bar{d}(t)^k$

Indeed:  $d(s+t) \leq d(s)\bar{d}(t) \leq 2d(s)d(t)$  and  $d(kt) \leq d(t(k-1))\bar{d}(t)$  and now we iterate by using the fact that  $d(s) \leq \bar{d}(s)$ .

**Proposition 3.8** (Upper bound on  $\epsilon$ -mixing time in terms of mixing time) For an Ergodic Markov Chain we have

$$t_{\text{mix}}(\epsilon) \leq \lceil -\log_2(\epsilon) \rceil t_{\text{mix}}$$

**Main idea:** Follows from direct computation using the fact that we know how to deal with the distance to stationarity at multiples of a certain time by using

$$d(kt) \leq \bar{d}(t)^k \leq (2d(t))^k$$

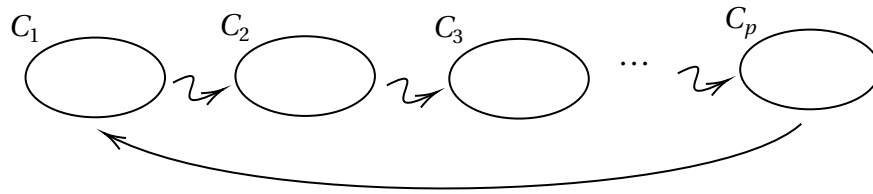
*Proof.* We proceed by showing that at time  $\lceil -\log_2(\epsilon) \rceil t_{\text{mix}}$ , the distance to stationarity falls below  $\epsilon$ . This is a straightforward computation:

$$\begin{aligned}
 d(\lceil -\log_2(\epsilon) \rceil t_{\text{mix}}) &\leq (2d(t_{\text{mix}}))^{\lceil -\log_2(\epsilon) \rceil} \\
 &\leq (2^{-1})^{\lceil -\log_2(\epsilon) \rceil} \\
 &\leq (2^{-1})^{-\log_2(\epsilon)} \leq \epsilon.
 \end{aligned}$$



**Remark 3.9** (Of when the mixing time could be  $\infty$ ) Throughout all these discussions we have made among other assumptions that the chain is aperiodic. We now show how, if the chain is periodic, it could be the case that the mixing time is infinite.

*Proof.* Suppose  $(X_t)$  is an irreducible and periodic Markov chain on a state space  $\Omega$ . Suppose that for some state  $x \in \Omega$  we have that the period  $\tau_x = \gcd\{t > 0 : P^t(x, x)\} = p > 1$ . Then by Background result [TRACK THIS RESULT], since the chain is irreducible, all states are periodic with period  $p$ , which means by Background result [TRACK THIS RESULT] that the state space  $\Omega$  can be partitioned into disjoint subsets  $C_1, \dots, C_p$  such that whenever you are in  $C_k$ , on the next move you must go to  $C_{k+1}$  (addition taken modulo  $p$ ).



By assumption the period is at least two, so there will be at least two partitioning states. Therefore at least one of this must have  $\pi$ -measure at most  $1/2$ , say this collection of states is  $C_k$ . This means that if we start at some state say  $x \in C_k$  in this particular collection of states, then at times multiple of  $k$ , we will be with probability 1 back in  $C_k$ . Therefore, the distance to stationarity at times multiple of the period is bounded below by:

$$d(pl) \geq |\mathbf{P}_x[X_{pl} \in C_k] - \pi(C_k)| = 1 - \pi(C_k) \geq \frac{1}{2}.$$

And since distance to stationarity is non-increasing, i.e: for any time  $t$  there will be a larger multiple of  $p$ , say  $pl'$  so that  $d(t) \geq d(pl')$ , it follows that for all  $0 < \epsilon < \frac{1}{2}$ , the distance to stationarity will never fall below  $\epsilon$ , so the  $\epsilon$ -mixing-time is  $\infty$ .



### 3.1 Examples

We now see a bunch of examples of how to bound TV-distances using the results of this section. To upper bound mixing times we have seen already a technique, namely the coupling inequality, but lower bounding mixing times is something quite harder.

**Example 3.10 (Random Card Transposition)** Let  $\sigma \in S_n$  be an arrangement of cards. Consider the following shuffling method. Pick two cards  $L_t$  and  $R_t$  uniformly at random, and swap them (even if you picked the same card both for  $L_t$  and  $R_t$ , i.e: do nothing). Then

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

**Main idea:** We will see some more techniques of lower bounding mixing times, but for now the only thing we can do is, as I say, *to get dirty*. This means that we need to work with the direct definition of the Total Variation distance. If we want to lower bound it for some time  $t$ , and more importantly, if we want a decent lower bound, we will need to find a set  $A$  where the  $\pi$ -measure and the  $P^t(x, \cdot)$  measure disagree quite a lot, i.e: either  $\pi(A)$  is large but the chain has a low probability of being there by time  $t$ , or conversely,  $A$  is  $\pi$ -small, but the chain has a large probability of being there by time  $t$ . This is what we will do now.

*Proof.* For a permutation  $\sigma \in S_n$ , let  $F(\sigma)$  be the number of fixed points of  $\sigma$ . The idea is that for a reasonably large number  $m$ , sets of the form  $\{\sigma : F(\sigma) > m\}$  have a very small  $\pi$ -mass, whereas if we start our random walk from the identity,  $\sigma_{\text{id}}$ , we start with  $n$  fixed points, and it probably take a long-ish time until we get rid of enough fixed points, which means that  $\sigma_t$  will be in  $A$  with high probability for a reasonably large time. We now make these intuitions precise and quantifiable.

First of all note that  $F(\sigma_t) \geq Z_{2t}$  where  $Z_t$  is the number of unselected coupons in a coupon-collector problem of size  $n$ . Indeed: all the coupons that haven't been touched give rise to a fixed point, but even if a coupon has been touched, it could have been sent back to its original place, contributing still to a fixed point. This motivates us to study some properties of  $Z_{2t}$ . Here we have to claims:

1.  $\mathbf{E}[Z_{2t}] = n(1 - 1/n)^{2t} =: \mu$ .

This is reasonably easy to show, if we let  $I_j(t)$  denote the indicator function that the  $j^{\text{th}}$  coupon has not been touched by time  $t$ , then it is clear that  $\mathbf{E}[I_j(2t)] = (1 - 1/n)^{2t}$ , and moreover it's clear to see that

$$Z_{2t} = \sum_{i=1}^n I_i(2t)$$

this finishes the first claim.

2.  $\text{Var}[Z_{2t}] \leq \mu$ . Showing this is a bit trickier, but it amounts to the following, of course:

$$\text{Var}(Z_{2t}) = \text{Var}\left(\sum_{i=1}^n I_i(2t)\right) = \sum_i \text{Var}(I_i(2t)) + \sum_{i \neq j} \text{Cov}(I_i(2t), I_j(2t))$$

The sum of variances can be immediately seen as  $\mu(1 - \mu/n)$ , indeed:

$$\text{Var}[I_i(2t)] = \mathbf{E}[I_i(2t)^2] - \mathbf{E}[I_i(2t)]^2$$

And the covariance, a bit trickier, is negative, indeed:

$$\text{Cov}(I_i(t), I_j(t)) = \mathbf{E}[I_i(t)I_j(t)] - \mathbf{E}[I_i(t)]\mathbf{E}[I_j(t)] = \left(1 - \frac{2}{n}\right)^t - \left(1 - \frac{1}{n}\right)^{2t} \leq 0$$

Therefore  $\text{Var}(Z_{2t}) \leq \mu$ .

We need one more technicality: namely that  $\mathbf{E}[F(\sigma)] = 1$  for  $\sigma$  picked uniformly at random. This is not too hard to see:

$$\mathbf{E}[F(\sigma)] = \mathbf{E}\left[\sum_{i=1}^n \mathbf{1}(\sigma(i) = i)\right] = \sum_{i=1}^n \frac{(n-1)!}{n!} = 1$$

Now we are ready. As hinted in the intuition section, we are going to formulate a set of permutations with more than a given amount of fixed points, we will choose the following:

$$A = \{\sigma : F(\sigma) \geq \mu/2\}$$

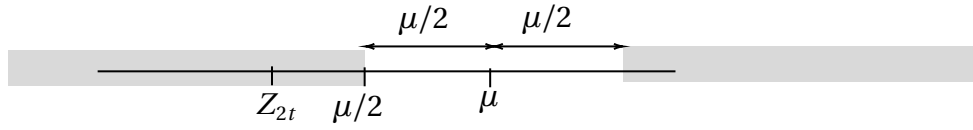
And so by Markov, and using the fact that  $\pi$  is uniform on  $S_n$ , we have that

$$\pi(A) \leq \frac{\mathbf{E}_\pi[F(\sigma)]}{\mu/2} = \frac{2}{\mu}$$

And so we see, that for  $n$  large enough,  $\mu$  will be large and the mass of this set will be small. Now we see how the probability of the walk being in  $A$  contrasts:

$$\begin{aligned} P^t(\sigma_{\text{id}}, A^c) &\stackrel{(1)}{\leq} \mathbf{P}[Z_{2t} \leq \mu/2] \\ &\stackrel{(2)}{\leq} \mathbf{P}[|Z_{2t} - \mu| \geq \mu/2] \\ &\stackrel{(3)}{\leq} \frac{\mu}{(\mu/2)^2} = \frac{4}{\mu} \end{aligned}$$

Where (1) comes from the fact that if  $\sigma_t$  has less than  $m$  fixed points, then there must be less than  $m$  untouched cards (this is the contrapositive of what we said at the start of the proof). Then step (2) comes from the following diagram:



And step (3) comes from Chebyshev's Inequality:  $\mathbf{P}[|X - \mu| \geq a] \leq \text{Var}(X)/a^2$  as well as the upper bound we found on the variance. Combining this all we have that

$$\pi(A) - P^t(\sigma_{id}, A) \geq 1 - \frac{6}{\mu}$$

and by looking at the definition of  $\mu$ , we can see how finding a large enough  $t$  that looks what is in the claim can render this difference lower bounded by  $\epsilon$ , and so obtaining the bound on the mixing time that we wanted. ♡

**Example 3.11 (Lower bounds and hitting times)** Let  $X$  be a Markov chain on a state space  $\Omega$  and denote by  $\tau_A$  the hitting time of a set  $A$ , i.e.:  $\tau_A = \inf\{t \geq 0 : X_t \in A\}$ . Then as we will see now:  $t_{\text{mix}} \geq c \max_{x \in \Omega} \mathbf{E}_x[\tau_A]$  for  $A$  such that  $\pi(A) \geq 1/8$ .

*Proof.* The key here is that we can relate  $\mathbf{E}_x[\tau_A]$  to any time  $t$  by using the Markov property:

$$\mathbf{E}_x[\tau_A] \leq t + \sum_{y \in A^c} P^t(x, y) \mathbf{E}_y[\tau_A]$$

Indeed:

$$\begin{aligned} \mathbf{E}_x[\tau_A] &= \mathbf{E}_x[\tau_A \mathbf{1}\{X_t \in A\}] + \mathbf{E}_x[\tau_A \mathbf{1}\{X_t \in A^c\}] \\ &\leq t + \sum_{y \in A^c} \mathbf{P}_x[X_t = y] \mathbf{E}_x[\tau_A \mathbf{1}\{X_t \in A^c\} | X_t = y] \\ &= t + \sum_{y \in A^c} P^t(x, y) \mathbf{E}_y[\tau_A] \end{aligned}$$

where we used the Markov property in the last line. Then if we let  $x' = \arg\max_{x \in \Omega} \mathbf{E}_x[\tau_A]$ , we see that for any  $t$ ,

$$\mathbf{E}_{x'}[\tau_A] \leq t + \mathbf{E}_{x'}[\tau_A] P^t(x', A^c).$$

Now we note that if  $t := t_{\text{mix}}(1/16) = 4t_{\text{mix}}(1/4)$ , we have that  $1/16 = d(t) \geq 1/8 - P^t(x', A)$ , and so  $P^t(x', A) \geq 1/16$  which means  $P^t(x', A^c) \leq 15/16$ , and so plugging into the above expression we

obtain that  $\frac{1}{16}\mathbf{E}_{x'}[\tau_A] \leq t_{\text{mix}}(1/16)$  which finally implies that

$$\frac{1}{64}\mathbf{E}_{x'}[\tau_A] \leq t_{\text{mix}}.$$





# Chapter 4

## Markovian Couplings

**Definition 4.1 (Notation)** Let  $f$  and  $g$  be two functions  $\mathbf{N} \rightarrow \mathbf{R}$ . Then we write

- $f \lesssim g$  if there is a constant  $C \geq 0$  such that  $f(n) \leq Cg(n)$  for all  $n \in \mathbf{N}$ . This is also usually written as  $f = O(g)$
- $f \asymp$  if  $f \lesssim g$  and  $g \lesssim f$ .
- $f \ll g$  if  $f(n)/g(n) \rightarrow 0$  as  $n \rightarrow \infty$ , also written sometimes as  $f = o(g)$ .

Recall that a coupling of distributions  $\mu$  and  $\nu$  is a pair  $(X, Y)$  of random variables on the same probability space, such that the marginal distribution of  $X$  is  $\mu$  and the marginal distribution of  $Y$  is  $\nu$ . We also saw how  $\|\mu - \nu\|_{\text{TV}}$  is characterised as the minimum over all couplings  $(X, Y)$  of  $\mu, \nu$ , of the probability that  $X$  and  $Y$  disagree, which provides an effective method of obtaining upper bounds on the distance. In this chapter we will extract more information by not just coupling distributions, but entire Markov chains. In particular, we will see how building two simultaneous copies of a Markov chain using a common source of randomness can be useful for getting bounds on the distance to stationarity.

**Definition 4.2 (Coupling of Markov Chains)** A coupling of Markov chains with transition matrix  $P$  is a process  $(X_t, Y_t)_{t=0}^{\infty}$  with the property that both  $(X_t)$  and  $(Y_t)$  are Markov chains with transition matrix  $P$ , although with possibly different starting distributions.

**Example 4.3 (Initial example)** Here is a “dumb example” of this, that showcases the power of coupling chains to obtain information about their distribution. Suppose that  $(X_t)$  is a simple random walk on the interval  $\{1, \dots, n\}$ , i.e: moves up and down with equal probability, and if the chain is at say  $n$  and attempts to move upwards, it stays put (similarly if the chain is at 1 and it tries to move downward). Then we have the “obvious” fact that if  $x \leq y$ , then  $P^t(x, n) \leq P^t(y, n)$ .

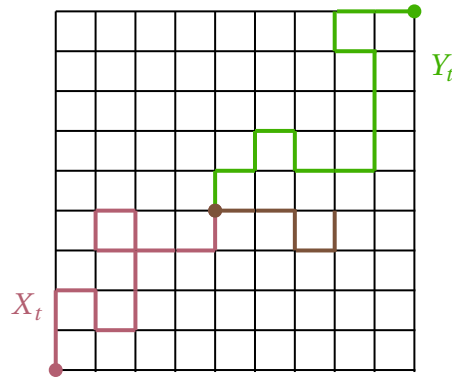


Figure 4.1: A coalescent coupling of two Random Walks  $X_t$  and  $Y_t$  on the grid.

This is “obvious” because if the chain starts at a lower initial position, then the probability that it reaches  $n$  in  $t$  steps cannot be greater than the probability when it starts at a higher initial configuration.

*Proof.* To prove this claim we will couple two chains  $(X_t)$  and  $(Y_t)$  on  $\{1, \dots, n\}$ . For this define a probability space on which there is defined a sequence  $(\Delta_i)$  of i.i.d random variables that take the value  $\pm 1$  with equal probability. Then it is clear that if  $x \leq y$  are two given starting positions, the chains  $(X_t)$  and  $(Y_t)$  that are started at  $x$  and  $y$  respectively and use the  $(\Delta_i)$ 's to perform the moves have the property that  $(X_t)$  and  $(Y_t)$  have transition probabilities  $P^t(x, \cdot)$  and  $P^t(y, \cdot)$  respectively and that one always has  $X \leq Y$ . Therefore, if  $X$  is at  $n$ , then so is  $Y$ , and so

$$P^t(x, n) = \mathbf{P}[X = n] \leq \mathbf{P}[Y = n] = P^t(y, n).$$

This “stupid” argument showcases the power of couplings.



**Definition 4.4** (Markovian coupling) A Markovian coupling is a coupling of Markov chains  $(X_t, Y_t)_{t \geq 0}$  which is itself a Markov chain on  $\Omega \times \Omega$  and

$$\mathbf{P}(X_1 = x \mid X_0 = x_0, Y_0 = y_0) = P(x_0, x) \quad \mathbf{P}(Y_1 = y \mid X_0 = x_0, Y_0 = y_0) = P(y_0, y)$$

**Remark 4.5** Any Markovian coupling with transition matrix  $P$  can be modified so that the two chains stay together after their first meeting, so that if  $X_s = Y_s$ , then  $X_t = Y_t$  for all  $t \geq s$ . For this to be possible without altering the law of the process, the condition of Markovian coupling is crucial, indeed, if we didn't have it, we would not know whether the transition probabilities of  $X_t$  and  $Y_t$  given that  $(X_{t-1}, Y_{t-1}) = (z, z)$  are the same. Of course, if we have a coupling of Markov chains that run independently before they meet, and then we run them together after they meet, we can still produce this property, since this independence effectively gives us what we need from

the Markovian coupling property.

**Example 4.6** (Example of a coupling of Markov chains that is not a Markovian coupling) To further understand the subtleties of these definitions, let us see an example of a coupling of Markov chains  $(X_t, Y_t)$  that is not a Markovian coupling.

Let  $Y_t$  be a Markov chain on the state space  $\{0, 1\}$  with transition probability  $P(x, y) = 1/2$  for all  $x, y$ , i.e: a memoryless Bernoulli binary string. Now take  $X_{t+1} = Y_t$ , and say  $X_0 = 0$ , then  $X_t$  is also a Markov chain with the same transition probabilities as  $Y_t$ , but the process  $(X_t, Y_t)$  is not a Markovian coupling because  $X_t$  is influenced by the values of  $Y_t$ . To be more precise, it is not at all the case that for any  $x, x_0, y_0$

$$\mathbf{P}(X_1 = x \mid X_0 = x_0, Y_0 = y_0) = 1/2$$

**Theorem 4.7** (Bound on TV via coalescent couplings, ♪) Let  $\{(X_t, Y_t)\}$  be coalescent Markov chain coupling, for which  $X_0 = x$  and  $Y_0 = y$ . Define

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

Then

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

Where  $\mathbf{P}_{x,y}(A) = \mathbf{P}(A \mid X_0 = x, Y_0 = y)$ .

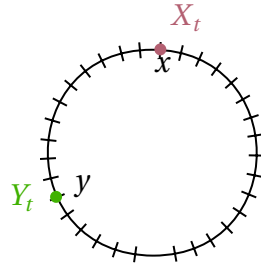
**Main idea:** We note that a coalescent Markov chain coupling started at  $x$  and  $y$  respectively is also a coupling for  $P^t(x, \cdot)$ ,  $P^t(y, \cdot)$  so we can use the coupling inequality. We then note that since the coupling is assumed to be coalescent,  $X_t$  and  $Y_t$  being different is equivalent to not having reached coalescence time by time  $t$ .

*Proof.* Since  $P^t(x, z) = \mathbf{P}_{x,y}(X_t = z)$  and  $P^t(y, z) = \mathbf{P}_{x,y}(Y_t = z)$  we have that  $(X_t, Y_t)$  is a coupling of  $P^t(x, \cdot)$  and  $P^t(y, \cdot)$ , therefore

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

Now we simply note that the events  $\{X_t \neq Y_t\}$  and  $\{\tau_{\text{couple}} > t\}$  are the same events due to assumption of the coupling being coalescent. ♡

Combining this previous fact with the bound obtained in Lemma 3.5 that tells us that  $d(t) \leq \bar{d}(t)$  and maximising over  $x$  and  $y$  gives that:

Figure 4.2: A coupling on  $\mathbf{Z}_n$ 

**Corollary 4.8** Suppose that for each pair of states  $x$  and  $y$  there is a Markov chain coupling  $(X_t, Y_t)$  with  $X_0 = x$  and  $Y_0 = y$ . Then for each such coupling, let  $\tau_{\text{couple}}$  be the first time the chains meet. Then

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

Why do we care about this Corollary? If our goal is to obtain upper bounds on the Mixing Time, since  $t_{\text{mix}} = \inf\{t \geq 0 : d(t) \leq 1/4\}$  we see that if we find some  $t^*$  such that  $d(t^*) \leq 1/4$ , then  $t_{\text{mix}} \leq t^*$ . If we manage to construct coalescent couplings for which  $\tau_{\text{couple}}$  is small, i.e: they coalesce rapidly, then we will obtain a good upper bound.

## 4.1 Examples

We now get dirty to see the power of this technique:

**Example 4.9** (Mixing time on lazy random walk on  $\mathbf{Z}_n$ ) Define a Lazy Simple Random Walk on the cycle  $\mathbf{Z}_n$  by setting the transition matrix to be  $(P + I)/2$  where only non-zero transitions in  $P$  are  $P(i, (i \pm 1) \bmod n) = 1/2$ . Then the mixing time of the Lazy Simple Random Walk on  $\mathbf{Z}_n$  satisfies

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

**Main idea:** The following idea is of great importance: we wish to construct a coupling on  $\mathbf{Z}_n$ , but we must do so in a way that we (almost surely) guarantee coalescence at some point, if we just ran the two chains as if they were, we might have the possibility that they jump over each other, so to fix this, we employ the following common technique of coupling lazy chains (this is why this example works with lazy chains): toss a fair coin and based on the outcome, only let one of the two chains move, this respects the laziness and prevents the chains from jumping over each other.

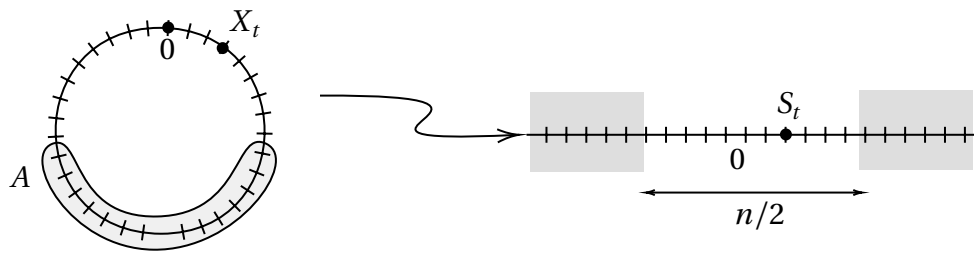


Figure 4.3: The picture to have in mind for the lower bound of the walk on  $\mathbf{Z}_n$

*Proof of upper bound.* 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  $Y$  makes a move. If at some point they are in the same location, we move them together. We see that the marginals for  $X$  and  $Y$  are indeed those of simple lazy random walks on  $\mathbf{Z}_n$ . Letting  $D_t$  denote the clockwise distance between the two particles, we note that  $D_t$  defines a simple random walk on the interior vertices of  $\{0, 1, \dots, n\}$  and gets absorbed at 0 or  $n$ . Thus in this case,  $\tau_{\text{couple}} = \min\{t \geq 0 : D_t \in \{0, n\}\}$ , and from Gambler's Ruin we know that  $\mathbf{E}_{x,y}(\tau_{\text{couple}}) = k(n-k)$  where  $k$  is the clockwise distance between the initial states  $x$  and  $y$ . Therefore using Corollary 4.8, we see that using Markov's Inequality and using the fact that  $k(n-k)$  is maximised at  $n^2/4$ :

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

Therefore for  $t^* = n^2$  we have that  $d(t^*) \leq 1/4$  and as such  $t_{\text{mix}} \leq t^* = n^2$ . ♡

**Main idea:** (for lower bound) Once again we are going to have to get dirty: start by noting that

$$\|P^t(x, \cdot) - \pi\|_{\text{TV}} \geq \pi(A) - P^t(x, A)$$

for all sets  $A$ , so to obtain a good lower bound, we would like a large set  $A$  (large in the sense of  $\pi$ -measure), but which hopefully has a small probability of the chain being there at time  $t$ . We can intuitively imagine, that if our chains starts say at the top of the cycle (say 0), and if  $n$  is large, for small enough values of  $t$ , the chain will have a hard time reaching the lower half of the cycle, so we may choose this one.

*Proof of lower bound.* To obtain a lower bound, we note that it is always the case that

$$d(t) \geq \pi(A) - \mathbf{P}_0(X_t \in A)$$

So if we can find a subset  $A$  of our state space such that  $\pi(A) - \mathbf{P}_0(X_{t^*} \in A) > 1/4$  for some time  $t^*$

then it will follow that  $t_{\text{mix}} \geq t^*$ . Consider the set  $A$  consisting on the lower half of the cycle, i.e:  $A = \{k \in \mathbb{Z}_n : |k| \geq n/4\}$ , then by letting  $S_t$  be a lazy random walk on  $\mathbb{Z}$  which tracks the steps that  $X_t$  has done (i.e: jumps +1 if  $X_t$  moves clockwise and jumps -1 if it moves counter-clockwise) then we have that if  $X_t$  is in the lower half of the cycle, then  $S_t$  is further away from 0 than  $n/4$ . Moreover we note that the variance of  $S_t$  is  $t/2$ , so

$$\mathbf{P}_0(X_t \in A) \leq \mathbf{P}_0(|S_t| > n/4) \leq \frac{16 \text{Var}(S_t)}{n^2} = \frac{8t}{n^2}$$

Choosing any  $t < n^2/32$ , for example  $t = n^2/33$  we then have that  $\mathbf{P}_0(X_t \in A) < 1/4$  but since  $A$  contains at least  $n/2$  vertices and  $\pi$  is easily checked to be the uniform distribution (also follows intuitively) we have that  $\pi(A) \geq 1/2$  so the claim follows, i.e: that  $t_{\text{mix}} \geq n^2/33$ .  $\heartsuit$

**Example 4.10 (Lazy random walk on binary tree)** Let  $T_k$  be a rooted finite binary tree with a root  $\rho$ . Notice there are  $n = 2^k - 1$  vertices on this tree, and each leaf has degree 1, every other vertex has degree 3. Then for this Markov Chain, we have that

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

**Main idea:** The idea is similar, we want a coupling that prevents the chains from jumping past each other. The idea is that once the chains reach the same level, they start moving together, then they will be coupled by the time they reach the root.

*Proof.* Construct the following coupling  $(X_t, Y_t)$  of two lazy random walks started from  $x_0, y_0$  on the tree. At each move, toss a fair independent coin to decide which of the two chains moves. The one who moves will move uniformly at random amongst its neighbours. This ensures the coupling respects the lazy transition matrix, and moreover, achieves that the particles don't jump past each other. Suppose that at some point they reach the same level (i.e: the same distance away from the root), once this happens we will move both particles together, i.e: keep tossing fair coins, but now if we get heads, we don't just move one and leave the other put, we move both of them. And so when they reach the root, they will have coalesced. Let  $L$  be the set of leaves at the bottom of the tree. Clearly, the coupling time  $\tau_{\text{couple}}$  is smaller than or equal to the time required for  $(X_t)$  to reach  $L$  and then reach the root, because in doing this trajectory it must have visited the same level as  $(Y_t)$ . Let  $\tau$  denote this time. We can now bound  $\tau$  by modelling the height evolution as a biased (biased since the chance of moving down a level is twice the chance of moving up because there are two leaves downwards and only one upwards) lazy simple random walk on the line segment  $\{0, \dots, k-1\}$ . This  $\tau$  is the so-called commute time from  $L$  to the root.

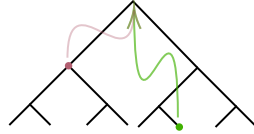


Figure 4.4: A picture of the coupling idea for the upper bound in the binary tree

We take it as a black box from now that  $\mathbf{E}[\tau] \leq 4n$ . Therefore  $\mathbf{P}_{x,y}[\tau_{\text{couple}} > t] \leq \frac{4n}{t}$ , which gives that whenever  $t > 16n$ ,  $d(t) < 1/4$  and so  $t_{\text{mix}} \leq 16n$ . ♡

We now move on to prove the lower bound. For this we will need a tool proven in Example Sheet 1: **Main idea:** The main idea for the lower bound is to use the technique proven in Example Sheet 1, that says that

$$t_{\text{mix}} \gtrsim \max_x \mathbf{E}_x[\tau_A]$$

For any  $A$  with  $\pi(A) \geq 1/8$ , where  $\tau_A$  is the hitting time of  $A$ .

*Lower bound.* This technique is pretty cracked, we can let  $A$  be the right half of the binary tree, and clearly  $\max_x \mathbf{E}_x[\tau_A]$  will be no less than the expected hitting time of  $A$  when starting from a leaf of the left hand side of the tree. Moreover, the hitting time of  $A$  when starting from the left hand side of the tree is precisely the hitting time of the root, which apparently is of order  $n$ . Therefore  $t_{\text{mix}} \gtrsim n$ . ♡

**Example 4.11 (Discrete torus)** Let  $X$  be a simple random walk on the  $d$ -dimensional discrete torus  $\mathbf{T} = (\mathbf{Z}/n\mathbf{Z})^d$ . Then  $t_{\text{mix}} \leq C(d)n^2$ .

*Proof.* The proof is of course by coalescent coupling, and we will inspire this proof with the arguments we did for the upper bound of the mixing time on the cycle. Writing

$$\mathbf{T} = \mathbf{Z}/n\mathbf{Z} \times \cdots \times \mathbf{Z}/n\mathbf{Z}$$

gives the idea that to sample this Markov chain, what we can do is first select one of the  $d$  coordinates at random, and then update that coordinate according to a lazy simple random walk on the cycle. More precisely, we will run two Markov chains  $\mathbf{X}_t = (X_t^1, X_t^2, \dots, X_t^d)$  and  $\mathbf{Y}_t = (Y_t^1, \dots, Y_t^d)$  as follows: select a coordinate  $k \in [d]$ , if the walks agree on that coordinate, i.e:  $X_t^k = Y_t^k$ , then move both walks together according to a LSRW on the corresponding cycle. If the two coordinates disagree however, we will choose one of the two walks at random, keep it still, and move the other one by  $\pm 1$  with probability  $1/2$ . This clearly respects a LSRW law.

We have now constructed a coalescent coupling, and so we turn our attention to bounding the coalescence time from above. We can define the random times

$$T_i = \min\{t \geq 0 : X_t^i = Y_t^i\}$$

i.e: the random times at which the walks agree on coordinate  $i$ . It follows then that the walks become fully coalescent at the last of those times, i.e:

$$\tau_{\text{couple}} = \max_{i \in [d]} T_i$$

Which of course can be bounded as  $\tau_{\text{couple}} \leq \sum_{i=1}^d T_i$ . We are now almost ready. The plan is to use Markov's Inequality, and so we need to know some bounds for  $\mathbf{E}[T_i]$ . This is no problem however, because we know that the distance  $|X_t^i - Y_t^i|$  between the walks in the  $i^{\text{th}}$  coordinate is distributed according to Gambler's ruin to reach 0 or  $n$ , where we interpret this Gambler's ruin as the distance, and we know that the time for coalescence on a LSRW on the cycle is bounded above by  $n^2/4$ . However, we have to note that we only update the  $i^{\text{th}}$  coordinate once its picked, which happens with probability  $1/d$ , and so the expected time for an update to occur is  $d$ . Therefore combining all this:

$$\mathbf{E}_{\mathbf{x}, \mathbf{y}} T_i \leq \frac{d n^2}{4}$$

We can now finish because using the coalescence time bound on mixing times:

$$d(t) \leq \max_{\mathbf{x}, \mathbf{y}} \mathbf{P}_{\mathbf{x}, \mathbf{y}}[\tau_{\text{couple}} > t] \leq \frac{\sum_{i=1}^d \max_{\mathbf{x}, \mathbf{y}} \mathbf{E}_{\mathbf{x}, \mathbf{y}}[T_i]}{t} \leq \frac{d^2 n^2}{4t}$$

and so choosing  $t = n^2 d^2$  gives  $d(t) \leq 1/4$  which means  $t_{\text{mix}} \leq n^2 d^2$ . ♥

The key takeaway from all these examples is that to upper bound mixing times with coupling arguments, one just considers a coupling where they coalesce, and then the hard part boils down to proving upper bounds for the expectation of this coalescence time. This is usually done with Gambler's ruin or other classic Markov chains. Remember that if  $(Z_t)$  is a simple random walk on  $\{1, \dots, n\}$ , Gambler's ruin says that the expected time to hit 1 or  $n$  is  $k(n-k)$  where  $k$  is the initial starting point.

Constructing couplings of Markov chains is useful sometimes even if the coupling we construct is not Markovian, but the philosophy here is that as long as you respect the transition probabilities, you have some leeway into how you sample your Markov chain, and this leeway is what helps you solve some problems, let see an example:



**Example 4.12** Let  $X$  be a Simple Random Walk on a graph  $G$ , and let  $Y$  be a Lazy Simple Random Walk on a graph  $G$ . Let  $R_X(t)$  (resp.  $Y$ ) to be the range of the walk  $X$  (resp.  $Y$ ) by time  $t$ , i.e:

$$R_X(t) = |\{X_1, \dots, X_t\}|$$

Then, for any starting distribution  $\mu$  on  $G$ , we have that  $\mathbf{P}_\mu[R_X(t) > m] \geq \mathbf{P}_\mu[R_Y(t) > m]$ . Which seems like a really obvious statement right? Since  $X$  is not lazy, it will move more and so it will visit more vertices.

**Main idea:** The idea here is to construct a coupling  $(X'_t, Y'_t)$  of the walks  $X$  and  $Y$ . This will be useful because in a coupling, the laws of the processes  $X$  and  $X'$  will be the same, which means that statements of the form  $\{R_X(t) > m\}$  and  $\{R_{X'}(t) > m\}$  will have the same probabilities, this is true because these statements are statements of the form

$$\{(X_1, \dots, X_t) \in A\} \quad \{(X'_1, X'_2, \dots, X'_t) \in A\}$$

where  $A$  is some special measurable set that describes the union of all possible state sets that have at least  $m$  elements. This is of course a measurable set, and since the two processes have the same law, the two events above have the same probability. Technicalities aside, this will help us because perhaps by sampling  $X'$  in a smart way that respects the law of  $X$ , we can actually compare in an easier way the ranges of  $X'$  and that of  $Y$ . Then the key is to simply set  $Y' = Y$  and then sample  $X'$  to be the "jumps of  $Y$ ".

*Proof.* We consider the following coupling of  $(X, Y)$ . Let  $Y' = Y$ , and defining the random times  $T_0 = 0$  and  $T_i = \min\{t > T_{i-1} : Y_t \neq Y_{t-1}\}$ , i.e: the set of jump times of  $Y$ . We can now sample  $X'_t = Y_{T_i}$ . I.e: we are sampling the motion of  $X'$  to be that of  $Y$  but skipping the waiting in between each jump. Then it is clear that the law of  $X'$  is that of a Simple Random Walk on  $G$ . But now it is easy to compare the ranges. Indeed:

$$\begin{aligned} \mathbf{P}_\mu[R_X(t) > m] &= \mathbf{P}_\mu[R_{X'}(t) > m] \\ &:= \mathbf{P}_\mu[|\{X'_1, X'_2, \dots, X'_t\}| > m] \\ &\stackrel{(1)}{=} \mathbf{P}_\mu[|\{Y_{T_1}, Y_{T_2}, \dots, Y_{T_t}\}| > m] \\ &\stackrel{(2)}{=} \mathbf{P}_\mu[|Y_0, Y_1, \dots, Y_{T_t}| > m] \\ &=: \mathbf{P}_\mu[R_Y(T_t) > m] \\ &\stackrel{(3)}{\geq} \mathbf{P}_\mu[R_Y(t) > m] \end{aligned}$$

Where (1) comes from the construction of  $X'$ , step (2) comes from the fact that  $Y$  does not move

at times different from  $\{T_i\}$  and so we may include them without interfering with the range, step (3) comes from the fact that  $t \leq T_t$ , this is because  $Y_t$  can't make  $t$  jumps in time less than  $t$ , and also using the fact that the range function is non-decreasing (set-wise) with  $t$ , this is clear because if you have a larger time you can't visit a smaller amount of states. ♡

Let's see one last example:

**Example 4.13 (Winning streak)** Let us describe the winning streak Markov chain. Imagine that after a night of questionable habits I limit my memory to a size of  $n$ . Then I keep tossing Bernoulli  $1/2$  random variables and try to remember the longest winning streak (i.e. of 1's) that I've had. Of course, after I've had  $n$  wins in a row, I could keep having more wins and I would still think that my winning streak is of  $n$ . It is clear that this Markov chain has transition probabilities

$$\begin{cases} P(i, 0) = 1/2 & i \leq n, \\ P(i, i+1) = 1/2 & i < n, \\ P(n, n) = 1/2. \end{cases}$$

The mixing time of this chain can be once again upper bounded with a coalescence time argument. The coupling in this case is quite simple, consider two winning streak chains  $(X_t)$  and  $(Y_t)$ , each with different starting points  $a, b \in \{1, \dots, n\}$ . We can run these chains by using the same sequence of Bernoulli  $1/2$  random variables, and so it is also clear that the coalescence time will be the first time a zero is obtained. Therefore by stochastically dominating  $\tau_{\text{couple}}$  with a Geometric  $1/2$  random variable, we see that  $\mathbf{P}[\tau_{\text{couple}} > t] \leq 2^{-t}$ , which gives that  $t_{\text{mix}}(\epsilon) \leq \lceil \log_2(1/\epsilon) \rceil$ .

# Chapter 5

## Strong Stationary Times

In this section we learn what a strong stationary time is, roughly speaking, a random time  $T$  "independent from the chain", after which the Markov chain is distributed according to its stationary distribution. We will show that

$$d(t) \leq \max_{x \in \Omega} \mathbf{P}_x(T > t)$$

and thus obtain another method of bounding mixing times.

We begin by recalling the definition of a filtration.

**Definition 5.1** Let  $(\Omega, \mathcal{F}, \mathbf{P})$  be a probability space. An increasing sequence of sub-sigma-algebras  $\{\mathcal{F}_t\}_{t \geq 0}$  is called a filtration.

**Definition 5.2** We say that a sequence of random variables  $(X_n)$  is a Markov chain with respect to  $\{\mathcal{F}_t\}$  if

$$\mathbf{P}_x(X_{t+1} = y \mid \mathcal{F}_t) = P(X_t, y)$$

**Remark 5.3** The usual sense of a Markov chain corresponds to being a Markov chain with respect to the natural filtration  $\mathcal{F}_t = \sigma(X_1, \dots, X_t)$ .

**Definition 5.4** Let  $\{\mathcal{F}_t\}$  be a filtration such that  $(X_n)$  is adapted to it and is a Markov Chain with respect to it. A Randomised Stopping Time  $T$  is a random variable such that  $\{T \leq t\} \in \mathcal{F}_t$ .

**Remark 5.5** The point of using a potentially larger sigma algebra instead of the natural filtration is that we may want to allow some extra source of randomness to interact with the process.

**Example 5.6 (Hitting time)** For a set  $A \subseteq \Omega$ , the random variable

$$\tau_A = \inf\{t \geq 0 : X_t \in A\}$$

is a stopping time, because

$$\{\tau_A \leq t\} = \bigcup_{i=1}^t \{\tau_A = i\} = \bigcup_{i=1}^t \{X_0 \notin A, X_1 \notin A, \dots, X_i \in A\}$$

and this event clearly belongs to  $\mathcal{F}_t$

**Definition 5.7 (Stationary time, Strong Stationary Time)** Let  $(X_t)$  be an irreducible Markov chain with stationary distribution  $\pi$ . A stationary time  $\tau$  for  $(X_t)$  is a stopping time, possibly depending on the starting position  $x$ , such that the distribution of  $X_\tau$  is  $\pi$ , i.e:

$$\mathbf{P}_x(X_\tau = y) = \pi(y)$$

If moreover, we have that  $X_\tau$  is independent of  $\tau$ , i.e: for all  $y \in S$

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

then we refer to  $\tau$  as a strong stationary time.

We hope to use strong stationary times to give other ways of bounding distance to stationarity and hence mixing times. Indeed, since at the strong stationary time  $\tau$ , the chain is distributed according to  $\pi$ , if there is a very high probability that for a time  $t$ ,  $\tau \geq t$ , then the distance between the current distribution and the invariant distribution should also be quite high. Conversely, if the probability that  $\tau \geq t$  is zero, meaning that  $\tau < t$  with probability 1, then it makes sense that at time  $t$ , we have already crossed the "threshold to stationarity" and as such we are already invariantly distributed so the distance should be zero. To give a formal proof, it will be convenient to talk about the

**Definition 5.8 (Separation distance, )** We define the separation distance  $s_x(t)$  as

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

and we also define  $s(t)$  by

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

**Lemma 5.9** (Separation distance upper bounds TV distance) For all  $x \in \Omega$ , we have that

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

Therefore  $d(t) \leq s(t)$ .

**Main idea:** The main idea is to use the equivalent characterisation of TV distance that says that for  $B = \{x : \mu(x) > \nu(x)\}$

$$\|\mu - \nu\|_{\text{TV}} = \mu(B) - \nu(B)$$

Then expressing this as a sum and extracting a factor of  $s_x(t)$  finishes the proof.

*Proof.*

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

♡

**Lemma 5.10** (Strong stationary time bounds separation distance) If  $\tau$  is a strong stationary time, then for all  $x \in \Omega$ , we have that for all  $t$ :

$$s_x(t) \leq \mathbf{P}_x(\tau > t)$$

**Main idea:** The heart of the proof resides in showing that

$$\mathbf{P}_x[\tau \leq t, X_t = y] = \mathbf{P}[\tau \leq t] \pi(y)$$

Doing this is just cleverly using the Law of Total Probability. To intuitively know why each step should be done, note that we have an event of the form  $\{\tau \leq t\}$  and our definitions concern events of the type  $\{\tau = s\}$ , and moreover, we will need to introduce somehow an event of the type  $\{X_s = z\}$  so that we can exploit the definition of a strong stationary time.

*Proof.* We first show that

$$\mathbf{P}_x(\tau \leq t, X_t = y) = \mathbf{P}(\tau \leq t)\pi(y)$$

This is just a mildly unpleasant calculation where we use the law of total probability a few times, then some conditional probability and finally the fact that  $\tau$  is a strong stationary time.

$$\begin{aligned} \mathbf{P}_x(\{X_t = y\} \cap \{\tau \leq t\}) &= \sum_{s \leq t} \mathbf{P}_x(\{X_t = y\} \cap \{\tau = s\}) \\ &= \sum_{s \leq t} \sum_{z \in \Omega} \mathbf{P}_x(\{X_t = y\} \cap \{\tau = s\} \cap \{X_s = z\}) \\ &= \sum_{s \leq t} \sum_{z \in \Omega} \underbrace{\mathbf{P}_x(\{X_t = y\} \mid \{\tau = s\} \cap \{X_s = z\})}_{P^{t-s}(z, y)} \underbrace{\mathbf{P}_x(\{\tau = s\} \cap \{X_s = z\})}_{\pi(z)\mathbf{P}_x(\tau = s)} \\ &= \sum_{s \leq t} \mathbf{P}_x(\tau = s) \sum_{z \in \Omega} \pi(z) P^{t-s}(z, y) = \mathbf{P}_x(\tau \leq t)\pi(y) \end{aligned}$$

Now that we have this, we know that

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

♡

As an immediate result we have that

$$d(t) \leq \max_{x \in \Omega} \mathbf{P}_x(\tau > t)$$

Which is what we claimed was true in the initial discussion. Before moving on to examples we give one more application of separation distance as a bound of distance to stationarity.

**Lemma 5.11** (Separation distance is upper bounded by distance to stationarity) We have that for a reversible chain

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

**Main idea:** The key here is that since  $s(2t)$  involves an expression of the form

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

and we can now reverse this second part of the product and then see an expectation, where we can use Jensen's inequality.

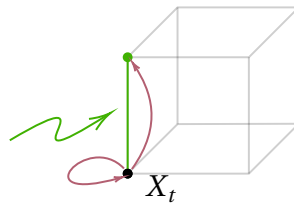


Figure 5.1: The refresh mechanism in the hypercube example: a dimension is chosen at random, in this illustration the vertical dimension, and then the walk must choose whether to stay in place or to jump

*Proof.* We have the following straightforward calculation

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

Where step (1) is Jensen, and step (2) is an awesome trick I had never seen before.



## 5.1 Examples

We will now present two examples where strong stationary times are used to bound the value of the mixing time.

**Example 5.12 (Mixing time of lazy random walk on hypercube)** The  $n$ -dimensional hypercube is the graph whose vertices are  $\{0, 1\}^n$  and two vertices are connected if and only if they differ by exactly one coordinate. Then for any  $\epsilon > 0$ , there is some constant  $C(\epsilon) > 0$  such that

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

**Main idea:** We aim to construct a strong stationary time. For this we invent a refresh mechanism on the dimensions of the hypercube and use as stationary time the random time when all coordinates have been refreshed at least once, then we use the coupon collector problem to estimate this time.

*Proof.* We first note that by the symmetry of the chain, the invariant distribution of the chain is going to be uniform. This can also be quickly checked with pen and paper. Now we set off to find a strong stationary time. First, we note that we can realise the lazy simple random walk by choosing at every step a coordinate, say

$$\overbrace{(0, 1, 0, 0, \dots, \underbrace{1}_{\uparrow}, 0, 1, \dots, 1)}^{n \text{ bits}}$$

and refreshing the coordinate, that is to say, replacing it by a uniform choice on  $\{0, 1\}$ . We define  $\tau_{\text{refresh}}$  to be the least time where all coordinates have been chosen at least once. Observe that once  $\tau_{\text{refresh}}$  is reached, all of the coordinates have been replaced with independent fair bits, the distribution on the chain is uniform on  $\{0, 1\}^n$ , that is  $X_{\tau_{\text{refresh}}}$  is an exact sample from the stationary distribution. From this we can show that  $\tau = \tau_{\text{refresh}}$  is a strong stationary time. Indeed:

$$\mathbf{P}_x(\tau = t, X_\tau = y) = \mathbf{P}_x(X_\tau = y \mid \tau = t) \mathbf{P}_x(\tau = t) = \pi(y) \mathbf{P}_x(\tau = t)$$

as required. Now we know that

$$d(t) \leq \max_{x \in \{0, 1\}^n} \mathbf{P}_x(\tau > t)$$

However, it is easy to see that  $\tau$  has the same distribution as the Coupon Collector Problem (note that I have written  $\mathbf{P}_x(\tau > t)$  but  $\tau$  is independent of the starting point really). From the coupon collector problem, we know that if we set  $t = n \log n + Cn$  (where  $C$  is for us to choose), then

$$d(t) \leq \exp(-C)$$

By making  $C = C(\epsilon)$  large enough, we make this right hand side less than  $\epsilon$ , and as such we see that

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



**Remark 5.13** This argument we did could have also been shown earlier in these notes with the technique of upper bounding mixing times by coupling: consider two chains  $(X_t)$  and  $(Y_t)$  on the hypercube with different starting points  $x, y \in \{0, 1\}^n$ . Then we can couple the chains by selecting one bit in both and updating it according to a uniform choice on  $\{0, 1\}$ , it is clear that by the time



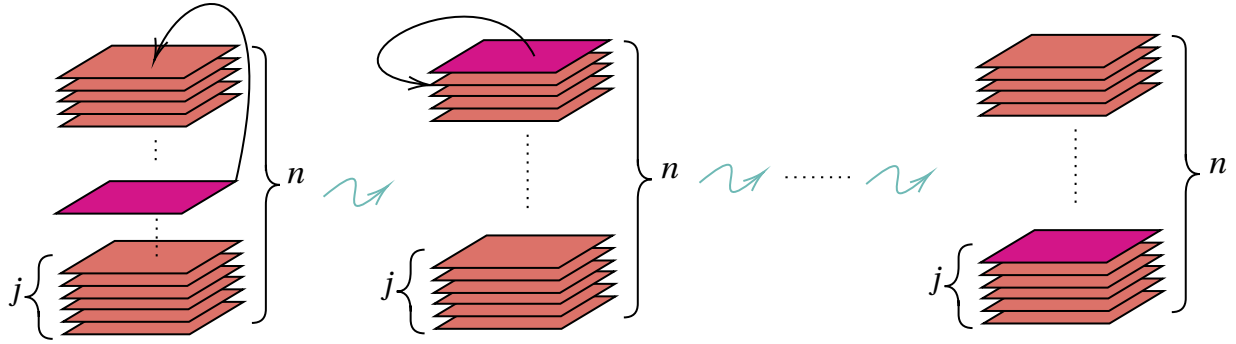


Figure 5.2: Picture to have in mind for the relationship between  $\tau_j$  and  $A_j$

we have selected all bits, the chains have met, therefore the same ideas of the Coupon Collector Problem can be used in this framework.

**Example 5.14 (Random-to-top shuffle mixing time)** Consider a deck of  $n$  cards with the random-to-top shuffling method. Then for all  $\epsilon > 0$ , there exists some constant  $C(\epsilon)$  such that

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

**Remark 5.15** Note that essentially this is a random walk on the group  $S_n$ , where each  $\sigma \in S_n$  represents a card shuffling. Moving card  $j$  to the top represents multiplying the current state  $\sigma$  on the left by  $(1\ 2 \cdots j)$ .

**Main idea:** Here the idea is to get dirty, and we use the set  $A$  of all permutations  $\sigma$  such that the bottom  $j$  cards are in increasing order (not in strict order, just increasing order). Then this set has a very small  $\pi$ -measure, and if we haven't picked more than  $n - j + 1$  cards we will still be on this set.

*Lower Bound.* First and foremost note that  $\pi$  is uniform (pen and paper, check that all the probability going into a state  $y \in S_n$  is 1). For the lower bound, we once again note that

$$d(t) = \max_{\sigma \in S_n} \|P^t(\sigma, \cdot) - \pi(\cdot)\|_{\text{TV}} \geq P^t(\text{id}, A) - \pi(A)$$

for all  $A$ . To obtain a good bound, what we then want, is a set  $A$  of shufflings such that  $\pi(A)$  is really small but  $P^t(\text{id}, A)$  is then large enough for some time  $t$ , then we can make the right hand side greater than  $\epsilon$  and bound the mixing time from below. The set of shufflings that we want is

$$A_j = \{\sigma \in S_n : \sigma(n-j) < \sigma(n-j+1) < \cdots < \sigma(n)\}$$

the set of shufflings where the bottom  $j$  cards are in increasing order (not strict order, just increasing order). Using that  $\pi$  is uniform, we see that

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

which gets small very quick. Intuitively, if there are enough cards and  $j$  is not a huge proportion of the cards, for small values of time, it will be very likely that we are still in a shuffling that preserves the order of the last  $j$  cards. Let us make this precise. Let  $X_t$  be our random walk on  $S_n$  and let  $T_j$  be the time by which  $n-j+1$  different cards have been picked. Then the claim is that if  $t < T_j$ , then  $X_t \in A_j$ . To see this, **look at the diagram**: When we start, all the cards are in increasing order. In our first step, we will introduce a disorder, represented by moving the purple card to the top, since the order went from  $1, 2, \dots, k-1, k, k+1, \dots$  to  $k, 1, 2, \dots, k-1, k+1, \dots$  (note that  $k-1, k+1$  is still in increasing order) (there is disorder unless we pick the top card but we don't care) and this disorder will only affect the  $j$  bottom cards after we have pushed the purple card all the way back down to the bottom  $j$  cards. In total this required  $n-j+1$  moves. In precise terms:

$$\mathbf{P}[X_t \in A_j] \geq \mathbf{P}[T_j > t] \text{ or also } \mathbf{P}[X_t \in A_{j-1}] \geq \mathbf{P}[T_j \geq t]$$

so we now set off to bound this last probability through our usual means, some sort of Markov's or Chebyshev's inequality. In this case, we are able to bound both the mean and the variance so we will use Chebyshev:

- Mean of  $T_j$ : note that this is morally the same object as the one in coupon collector, and likewise, we can express  $T_j$  as a sum of independent geometrically distributed random variables,  $T_j = Q_1 + Q_2 + \dots + Q_{n-j+1}$  where  $Q_i$  represents the additional cards we need to select in order to collect one more kind of card. We have that if we have already selected  $i-1$  kinds of cards, the probability of choosing a new one is precisely

$$\frac{N-i+1}{N}$$

So  $\mathbf{E}[Q_i] = \frac{N}{N-i+1}$  therefore:

$$\begin{aligned} \mathbf{E}[T_j] &= \mathbf{E}[Q_1] + \dots + \mathbf{E}[Q_{n-j+1}] \\ &= 1 + \frac{n}{n-1} + \dots + \frac{n}{n-(n-j+1)+1} \\ &= n \sum_{k=j}^n \frac{1}{k} \geq n \int_j^n \frac{dk}{k} = n[\log n - \log j] \end{aligned}$$

- Variance of  $T_j$ : we recall momentarily that if  $X \sim \text{Geo}(p)$  then  $\text{Var}(X) = \frac{1-p}{p^2} \leq \frac{1}{p^2}$ , so from

this and the fact that the  $Q_i$ 's are independent we gather that

$$\begin{aligned}
 \text{Var}(T_j) &= \sum_{k=1}^{n-j+1} \text{Var}(Q_k) \\
 &\leq \sum_{k=1}^{n-j+1} \frac{n^2}{(n-k+1)^2} \\
 &= \sum_{k=j}^n \frac{n^2}{k^2} \\
 &\leq \sum_{k=j}^n \frac{n^2}{k(k-1)} \\
 &\leq \frac{n^2}{j-1}
 \end{aligned}$$

Where the last inequality comes from expressing  $1/k(k-1)$  as a partial fraction and telescoping the sum.

We are now ready to put this together and employ Chebyshev's inequality. We have the following (watch out for the highlighted steps)

$$\begin{aligned}
 \mathbf{P}[T_j \geq n \log n - Cn] &\stackrel{(!)}{\geq} \mathbf{P}[|T_j - \mathbf{E}T_j| < n(C - \log j)] \\
 &= 1 - \mathbf{P}[|T_j - \mathbf{E}T_j| \geq n(C - \log j)] \\
 &\stackrel{(2)}{\geq} 1 - \frac{\text{Var}(T_j)}{n^2(C - \log j)^2} \\
 &\stackrel{(3)}{\geq} 1 - \frac{n^2}{(j-1)n^2(C - \log j)^2} \\
 &\geq 1 - \frac{1}{j-1}
 \end{aligned}$$

Here step (!) comes from the following observation: note that if  $|T_j - \mathbf{E}T_j| < n(C - \log j)$  then  $T_j \in (\mathbf{E}T_j - n(C - \log j), \mathbf{E}T_j + n(C - \log j))$  which in particular means that  $T_j > \mathbf{E}T_j - n(C - \log j) \geq n \log n - Cn$  due to the bound on the expectation. Therefore  $\{|T_j - \mathbf{E}T_j| < n(C - \log j)\} \subseteq \{T_j \geq n \log n - Cn\}$ . Then step (2) comes from Chebyshev's inequality, and step (3) comes from the bound on the variance we obtained. The last step is just a brutal approximation by dropping the  $(C - \log j)^2$  whenever  $C \geq \log j + 1$ . We are now almost done with the lower bound, we put this

all together and note that

$$\begin{aligned} d(n \log n - Cn) &\geq P^{n \log n - Cn}(id, A_{j-1}) - \pi(A_{j-1}) \\ &\geq 1 - \frac{1}{j-1} - \frac{1}{(j-1)!} \\ &\geq 1 - \frac{2}{j-1} \end{aligned}$$

and by taking  $j = \lceil \exp(C-1) \rceil$  to satisfy the constraint on  $C$  we mentioned above, we get that

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

and so by taking  $C = C(\epsilon)$  to be a large enough number, we can guarantee that this right hand side is larger than  $\epsilon$  and so we have that

$$t_{\text{mix}}(\epsilon) \geq n \log n - C(\epsilon)n$$

♡

I suppose that the moral of the story is that lower bounds with a hands-on approach are hard. Thankfully for the upper bound we can just use the theory we have now developed, i.e: find a strong stationary time.

*Upper Bound.* We find a strong stationary time. The guess for this would be the time  $\tau$  at which all the cards have been selected at least once. This is clearly distributed to coupon collector and so we can obtain an explicit formulation for  $\mathbf{P}[\tau > t]$ . Indeed, if we let  $A_i$  be the event that the  $i^{\text{th}}$  card hasn't been selected by time  $t$ , then

$$\begin{aligned} \mathbf{P}[\tau > t] &= \mathbf{P}\left[\bigcup_i A_i\right] \\ &\leq \sum_i \mathbf{P}[A_i] \\ &= \sum_i \left(1 - \frac{1}{n}\right)^t \leq n \exp(-t/n) \end{aligned}$$

which means that for any  $C$ , we have that

$$\mathbf{P}[\tau > n \log n + Cn] \leq \exp(-C)$$

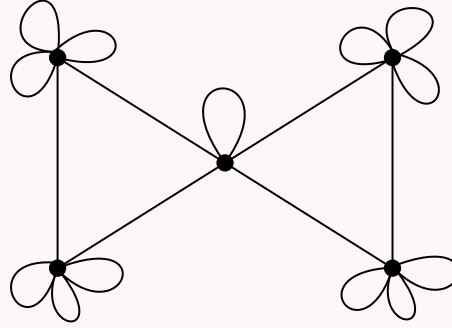
which can be made as small as we want, i.e: we can choose a  $C = C(\epsilon)$  large enough so that the right hand side is less than  $\epsilon$ . Of course this does not finish the problem, we still need to see why  $\tau$  is a strong stationary time. To do so, we introduce the following notation. Let  $k(t)$  be the

number of cards we have picked by time  $t$ , and let  $C_k$  be the random set of the top  $k$  cards. We will prove by induction on  $t$  that the set  $C_{k(t)}$  is a uniform subset of  $\{1, \dots, n\}$  of size  $k$  where the ordering of the cards is also uniform. The base case is clear ( $t = 1$ ). Suppose that for some time  $t$ , we have picked  $k$  cards, and that the set  $C_k$  is uniformly random, both in the sense of which cards it has and the ordering of the cards in it. Then at time  $t + 1$  we have two cases, either we pick a card from within  $C_k$ , in which case it is obvious that the order remains uniformly at random, or we pick a card from the remaining  $\{1, \dots, n\} \setminus C_k$  cards. We note that these remaining cards are not in random order, but the cards themselves are random, so by picking one of those cards, and placing it on the top, now the set  $C_{k+1}$  has  $k + 1$  cards which are randomly chosen, and the order is indeed uniformly at random, because the first card is random, and the next  $k$  cards are both random cards and with random order, therefore the order of the total  $k + 1$  cards is also random. (Just to reiterate, the first card could have been any card, and the next  $k$  are all random and have random order). Therefore by the time we have picked all cards, the entire deck is a uniformly at random permutation of  $\{1, \dots, n\}$ . ♡

**Remark 5.16 (On the top-to-random shuffle)** We have just described the random-to-top shuffle method, but we could talk about a very similar chain, the top-to-random shuffle, where instead of choosing a card at random and placing it on the top, we get the card on the top and place it somewhere at random. If  $P$  and  $\hat{P}$  are the transition matrices on the group  $\mathcal{S}_n$  of the random-to-top and top-to-random shuffle, it is clear that for any permutation  $a \in \mathcal{S}_n$ , we have that  $P(\text{id}, a) = \hat{P}(\text{id}, a^{-1})$ , it is also clear that in both of these chains the invariant distribution is uniform (as a matter of fact, any random walk on a group where the increments are given by group multiplication has uniform invariant distribution), therefore

$$\|P(\text{id}, \cdot) - \pi\|_{\text{TV}} = \frac{1}{2} \sum_{\sigma \in \mathcal{S}_n} |P(\text{id}, \sigma) - \pi(\sigma)| = \frac{1}{2} \sum_{\sigma \in \mathcal{S}_n} |\hat{P}(\text{id}, \sigma) - \pi(\sigma)| = \|\hat{P}(\text{id}, \cdot) - \pi\|_{\text{TV}}.$$

**Example 5.17 (The flower graph)** To endure the horrors of mixing times, we study Markov chains on pretty graphs, like the flower graph: consider two copies of the complete graph  $K_n$ , and glue them at a single vertex  $v^*$ . Then add  $n$  loops to each vertex different from  $v^*$  and add one final loop at  $v^*$ , see the example for  $n = 3$ :



On this graph we have that the mixing time of a simple random walk is order  $n$ .

*Upper bound.* To prove the upper bound we come up with the following strong stationary time. Let  $\tau_{v^*}$  be the hitting time of the vertex  $v^*$ , and note that once the walk is at  $v^*$ , the next state of the walk can be any vertex of this graph, and each occurs with equal probability. Note that this graph is such that each vertex has equal degree, so the stationary distribution is indeed uniform, and so  $\tau := \tau_{v^*} + 1$  is going to be a strong stationary time: strong because the state of the walk at  $\tau_{v^*} + 1$  is independent of this time, and stationary because of what we have just described. Then, we have that

$$\begin{aligned} \mathbf{P}[\tau > t] &\leq \mathbf{P}[\tau_{v^*} > t] \\ &\leq \left(1 - \frac{1}{2n-1}\right)^{t-1} \\ &\leq \left(1 - \frac{1}{2n-1}\right) \left(1 - \frac{1}{2n-1}\right)^{t-2} \\ &\leq \left(1 - \frac{1}{2n-1}\right)^{t-2} \end{aligned}$$

and so if we choose  $t = 4n$ , this thing above is less than or equal to  $e^{-2} < 1/4$ . And so  $t_{\text{mix}} \leq 4n$ .

♡

Hint for the lower bound: consider the set of vertices comprising one of the copies of  $K_n$ .

*Lower bound.* For the lower bound, as usual, we get dirty. The hint suggests that we need to deal with a set of large  $\pi$ -measure, and small probability of being there by time  $t$ . In particular, let  $y \in V$  be any vertex of the flower-graph distinct from  $v^*$ , and let  $A$  be the complete graph which does not contain  $y$ . Then  $d(t) \geq \pi(A) - \mathbf{P}_y[X_t \in A]$ , where  $(X_t)$  is the random walk on this graph.

Now we have the following easy calculation:

$$\begin{aligned}
 \mathbf{P}_y[X_t \in A] &\leq \mathbf{P}_y\left[\bigcup_{1 \leq s \leq t} \{X_s = v^*\}\right] \\
 &\leq \sum_{1 \leq s \leq t} \mathbf{P}_y[X_s = v^*] \\
 &\leq \sum_{1 \leq s \leq t} \sum_{x \in V} \mathbf{P}_y[X_s = v^* \mid X_{s-1} = x] \mathbf{P}_y[X_{s-1} = x] \\
 &= \frac{1}{2n-1} \sum_{1 \leq s \leq t} 1 = \frac{t}{2n-1}.
 \end{aligned}$$

Moreover, note that  $\pi$  being uniform implies that  $\pi(A) = \frac{n}{2n-1}$ . Therefore we wish to choose  $t$  so that

$$\frac{n}{2n-1} - \frac{t}{2n-1} \geq \frac{1}{4},$$

which rearranging gives that  $t \leq \frac{n}{2} \left(1 - \frac{1}{4n}\right)$ , (i.e, if  $t$  is not greater than this number, then the distance to stationarity is greater than or equal to  $1/4$ ), and so we conclude that

$$t_{\text{mix}} \geq \frac{n}{2}(1 + o(1))$$







# Chapter 6

## Cutoff

In example 5.14, we have shown that in some sense, all the mixing happens at time  $n \log n$ . This motivates the following definition and concept which will come to stay for the rest of the course.

**Definition 6.1 (Cutoff)** A sequence  $X^n$  of Markov chains has cutoff if for all  $1 > \epsilon > 0$ ,

$$\lim_{n \rightarrow \infty} \frac{t_{\text{mix}}^{(n)}(\epsilon)}{t_{\text{mix}}^{(n)}(1-\epsilon)} = 1$$

The intuition behind this definition is that as the "size of the chain" ( $n$  doesn't need to be size but it helps to think about it) grows, then the times for which you are arbitrarily close or far to stationarity are pretty much the same time, meaning that all the mixing, happens very abruptly. In particular, as we will see in the following remark, having cutoff is equivalent to saying that all the mixing occurs at  $t_{\text{mix}}$ .

**Remark 6.2** A perhaps more clear illustration is that an equivalent characterisation for cutoff is, writing  $d_n(t)$  for  $d(t)$  with respect to  $X^{(n)}$ , that there is cutoff if and only if we have that

$$\lim_{n \rightarrow \infty} d_n(c t_{\text{mix}}^{(n)}) = \begin{cases} 1 & c < 1 \\ 0 & c > 1 \end{cases} \quad (\star)$$

**Main idea:** Work with epsilonotics and use the definition of mixing times in both directions, i.e: if  $d(t) < \epsilon$  then  $t_{\text{mix}}(\epsilon) < t$  and if  $t_{\text{mix}}(\delta) > s$ , then  $d(s) > \delta$ .

*Proof.* Assume first that that expression  $(\star)$  holds. Let  $\gamma > 0$ , then we can write  $c > 1$  as  $c = 1 + \gamma$ . Since  $\lim_{n \rightarrow \infty} d_n(c t_{\text{mix}}^{(n)}) = 0$ , then by definition, for any  $\epsilon > 0$  we have that there is some  $n$  large enough so that  $d_n((1 + \gamma) t_{\text{mix}}^{(n)}) < \epsilon$ , this by definition of mixing time means that  $t_{\text{mix}}^{(n)}(\epsilon) \leq (1 + \gamma) t_{\text{mix}}^{(n)}$ . Similarly we gather that for  $n$  large enough,  $t_{\text{mix}}^{(n)}(1 - \epsilon) \geq (1 - \gamma) t_{\text{mix}}^{(n)}$ . From this

we conclude that

$$\frac{t_{\text{mix}}^{(n)}(\epsilon)}{t_{\text{mix}}^{(n)}(1-\epsilon)} \leq \frac{1+\gamma}{1-\gamma} \rightarrow 1 \quad \gamma \downarrow 0$$

But of course we always have that  $t_{\text{mix}}^{(n)}(1-\epsilon) \leq t_{\text{mix}}^{(n)}(\epsilon)$  so dividing by  $t_{\text{mix}}^{(n)}(1-\epsilon)$  gives that the desired ratio is always lower bounded by 1 so we have the desired defining limit of cutoff. Now conversely, suppose we have cutoff. Then start by fixing a  $\gamma > 0$ . Then by definition of cutoff, for any  $\epsilon > 0$ , there is some  $n$  large enough so that

$$\left| \frac{t_{\text{mix}}^{(n)}(\epsilon)}{t_{\text{mix}}^{(n)}(1-\epsilon)} - 1 \right| < \gamma.$$

From this we gather two things: first that

$$t_{\text{mix}}^{(n)}(\epsilon) < (1+\gamma)t_{\text{mix}}^{(n)}(1-\epsilon) \stackrel{(!)}{\leq} (1+\gamma)t_{\text{mix}}^{(n)}$$

Where (!) comes from assuming that  $\epsilon < 3/4$  without loss of generality (we will later send  $\epsilon$  to zero). This means that by non-increasingness of  $d_n(t)$ ,

$$\lim_{n \rightarrow \infty} d_n[(1+\gamma)t_{\text{mix}}^{(n)}] < \epsilon \rightarrow 0 \quad \epsilon \downarrow 0$$

Similarly we will have that

$$\lim_{n \rightarrow \infty} d_n[(1-\gamma)t_{\text{mix}}^{(n)}] > 1-\epsilon \rightarrow 1 \quad \epsilon \downarrow 0$$



**Definition 6.3 (Cutoff window)** A sequence of Markov chains  $X^n$  is said to have cutoff with cutoff window of order  $w_n$  if  $w_n \ll t_{\text{mix}}^{(n)}$  and if for any  $1 > \epsilon > 0$  there exists some  $C(\epsilon) > 0$  such that

$$t_{\text{mix}}^{(n)} - C(\epsilon)w_n \leq t_{\text{mix}}^{(n)}(\epsilon) \leq t_{\text{mix}}^{(n)} + C(\epsilon)w_n$$

**Remark 6.4** Here is some intuition behind this. Let  $\epsilon > 0$  be a small mixing threshold you want to reach. We want to look at the time window where all the mixing happens. This time window has length  $t_{\text{mix}}(\epsilon) - t_{\text{mix}}(1-\epsilon)$ , but by definition of cutoff window, we have that

$$t_{\text{mix}}(\epsilon) - t_{\text{mix}}(1-\epsilon) \leq (C(\epsilon) + C(1-\epsilon))w_n$$

Or in other words  $t_{\text{mix}}^{(n)}(\epsilon) - t_{\text{mix}}^{(n)}(1 - \epsilon) \lesssim w_n \ll t_{\text{mix}}^{(n)}$ . I.e: the relative length of the mixing window becomes negligible as  $n \rightarrow \infty$ .



# Chapter 7

## $\mathcal{L}^p$ distance

Recall from our equivalent characterisations of TV distance, that the TV distance between two distributions  $\mu$  and  $\nu$ , is nothing but half the  $\mathcal{L}^1$  distance between them. We start by defining the  $\mathcal{L}^p$  norm.

**Definition 7.1** ( $\mathcal{L}^p$  norm) Let  $\pi$  be a probability distribution on  $\Omega$  and  $f : \Omega \rightarrow \mathbf{R}$ . Then we define as usual

$$\|f\|_p \equiv \|f\|_{p,\pi} := \begin{cases} \left( \sum_{x \in \Omega} |f(x)|^p \pi(x) \right)^{1/p} & p \in [1, \infty) \\ \max_{x \in \Omega} |f(x)| & p = \infty \end{cases}$$

**Definition 7.2** (Inner product) Let  $f, g$  be two functions  $\Omega \rightarrow \mathbf{R}$ , then we can also define the inner product

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

We can use these  $\mathcal{L}^p$  norms to talk about the distance from  $P^t(x, \cdot)$  to  $\pi$  in a different way. Namely, we can talk about the generalisation of  $d(t)$ :

$$d_p(t) := \max_{x \in \Omega} \left\| \frac{P^t(x, \cdot)}{\pi(\cdot)} - 1 \right\|_p$$

which equals

$$d_p(t) = \max_{x \in \Omega} \left( \sum_{y \in \Omega} \left| \frac{P^t(x, y)}{\pi(y)} - 1 \right|^p \pi(y) \right)^{1/p} \quad \text{or} \quad d_\infty(t) = \max_{x, y \in \Omega} \left| \frac{P^t(x, y)}{\pi(y)} - 1 \right|$$

(It is quite clear how this generalises  $d(t)$ ). We have now some facts about these distances

**Proposition 7.3**  $d_p(t)$  is increasing in  $p$ .

**Main idea:** As with increasing/decreasing quantities involving expectation-like quantities, we want to use Jensen's inequality. Of course, we may also want to note that

$$d_p(t) = \max_{x \in \Omega} \mathbf{E}_\pi[|P^t(x, Y)/\pi(Y) - 1|^p]^{1/p}$$

*Proof.* We use Jensen's Inequality of course. Suppose  $p < r$ , then  $x \mapsto x^{p/r}$  is concave, so we have

$$\mathbf{E}_\pi[|f(x)|^p] = \mathbf{E}_\pi\left[\left(|f(x)|^r\right)^{p/r}\right] \leq \mathbf{E}_\pi[|f(x)|^r]^{p/r}$$

So taking  $p^{\text{th}}$  roots gives the result. ♡

It is clear as well how  $d_p(t) \leq d_\infty(t)$  for all  $p \in [1, \infty]$ , so we also have the following bounds:

$$2d(t) = d_1(t) \leq d_2(t) \leq d_\infty(t)$$

We can now sensibly define a family of mixing times

**Definition 7.4** ( $\mathcal{L}^p$  mixing times) Let  $p \in [1, \infty]$ . Then we define the  $\mathcal{L}^p$  mixing time by

$$t_{\text{mix}}^{(p)}(\epsilon) = \inf\{t \geq 0 : d_p(t) \leq \epsilon\}.$$

**Proposition 7.5** (Relating  $d_2$  and  $d_\infty$  for reversible chains) For a reversible Markov chain, we have

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

**Main idea:** We use inner product algebra to easily show the second equality, and we use Cauchy Schwarz to show that  $d_\infty(2t) \leq d_2(t)^2$ . We then observe that the thing on the right hand side is trivially less than or equal to  $d_\infty(2t)$  and so we are done.

*Proof.* Recall that we may think of the  $\mathcal{L}^p$ -distances as  $d_2(t) = \max_x \|q_t(x, \cdot) - 1\|_2$  where  $q_t(x, y) =$

$P^t(x, y)/\pi(y)$ . Now:

$$\begin{aligned}\|q_t(x, \cdot) - 1\|_2^2 &= \langle q_t(x, \cdot) - 1, q_t(x, \cdot) - 1 \rangle_\pi \\ &= \langle q_t(x, \cdot), q_t(x, \cdot) \rangle_\pi - 2 \langle q_t(x, \cdot), 1 \rangle_\pi + 1 \\ &\stackrel{(!)}{=} q_{2t}(x, x) - 1\end{aligned}$$

Where (!) comes from the fact that  $\langle q_t(x, \cdot), 1 \rangle_\pi$  can be immediately seen to be 1, and  $\langle q_t(x, \cdot), q_t(x, \cdot) \rangle$  can be expanded as an inner product, and reversing one of the matrices, we get that it is equal to  $q_{2t}(x, x)$ . Maximising over  $x$  gives the second inequality. Now we can in some sort reuse this argument, and we have that

$$\begin{aligned}|q_{2t}(x, y) - 1| &= |\langle q_t(x, \cdot) - 1, q_t(y, \cdot) - 1 \rangle| \\ &\leq \|q_t(x, \cdot) - 1\|_2 \|q_t(y, \cdot) - 1\|_2\end{aligned}$$

Maximising over  $x$  and  $y$  gives that  $d_\infty(2t) \leq (d_2(t))^2$ . But of course, it is trivial to see that

$$\max_x \frac{P^{2t}(x, x)}{\pi(x)} - 1 \leq \max_{x, y} \left| \frac{P^{2t}(x, y)}{\pi(y)} - 1 \right| = d_\infty(2t)$$

Therefore we have that

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

so everything is an equality. ♡

As an immediate corollary:

**Corollary 7.6** (TV,  $\mathcal{L}^2$ , and  $\mathcal{L}^\infty$  for a reversible chain) For a reversible chain, we have that

$$t_{\text{mix}}(\epsilon) \leq t_{\text{mix}}(\epsilon/2) = t_{\text{mix}}^{(1)}(\epsilon) \leq t_{\text{mix}}^{(2)}(\epsilon) \leq t_{\text{mix}}^{(\infty)}(\epsilon) \leq 2 t_{\text{mix}}^{(2)}(\sqrt{\epsilon})$$

Moreover,  $t_{\text{mix}}^{(2)}(\epsilon) = \lceil \frac{1}{2} t_{\text{mix}}^{(\infty)}(\epsilon^2) \rceil$





# Chapter 8

## Spectral Decomposition

We now focus on observing the link between mixing times, and eigenfunctions. Note that since vectors in  $\mathbf{R}^\Omega$  can be thought of as functions  $\Omega \rightarrow \mathbf{R}$ , we use the term eigenfunction and eigenvector interchangeably. From now on we will assume chains are reversible, that is to say for all  $x, y \in \Omega$ ,  $\pi(x)P(x, y) = \pi(y)P(y, x)$ .

**Theorem 8.1 (Spectral decomposition of reversible chains)** Let  $P$  be a reversible chain with respect to  $\pi$ , then

- The inner product space  $(\mathbf{R}^\Omega, \langle \cdot, \cdot \rangle_\pi)$  admits an orthonormal basis of real-valued eigenfunctions (of  $P$ )  $\{f_j\}_{j=1}^{|\Omega|}$  corresponding to real eigenvalues  $\{\lambda_j\}$ .
- The matrix  $P$  can be decomposed as

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

- The eigenfunction  $f_1$  corresponding to the eigenvalue 1 (note that 1 is an eigenvalue for the function  $\pi$  by definition), can be taken to be the constant vector  $f = (1, \dots, 1)^T$ . In which case:

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

**Main idea:** Everything follows once we see that we have a matrix  $P$  that satisfies  $\pi(x)P(x, y) = \pi(y)P(y, x)$  and we think of how we can get a symmetric matrix out of this in order to use the spectral theorem for real symmetric matrices. Said matrix can be

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

*Proof.* We will start by defining the matrix  $A(x, y) = \frac{\sqrt{\pi(x)}}{\sqrt{\pi(y)}} P(x, y)$ . Reversibility means that  $\pi(x)P(x, y) = \pi(y)P(y, x)$  so

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

Symmetric matrices admit an orthonormal basis  $\{\phi_j : j \in [|\Omega|]\}$  of eigenfunctions with real eigenvalues  $\lambda_j$ . We check that  $\sqrt{\pi}$  is an eigenfunction of  $A$  with eigenvalue 1:

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

Where we have used the definition of  $A$ , as well as the fact that  $P$  is a stochastic matrix. We label  $\phi_1(x) := \sqrt{\pi(x)}$  and  $\lambda_1 := 1$ . Notice that if  $D_\pi$  is defined as a diagonal matrix with  $D_\pi(x, x) = \pi(x)$ , then it is easy to see that  $A = D_\pi^{1/2} P D_\pi^{-1/2}$ . Now we can use this to note that

$$P(D_\pi^{-1/2} \phi_j) = D_\pi^{-1/2} A \phi_j = \lambda_j D_\pi^{-1/2} \phi_j$$

Showing that  $D_\pi^{-1/2} \phi_j$  is an eigenfunction of  $P$  with eigenvalue  $\lambda_j$ . Notice that this implies that  $\mathbf{1}$  is an eigenfunction of  $P$  with eigenvalue 1, which is obviously true by  $P$  being stochastic, as we would have expected. We now verify that  $\{f_j\}$  is indeed orthonormal with respect to the inner product  $\langle \cdot, \cdot \rangle_\pi$ :

$$\langle f_i, f_j \rangle_\pi = \sum_x f_i(x) f_j(x) \pi(x) = \sum_x \frac{\phi_i(x)}{\sqrt{\pi(x)}} \frac{\phi_j(x)}{\sqrt{\pi(x)}} \pi(x) = \sum_x \phi_i(x) \phi_j(x) = \delta_{i,j}$$

Where in this last equality we have used the fact that  $\{\phi_i\}$  formed an orthonormal basis with respect to the usual inner product. This proves the first claim. Now we set off to prove the desired decomposition:

To decompose  $P^t$ , we use the trick that  $P^t(x, y) = (P^t \delta_y)(x)$ , where  $\delta_y$  should be thought of as  $(0, \dots, 1, 0, \dots, 0)^T$  with the 1 at the  $y^{\text{th}}$  position, therefore  $P^t \delta_y$  is the  $y^{\text{th}}$  column of  $P^t$  and then we can easily pick out the  $x^{\text{th}}$  entry of this vector. We can now decompose  $\delta_y$  as a sum of our eigenfunctions, for they establish a basis:

$$\delta_y = \sum_j \langle \delta_y, f_j \rangle_\pi f_j = \sum_j \left( \sum_i \delta_y(i) f_j(i) \pi(i) \right) f_j = \sum_j f_j(y) \pi(y) f_j$$

Now we put this all together and using the fact that  $f_j$  is an eigenfunction of  $P^t$  with eigenvalue

$\lambda^t$ :

$$\begin{aligned}
 P^t(x, y) &= (P^t \delta_y)(x) = \sum_z P^t(x, z) \delta_y(z) = \sum_z \sum_j P^t(x, z) f_j(y) \pi(y) f_j(z) \\
 &= \sum_j \left( \sum_z P^t(x, z) f_j(z) \right) f_j(y) \pi(y) \\
 &= \sum_j \lambda^t f_j(x) f_j(y) \pi(y)
 \end{aligned}$$

Dividing through by  $\pi(y)$  gives the desired decomposition



This decomposition is nice, because in this basis, we can find a very elegant expression for  $P^t f$  for any  $f : \Omega \rightarrow \mathbf{R}$ :

**Corollary 8.2** Let  $f : \Omega \rightarrow \mathbf{R}$ . Then

$$P^t f = \sum_j \lambda_j^t \langle f, f_j \rangle f_j$$

**Main idea:** This is a rather brainless computation in which one has to start from the definition of  $P^t f$  and then use the decomposition of  $P$  we have just proven above.

*Proof.* We start by decomposing

$$f(x) = \sum_i \langle f, f_i \rangle f_i(x)$$

And now

$$\begin{aligned}
 (P^t f)(x) &= \sum_y P^t(x, y) f(y) \stackrel{(1)}{=} \sum_y \lambda^t \left( \sum_j f_j(x) f_j(y) \pi(y) \right) \left( \sum_i \langle f, f_i \rangle f_i(y) \right) \\
 &= \sum_i \lambda^t \langle f, f_i \rangle \sum_j f_j(x) \underbrace{\left( \sum_y f_j(y) f_i(y) \pi(y) \right)}_{\langle f_j, f_i \rangle_\pi} \\
 &= \sum_i \lambda^t \langle f, f_i \rangle \left( \sum_j f_j(x) \delta_{i,j} \right) \\
 &= \sum_i \lambda^t \langle f, f_i \rangle f_i(x)
 \end{aligned}$$



**Remark 8.3** We know from Perron's Theorem that all eigenvalues of a stochastic matrix are in  $[-1, 1]$ , and moreover, if we assume the chain is lazy, it is a fact that  $-1$  is not an eigenvalue, therefore when our chain is reversible, using the Theorem above, we have that

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

and so as  $t \rightarrow \infty$ , the big sum will disappear, because the powers of the eigenvalues will go to zero. This gives another proof of the Ergodic Theorem.

## 8.1 Examples

**Example 8.4** For a state  $x \in \Omega$ , let  $T(x) = \{t : P^t(x, x) > 0\}$ . Show that  $T(x) \subseteq 2\mathbb{Z}$  if and only if  $-1$  is an eigenvalue.

*Proof.* Suppose that  $-1$  is an eigenvalue with eigenfunction  $f$ . Let  $x = \operatorname{argmax}_{y \in \Omega} |f(y)|$ . Then we have that

$$|f(x)| = |Pf(x)| \leq \sum_y P(x, y) |f(y)| \leq |f(x)|.$$

From which we gather, that whenever  $P(x, y) > 0$ , we must have that  $|f(y)| = |f(x)|$ . We could repeat this argument with  $P^t$  instead of  $P$  and then the eigenvalue would have been  $(-1)^t$ . And since the chain is irreducible, there is some  $t$  such that  $P^t(x, y) > 0$  for all  $y$ . This tells us that for all  $y \in \Omega$ ,  $|f(y)| = |f(x)|$ . For simplicity, suppose without loss of generality, that  $|f(x)| = 1$ . Then we can partition  $\Omega$  into two disjoint sets  $A$  and  $B$ , namely

$$A = \{x : f(x) = 1\}$$

and  $B$  the same but with  $-1$ . We now show that the chain jumps from  $A$  to  $B$  and from  $B$  to  $A$  successively, thus showing that  $P^t(x, x) > 0$  only if  $t \in 2\mathbb{Z}$ . Let  $x \in A$ , then

$$-1 = (Pf)(x) = \sum_{x, y} P(x, y) f(y)$$

since  $f(y) \in \{+1, -1\}$ , it must be that whenever  $P(x, y) > 0$ ,  $y \in B$ . Similarly we can show that if  $x \in B$ , then the next step of the chain will be in  $A$ . Now we can do the converse. Much like we did in Remark 3.9, if for some  $x$ ,  $\gcd\{T(x)\} = 2m$  for some  $m$ , then due to irreducibility, all states have period  $2m$ . Which means we can partition the states-space into  $2m-1$  components, such that when you are on component  $C_i$ , the next step must be on component  $C_{i+1}$ . We can

then make  $f(x)$  take alternating values  $\pm 1$  on consecutive components. Which will give  $-1$  as an eigenvalue. ♡

**Example 8.5** Let  $P$  be a reversible transition matrix. Show that

$$\left| \frac{P^t(x, x)}{\pi(x)} - 1 \right| \leq \sqrt{\left( \frac{P^t(x, x)}{\pi(x)} - 1 \right) \left( \frac{P^t(y, y)}{\pi(y)} - 1 \right)}$$

*Proof.* By using spectral decomposition and Cauchy-Schwarz, and then spectral decomposition in reverse:

$$\begin{aligned} \left| \frac{P^t(x, x)}{\pi(x)} - 1 \right| &= \left| \sum_{i \geq 2} (\lambda_i^t)^{1/2} f_i(x) (\lambda_i^t)^{1/2} f_i(y) \right| \\ &\leq \sqrt{\left( \frac{P^t(x, x)}{\pi(x)} - 1 \right) \left( \frac{P^t(y, y)}{\pi(y)} - 1 \right)} \end{aligned}$$

♡



# Chapter 9

## The Relaxation Time

For a reversible transition matrix  $P$ , we know by Perron's Theorem, that we can label the eigenvalues in decreasing order

$$1 = \lambda_1 > \lambda_2 \geq \dots \geq \lambda_{|\Omega|} \geq -1$$

And by defining  $\lambda^* = \max\{|\lambda| : \lambda \text{ is an eigenvalue different from } 1\}$ , we can define

**Definition 9.1 (Spectral gaps)** The difference  $\gamma^* = 1 - \lambda^*$  is the absolute spectral gap. The difference  $\gamma = 1 - \lambda_2$  is the spectral gap.

**Definition 9.2 (Relaxation time)** The relaxation time  $t_{\text{rel}}$  is defined as

$$t_{\text{rel}} = \frac{1}{\gamma^*}$$

We now prove an upper and lower bound on the mixing time for reversible chains in terms of the relaxation time and the stationary distribution of the chain. Here's in plain English why these results should morally be true: from the spectral decomposition we have shown, we know that

$$P^t(x, y) = \pi(y) \left( 1 + O(\lambda_2^t) + O(\lambda_3^t) + \dots \right)$$

Therefore, assuming say that  $-1$  is not an eigenvalue for argument's sake, we see that the speed of convergence to  $\pi$  is in some sense limited by the largest of the eigenvalues, since the small eigenvalues will decay very quickly. The absolute spectral gap quantifies how much the eigenvalue 1 dominates the other eigenvalues. A big spectral gap will mean that the convergence will happen fast, so the "relaxation time" will be small. In some sense, the relaxation time quantifies how long the chain takes to forget its initial distribution say, and starts "relaxing" into its invariant distribution.

**Theorem 9.3 (Upper and Lower bounds on mixing time in terms of relaxation time)** Let  $P$  be the transition matrix of a reversible, irreducible Markov chain and let  $\pi_{\min} := \min_x \pi(x)$ . Then

$$t_{\text{mix}}(\epsilon) \leq \log\left(\frac{1}{\epsilon \pi_{\min}}\right) t_{\text{rel}}$$

Moreover, if the chain is aperiodic, then

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

To prove these bounds, although we could prove them directly, we go on a slight detour and explore some other bounds first, which will reinforce the intuition given about spectral gaps and relaxation times, and give a more general overview, particularly in terms of an arbitrary starting distribution  $\nu$ :

**Theorem 9.4 (Poincaré Inequality)** Let  $P$  be a reversible matrix with respect to the invariant distribution  $\pi$ . Then for any starting distribution  $\nu$ :

$$\|\mathbf{P}_\nu(X_t = \cdot) - \pi\|_2 \leq (1 - \gamma^*)^t \|\nu - \pi\|_2 \leq \exp\left(-\frac{t}{t_{\text{rel}}}\right) \|\nu - \pi\|_2$$

Once again, note that if  $\gamma^*$  is huge, i.e: the stationary distribution dominates as an eigenfunction, then  $1 - \gamma^*$  is very small so as  $t$  grows, then the distance between the current distribution of the chain and the invariant distribution gets crushed. The proof of this inequality is just a bunch of algebra that uses the spectral decomposition we found above

*Proof.* We start with the definition of the quantity of interest, as as is usual, we square it for convenience:

$$\begin{aligned} \|\mathbf{P}_\nu[X_t \in \cdot] - \pi\|_2^2 &:= \sum_y \left\{ \frac{\mathbf{P}_\nu[X_t = y]}{\pi(y)} - 1 \right\}^2 \pi(y) \\ &= \left( \sum_y \frac{\mathbf{P}_\nu[X_t = y]^2}{\pi(y)} \right) - 1 \end{aligned}$$



And now we study this

$$\begin{aligned}
\left( \sum_y \frac{\mathbf{P}_\nu[X_t = y]^2}{\pi(y)} \right) &= \sum_y \frac{1}{\pi(y)} \left( \sum_x \nu(x) \mathbf{P}_x[X_t = y] \right)^2 \\
&= \sum_y \frac{1}{\pi(y)} \left( \sum_x \nu(x) \mathbf{P}_x[X_t = y] \right) \left( \sum_{x'} \nu(x') \mathbf{P}_{x'}[X_t = y] \right) \\
&= \sum_y \frac{1}{\pi(y)} \left( \sum_x \nu(x) P^t(x, y) \right) \left( \sum_{x'} \nu(x') P^t(x', y) \right) \\
&= \sum_y \frac{1}{\pi(y)} \sum_{x, x'} \nu(x) \nu(x') \left( \sum_{i, j} \lambda_i^t \lambda_j^t \pi(y)^2 f_i(x) f_i(y) f_j(x') f_j(y) \right) \\
&= \sum_{x, x'} \nu(x) \nu(x') \sum_{i, j} \lambda_i^t \lambda_j^t f_i(x) f_j(x') \sum_y f_i(y) f_j(y) \pi(y) \\
&= \sum_{x, x'} \nu(x) \nu(x') \sum_{i, j} \lambda_i^t \lambda_j^t f_i(x) f_j(x') \delta(i, j) \\
&= \sum_{x, x'} \nu(x) \nu(x') \sum_i \lambda_i^{2t} f_i(x) f_i(x') \\
&= 1 + \sum_{x, x'} \nu(x) \nu(x') \sum_{i=2}^{|\Omega|} \lambda_i^{2t} f_i(x) f_i(x') \\
&\leq 1 + \lambda_*^{2t} \sum_{x, x'} \nu(x) \nu(x') \sum_{i=2}^{|\Omega|} f_i(x) f_i(x')
\end{aligned}$$

Now note that

$$\begin{aligned}
\frac{P^0(x, x')}{\pi(x')} &= \sum_i \lambda_i^0 f_i(x) f_i(x') \\
&= 1 + \sum_{i \geq 2} f_i(x) f_i(x')
\end{aligned}$$

and so

$$\sum_{i \geq 2} f_i(x) f_i(x') = \frac{\text{Id}(x, x')}{\pi(x')} - 1$$

Returning to the main computation we continue with:

Finally we have that

$$\begin{aligned}\|\mathbf{P}_\nu[X_t \in \cdot] - \pi\|_2^2 &\leq \lambda_*^{2t} \|\nu - \pi\|_2^2 \\ &= (1 - \gamma^*)^{2t} \|\nu - \pi\|_2^2 \\ &\leq \exp\left(-\frac{2t}{t_{\text{rel}}}\right) \|\nu - \pi\|_2^2\end{aligned}$$

Taking square roots finishes the proof. We have used the inequality  $1 - x \leq \exp(-x)$

♡

And the proof of the Poincaré inequality can be easily modified to give

**Lemma 9.5** Let  $P$  be a reversible chain, and using the usual convention for eigenvalues, we have that

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

*Proof.* Repeat the proof of Poincaré's inequality with  $\nu = \delta_x$ .

♡

Armed with these inequalities, we now prove the goal of this section

**Theorem 9.6 (Upper bound on mixing time in terms of relaxation time)** Let  $P$  be reversible with respect to the invariant distribution  $\pi$  and let  $\pi_{\min} := \min_x \pi(x)$ . Then for all  $\epsilon \in (0, 1)$  we have that

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

**Main idea:** The proof follows almost immediately after bounding  $\mathcal{L}^\infty$  distance by some function of  $\mathcal{L}^2$  distance (which we know we can do by Proposition 7.5), and then applying Poincaré's inequality with a little bit of algebra.

*Proof.* Since TV distance is proportional to  $\mathcal{L}^1$  distance and  $\mathcal{L}^p$  norms are monotonic, then it suffices to prove the second inequality. Moreover, we know from Proposition 7.5, that

$$t_{\text{mix}}^{(\infty)}(\epsilon) \leq 2 t_{\text{mix}}^{(2)}(\sqrt{\epsilon})$$

so it suffices to show that

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

But now applying a direct consequence of Poincaré's Inequality:

$$\begin{aligned}\|P^t(x, \cdot) - \pi\|_2 &\leq \exp\left(-\frac{t}{t_{\text{rel}}}\right) \|\mathbf{1}_x - \pi\|_2 \leq \exp\left(-\frac{t}{t_{\text{rel}}}\right) \left(\frac{1}{\pi(x)} - 1\right)^{1/2} \\ &\leq \exp\left(-\frac{t}{t_{\text{rel}}}\right) \frac{1}{\sqrt{\pi(x)}} \\ &\leq \exp\left(-\frac{t}{t_{\text{rel}}}\right) \frac{1}{\sqrt{\pi_{\min}}}\end{aligned}$$

Now choose  $t^* = \frac{1}{2} t_{\text{rel}} \log\left(\frac{1}{\epsilon \pi_{\min}}\right)$  and the above inequality shows that  $t_{\text{mix}}^{(2)}(\sqrt{\epsilon}) < t^*$  and thus proves the Theorem.  $\heartsuit$

We now have the lower bound:

**Theorem 9.7** If moreover we assume that the chain is aperiodic, then  $t_{\text{mix}}(\epsilon) \geq (t_{\text{rel}} - 1) \log\left(\frac{1}{2\epsilon}\right)$

**Remark 9.8 (Aperiodicity?)** Why do we need aperiodicity? If we don't assume aperiodicity, one of the background results shows that  $-1$  could be an eigenvalue, which means that  $t_{\text{rel}}$  could be infinite. Of course, aperiodicity also implies that there is convergence to the stationary distribution so hence why we need  $t_{\text{rel}}$  to be finite.

We now prove the Theorem

**Main idea:** The key here is that since  $\sum_y f(y) \pi(y) = 0$ , one can rewrite cleverly  $|\lambda^t f(x)|$  for any  $x \in \Omega$  and eigenvalue  $\lambda$  and show that  $|\lambda^t f(x)| \leq 2 \|f\|_{\infty} d(t)$ . Then since state space is finite, there is some  $x^*$  for which  $|f(x^*)| = \|f\|_{\infty}$ , and so we have that for any eigenvalue  $\lambda$ :  $|\lambda|^t \leq 2d(t)$ . Now in particular we can choose  $|\lambda| = \lambda^*$  and rearrange.

*Proof.* Suppose  $f$  is an eigenfunction whose eigenvalue  $\lambda$  is not 1. Since eigenfunctions are orthogonal with respect to  $\langle \cdot, \cdot \rangle_{\pi}$ , and  $\mathbf{1}$  is different eigenfunction to  $f$ , it follows that

$$\sum_y \pi(y) f(y) = \langle \mathbf{1}, f \rangle_{\pi} = 0$$

Thus

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

State space being finite means that for some  $x^*$  we have that  $|f(x^*)| = \|f\|_{\infty}$  and so using that  $x^*$  in the above expression yields

$$|\lambda|^t \leq 2d(t)$$

or in other words,

$$|\lambda|^{t_{\text{mix}}(\epsilon)} \leq 2\epsilon$$

Rearranging gives

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

Since generally we have that  $\log(1+u) \leq u$ , we see that

$$t_{\text{mix}}(\epsilon) \left(\frac{1}{|\lambda|} - 1\right) \geq \log\left(\frac{1}{2\epsilon}\right)$$

In other words:

$$t_{\text{mix}}(\epsilon) \geq \log\left(\frac{1}{2\epsilon}\right) \frac{|\lambda|}{1-|\lambda|}$$

Since this worked for all  $\lambda \neq 1$ , it also works for  $\lambda^*$ , and so we have that

$$t_{\text{rel}}(\epsilon) \geq \log\left(\frac{1}{2\epsilon}\right) \frac{\lambda^*}{1-\lambda^*} = \log\left(\frac{1}{2\epsilon}\right) \frac{\lambda^* - 1 + 1}{\gamma^*} = \log\left(\frac{1}{2\epsilon}\right) (t_{\text{rel}} - 1)$$

♡

**Corollary 9.9** ( $\lambda_*$  as the limit of  $d(t)^{1/t}$  for ergodic chains) For a reversible, irreducible and aperiodic chain, we have that

$$\lim_{t \rightarrow \infty} d(t)^{1/t} = \lambda_*$$

*Proof.* From the proof of the previous Theorem, we know that  $|\lambda|^t \leq 2d(t)$ , so this gives that for all  $t$ ,  $d(t)^{1/t} 2^{1/t} \geq \lambda_*$ . Thus  $\liminf d(t)^{1/t} \geq \lambda_*$ . On the other hand we use the monotonicity of the  $\mathcal{L}^p$  norm, and the last lines of the proof of the Poincaré inequality

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

Indeed: from the proof of the Poincaré inequality we had that

$$\begin{aligned} d_2(t) &= \max_x \|P^t(x, \cdot) - \pi\|_2 \\ &= \max_x \|\mathbf{P}_{\delta_x}[X_t \in \cdot] - \pi\|_2 \\ &\leq \max_x (1-\gamma_*)^t \|\delta_x - \pi\|_2 \end{aligned}$$

And

$$\|\delta_x - \pi\|_2^2 \leq \sum_y \frac{\delta_x(y)^2}{\pi(y)} = \frac{1}{\pi(x)}$$

This gives the inequality claimed above, and so taking  $t^{th}$  roots, and taking limsup finishes the claim. ♥

## 9.1 A necessary condition for cutoff

We now use an application of the bounds derived from spectral theory to show a necessary condition for cutoff. Recall that a sequence  $\{X^{(n)}\}$  of Markov Chains exhibits cutoff if given any  $\epsilon \in (0, 1)$  we have that

$$\lim_{n \rightarrow \infty} \frac{t_{\text{mix}}^{(n)}(\epsilon)}{t_{\text{mix}}^{(n)}(1 - \epsilon)} = 1$$

It is obvious that if cutoff holds, then the following weaker condition also must hold:

**Definition 9.10** (Pre-cutoff ) A sequence of Markov chains  $(X_n)$  exhibits pre-cutoff if

$$\sup_{\epsilon \in (0, 1/2)} \limsup_{n \rightarrow \infty} \frac{t_{\text{mix}}^{(n)}(\epsilon)}{t_{\text{mix}}^{(n)}(1 - \epsilon)} < \infty$$

We now state a condition that unless true, will prevent pre-cutoff and hence cutoff from happening:

**Definition 9.11** (Product condition ) A sequence of reversible Markov chains  $(X_n)$  exhibits the product condition if

$$t_{\text{mix}}^{(n)} \gg t_{\text{rel}}^{(n)}$$

**Proposition 9.12** (No cutoff unless product condition holds) Let  $\{X^{(n)}\}$  be a sequence of reversible aperiodic Markov chains with mixing times  $t_{\text{mix}}^{(n)}$  and relaxation times  $t_{\text{rel}}^{(n)}$ . Suppose that  $t_{\text{mix}}^{(n)} \rightarrow \infty$  as  $n \rightarrow \infty$  but  $t_{\text{rel}}^{(n)} / t_{\text{mix}}^{(n)} \not\rightarrow 0$ . Then there is no pre-cutoff, and therefore no cutoff.

*Proof.* From one of the upper bounds derived from spectral theory, we know that

$$t_{\text{mix}}^{(n)}(\epsilon) \geq (t_{\text{rel}}^{(n)} - 1) \log\left(\frac{1}{2\epsilon}\right)$$

Dividing through by  $t_{\text{mix}}$  we have that

$$\frac{t_{\text{mix}}^{(n)}(\epsilon)}{t_{\text{mix}}} \geq \left( \frac{t_{\text{rel}}^{(n)}}{t_{\text{mix}}} - \frac{1}{t_{\text{mix}}} \right) \log\left(\frac{1}{2\epsilon}\right)$$

Since we assume there is no product condition, we have that  $t_{\text{rel}}^{(n)} / t_{\text{mix}}^{(n)}$  does not converge to zero, that is to say, there exists some constant  $c > 0$  and some infinite subset  $J \subseteq \mathbb{N}$  such that for

all  $n \in J$ , one has that

$$\frac{t_{\text{rel}}^{(n)}}{t_{\text{mix}}^{(n)}} > c$$

Moreover, since  $1/t_{\text{mix}}^{(n)} \rightarrow 0$ , we have that

$$\limsup_{n \rightarrow \infty} \frac{t_{\text{mix}}^{(n)}(\epsilon)}{t_{\text{mix}}^{(n)}} \geq c \log\left(\frac{1}{2\epsilon}\right)$$

But also  $t_{\text{mix}}^{(n)}(1-\epsilon) \leq t_{\text{mix}}^{(n)}$  for  $\epsilon \in (0, 1/2)$  so we have that

$$\limsup_{n \rightarrow \infty} \frac{t_{\text{mix}}^{(n)}(\epsilon)}{t_{\text{mix}}^{(n)}(1-\epsilon)} \geq c \log\left(\frac{1}{2\epsilon}\right)$$

But now taking the supremum of  $\epsilon$  over  $(0, 1/2)$  gives that the right hand side  $\rightarrow \infty$  and as such we have no pre-cutoff and as such no cutoff. ♥

**Remark 9.13** Recall from the upper bound on mixing time

$$t_{\text{mix}}(\epsilon) \leq t_{\text{rel}} \log\left(\frac{1}{\epsilon \pi_{\min}}\right) \equiv t_{\text{rel}} C(\epsilon)$$

that automatically  $t_{\text{mix}}^{(n)}(\epsilon) \lesssim t_{\text{rel}}^{(n)}$ . The result we have just shown says that if we also have  $t_{\text{mix}}^{(n)} \gtrsim t_{\text{rel}}^{(n)}$ , that is to say, if we have  $t_{\text{mix}}^{(n)} \asymp t_{\text{rel}}^{(n)}$ , then we have no cutoff. This gives a relatively straightforward way to see whether a given chain does not exhibit cutoff.

Let us see an example of this result in practice:

**Example 9.14 (No cutoff for a lazy walk on a cycle)** Recall from Example 4.9 that for a lazy simple random walk on  $\mathbf{Z}_n$ , the mixing time is of order

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

Let us show by computing the order of  $t_{\text{rel}}^{(n)}$ , that this chain does not exhibit cutoff.

**Main idea:** We need to compute the order of the relaxation time. To do this, we must evaluate the eigenvalues of the transition matrix, and get the order of the second largest eigenvalue.

*Proof.* We first note that the transition matrix of this chain is

$$P(x, x) = \frac{1}{2} \quad P(x, x \pm 1) = \frac{1}{4}$$

Where addition and subtraction is understood to be modulo  $n$ . Therefore if  $(f, \lambda)$  is an eigenpair, we have that

$$\lambda f(x) = (Pf)(x) = \sum_{y \in \mathbb{Z}_n} P(x, y) f(y) = \frac{1}{2} f(x) + \frac{1}{4} (f(x+1) + f(x-1))$$

Thinking of  $\mathbb{Z}_n$  as  $\{\omega^j : j \in [n]\}$  where  $\omega$  is the  $n^{\text{th}}$  root of unity, gives an easy way to verify that

$$f_k(\omega^j) = \omega^{kj}$$

is an eigenfunction with eigenvalue

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

Indeed:

$$(Pf_k)(j) := (Pf_k)(\omega^j) = \frac{1}{2} \omega^{kj} + \frac{1}{4} \{ \omega^{(k+1)j} + \omega^{(k-1)j} \} = \frac{\omega^{kj}}{2} \{ 1 + \omega^k + \omega^{-k} \} = \lambda_k \omega^{kj} = \lambda_k f_k(j)$$

Of course the largest eigenvalue is  $\lambda_0 = 1$  as expected, and as a cute observation note that all eigenvalues are non-negative, as expected by the fact that the chain is lazy. It is clear that the largest eigenvalues other than  $\lambda_0$  will be  $\lambda_1$  which also is the same as  $\lambda_{n-1}$ . Therefore, to compute the order of the relaxation time, we want to compute the order of  $\lambda_1$ , which is

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

And so  $t_{\text{rel}}^{(n)} \asymp n^2$ , showing that  $t_{\text{rel}} \asymp t_{\text{mix}}$ , and as such showing that this chain cannot exhibit cutoff. ♡

## 9.2 Examples

Let us show how to compute the eigenvalues of the transition matrix of some simple classes of Markov chains:

**Example 9.15 (Walk reflected on the boundaries)** Consider a simple random walk on the  $n$ -vertex segment  $\{v_0, \dots, v_{n-1}\}$  with reflecting boundary conditions, i.e: when the walk is at  $v_0$ , it jumps on the next step to  $v_1$  with probability one, and when it is at  $v_{n-1}$ , it jumps to  $v_{n-2}$  with probability one. Then for  $0 \leq j \leq n-1$ , the functions

$$f_j^\#(v_k) = \cos\left(\frac{\pi j k}{n-1}\right),$$

are eigenfunctions with eigenvalue  $\cos(\pi j / (n-1))$

*Proof.* The key idea is that if we let  $\omega = e^{i\pi/(n-1)}$ , and let  $P$  be the transition matrix of a simple random walk on the  $2(n-1)$  cycle identified with the group  $W_{2(n-1)} = \{\omega, \omega^2, \dots, \omega^{2(n-1)}\}$ , we can introduce an equivalence relation on the state space  $W_{2(n-1)}$  that declares  $\omega^k$  and  $\omega^{-k}$  the same element, and so the random walk  $(X_t)$  on  $W_{2(n-1)}$  can be projected to the random walk  $([X_t])_t$  on  $\{v_0, \dots, v_{n-1}\}$ . With a transition matrix  $P^\#([x], [y]) = \sum_{y \in [y]} P(x, y)$ , where  $x$  is taken as any representative of the coset  $[x]$ . We now make two observations:

- For the simple random walk on the  $n$ -cycle, if  $\omega$  was the  $n^{\text{th}}$  root of unity, then  $\varphi_j(\omega^k) = \exp(2\pi i k / n)$  was seen to be an eigenfunction with eigenvalue  $\cos(2\pi j / n)$ . Since the eigenvalue is real, then it means that both real and imaginary parts of  $\varphi_j$  are eigenvalues, so in particular,  $f_j(\omega^k) = \cos\left(\frac{2\pi j k}{n}\right)$  is an eigenfunction.
- If we define  $f_j^\#([x]) = f_j(x)$  (notice this is well defined since  $\cos(x) = \cos(-x)$ ), then we see that

$$(P^\# f_j^\#)([x]) = \sum_{[y] \in W_{2(n-1)}/\sim} P^\#([x], [y]) = \sum_{[y] \in W_{2(n-1)}/\sim} \sum_{z \in [y]} P(x, z) f_j(z) = \cos\left(\frac{\pi j}{n-1}\right) f_j(x) = \cos\left(\frac{\pi j}{n-1}\right) f_j^\#([x]).$$





# Chapter 10

## Transitive chains

Let us give yet another expression for

$$\|P^t(x, \cdot) - \pi\|_2$$

in terms of eigenvalues, but this time for a specific kind of chains

**Definition 10.1** (Transitive chain) A Markov chain with transition matrix  $P$  is said to be transitive if it looks the same from all points, i.e. for any pair  $(x, y)$  of states, there exists a map  $\phi \equiv \phi_{x,y} : \Omega \rightarrow \Omega$  such that

- $\phi(x) = y$ .
- $P(\phi(z), \phi(w)) = P(z, w)$  for all  $z, w \in \Omega$ .

**Lemma 10.2** (Transitive chains have uniform stationary distributions)

*Proof.* Let  $U(x)$  denote the uniform distribution on  $\Omega$ . Pick any two states  $x$  and  $y$ , let us show that when initialising the chain with distribution  $U$  and doing one step, the mass going into  $x$  and  $y$  is the same. This will then easily show that  $U = \pi$ .

Let  $\phi = \phi_{x,y}$  be the map mentioned above. Then

$$\begin{aligned} \sum_{z \in \Omega} U(z)P(z, x) &= \sum_{z \in \Omega} U(\phi(z))P(\phi(z), \phi(x)) \\ &= \sum_{z \in \Omega} U(\phi(z))P(\phi(z), y) \\ &= \sum_{\omega \in \Omega} U(\omega)P(\omega, y) \end{aligned}$$

Therefore

$$\sum_{x \in \Omega} \sum_{z \in \Omega} U(z) P(z, x) = \sum_x C = C |\Omega|$$

But also, by swapping the order of the sum

$$\sum_{x, z} U(z) P(z, x) = 1$$

so

$$\sum_z U(z) P(z, x) = |\Omega|^{-1} = U(x)$$



**Lemma 10.3** (Transition steps of transitive chains has constant diagonals) Let  $P$  be the transition matrix of a transitive chain, then for any  $n \geq 0$ , any  $x, y \in \Omega$ , we have that  $P^n(x, x) = P^n(y, y)$ .

*Proof.* One shows immediately by induction that for any  $z, w$ , setting  $\phi = \phi_{xy}$  as the "transitive map", we have that  $P^n(z, w) = P^n(\phi z, \phi w)$ , from this we plug  $z = x$  and  $w = x$ .



**Lemma 10.4** (Eigenvalue decomposition for transitive reversible chain) Let  $P$  be reversible and transitive. Then for all  $x \in \Omega$  we have that

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

**Main idea:** Combine the facts that

$$\|P^t(x, \cdot) - \pi\|_2^2 = \sum_{i=2}^{|\Omega|} f_i^2(x) \lambda_i^{2t} \quad \|P^t(x, \cdot) - \pi\|_2^2 = \frac{P^{2t}(x, x)}{\pi(x)} - 1$$

With the fact that transitive matrices have constant diagonals for all powers, as well as  $\pi$  being uniform to see that  $\|P^t(x, \cdot) - \pi\|_2^2$  is independent of  $x$ .

*Proof.* We know from Lemma 9.5 that

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

But also we know that from Proposition 7.5

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

We know that for any  $t$ , we have that  $P^t(x, x)$  is constant for all  $x \in \Omega$ . Moreover, we know that  $\pi$  is uniform, so in fact

$$\|P^t(x, \cdot) - \pi\|_2^2$$

is independent of  $x$ . From this we gather that

$$\frac{1}{n} \sum_{x \in \Omega} \sum_{i=2}^{|\Omega|} f_i^2(x) \lambda_i^{2t} = \sum_{i=2}^{|\Omega|} f_i^2(x) \lambda_i^{2t}$$

But also, rearranging and noting that  $\{f_i\}$  is a collection of orthonormal vectors:

$$\frac{1}{n} \sum_{x \in \Omega} \sum_{i=2}^{|\Omega|} f_i(x) \lambda_i^{2t} = \sum_{i=2}^{|\Omega|} \lambda_i^{2t} \underbrace{\left( \sum_{x \in \Omega} \frac{1}{n} f_i(x) f_i(x) \right)}_{\langle f, f \rangle_\pi = 1}$$

as required. ♡

## 10.1 Examples

**Example 10.5** Let  $X$  be a transitive chain on a  $d$ -regular graph. Let  $\lambda$  be an eigenvalue of the chain of multiplicity one, with corresponding eigenvector  $f$  that has been rescaled such that  $\max_x |f(x)| = 1$ . Show that  $f(x) \in \{+1, -1\}$  for all  $x \in \Omega$  and that there exists some  $p \in \{0, 1, \dots, d\}$  for which

$$\lambda = 1 - \frac{2p}{d}$$

*Proof.* Let  $x \in \Omega$  be some state for which  $f(x) \in \{1, -1\}$  and let  $y \in \Omega$  be given. Let  $\hat{f} = f \circ \phi_{x,y}$ . Note that

$$\begin{aligned} (P\hat{f})(z) &= \sum_w P(z, w) f(\phi(w)) \\ &= \sum_w P(\phi(z), \phi(w)) f(\phi(w)) \\ &= \sum_{w'} P(\phi(z), w') f(w') \\ &= \lambda f(\phi(z)) \end{aligned}$$

since  $\lambda$  is of multiplicity one, it must mean that  $\hat{f}(x) = C f(x)$  for all  $x$ , but since  $\hat{f} = f(\phi(\cdot))$ , it must be that  $\hat{f}(\cdot) = \pm f(\cdot)$  and in particular,  $f(y) = \pm \hat{f}(y) = \pm f(\phi(y)) = \pm f(x) \in \{1, -1\}$ . To show the second claim, without loss of generality suppose  $x \in \Omega$  is such that  $f(x) = 1$ , then

$$\lambda = (Pf)(x) = \sum_{y \sim x} \frac{1}{d} f(y)$$

we know that  $f(y) \in \{1, -1\}$  and thus if  $p$  neighbours of  $x$  are sent to  $+1$  and  $d - p$  are sent to  $-1$ , we'll have that

$$\lambda = 1 - \frac{2p}{d}.$$

♡

## 10.2 Wilson's Method

We now present another method to bound (from below) the mixing times.

**Theorem 10.6 (Wilson's Method)** Let  $(X_t)$  be an irreducible aperiodic Markov chain with state space  $\Omega$  and transition matrix  $P$ . Let  $f$  be an eigenfunction with eigenvalue  $\lambda$  such that

$$1/2 < \lambda < 1$$

Then for a fixed  $\epsilon > 0$  let  $R > 0$  satisfy

$$\mathbf{E}_x [|f(X_1) - f(x)|^2] \leq R \quad \forall x \in \Omega$$

Then

$$t_{\text{mix}}(\epsilon) \geq \max_{x \in \Omega} \left\{ \frac{1}{2 \log(1/\lambda)} \left[ \log \left( \frac{(1-\lambda)f(x)^2}{2R} \right) + \log \left( \frac{1-\epsilon}{\epsilon} \right) \right] \right\}$$

**Example 10.7 (LSRW on Hypercube has cutoff)** Let  $\{X_t^{(n)}\}$  be a Lazy Simple Random Walk on the  $n$ -dimensional hypercube:  $\{0,1\}^n$ . The sequence exhibits cutoff at time  $\frac{1}{2}n \log n$  with a window of order  $n$ . Moreover, the relaxation time is  $t_{\text{rel}} = n$ .

*Exposition and waffle.* We note that the hypercube is an example of a Product Chain (see Appendix), and so we can start by finding eigenfunctions of the "sub-chains", i.e: LSRW on  $\{0,1\}$ , which has transition matrix  $P(x,y) = 1/2$  for all  $x,y \in \{0,1\}$ . Of course, the function  $g_1(x) = 1$  for all  $x$  is an eigenfunction with eigenvalue 1, but then we also have  $g_2(x) = 1 - 2x$ :

$$(Pg_2)(x) = \sum_{y \in \{0,1\}} P(x,y)g_2(y) = \frac{1}{2}(1-1) = 0$$

so  $g_2$  is an eigenfunction with eigenvalue zero. Now using the known results about product chains from the Appendix, we see that

$$f_I = \prod_{i=1}^n f_i(x_i)$$

where each  $f_i$  is  $g_1$  or  $g_2$ , and  $I = \{i : f_i = g_2\}$ , is an eigenfunction of the chain on the hypercube, with eigenvalue

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

It is clear that the invariant distribution of each of the smaller chains is simply  $\pi_j$  a uniform distribution, so its also easy to check that  $\langle g_i, g_j \rangle_{\pi_j} = \delta(i, j)$ , so in fact by another result of the Appendix,

$$B = \left\{ f_I = \prod_{i \in I} g_2(x_i) : I \subseteq \{1, \dots, n\} \right\}$$

are all the orthonormal eigenfunctions forming an orthonormal basis. Note that for  $I = \emptyset$ ,  $\lambda_{\emptyset} = 1$ , and for  $|I| = 1$ , we have  $\lambda_* = \lambda_I = 1 - \frac{1}{n}$ , and so  $t_{\text{rel}} = n$  as required. Recall that using a strong stationary time argument we showed that  $t_{\text{mix}}(\epsilon) \leq \frac{1}{2}n \log n + C(\epsilon)n$ . We will reprove this fact using some results we have just proven and compare different bounds that we obtain using different results.

Since we have computed the relaxation time  $t_{\text{rel}} = n$ , we might as well see what bound we get using this. Recall from Theorem 9.6 that

$$t_{\text{mix}}(\epsilon) \leq n \log \frac{1}{\epsilon \pi_{\min}}$$

It is easy to compute  $\pi_{\min}$  because its obvious that the chain is transitive and so the distribution is uniform. Hence  $\pi_{\min} = 2^{-n}$ . From this we get that  $t_{\text{mix}}(\epsilon) \lesssim n^2$ . But observe that this is

actually a pretty shit bound compared to what we have already obtained before. Now we will see an example of the tradeoff between being lazy and computing just the second eigenvalue to actually computing the full spectrum like we have done. Using a corollary to Poincaré's inequality for transitive chains:

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

So choosing  $t = \frac{n \log n}{2} + Cn$  means we can choose  $C$  large enough and bound TV distance by anything as small as we wish, and we get the same bound we got before. Now we set off to get the lower bound with Wilson's method. If you look at the set  $B$  defined above, you will convince yourself that the functions  $g_i = 1 - 2x_i$  are eigenfunctions of the LSRW on the hypercube each with eigenvalue  $1 - \frac{1}{n}$ , therefore if  $P$  is transition matrix of the LSRW on the hypercube

$$P \left( \underbrace{\sum_{i=1}^n 1 - 2x_i}_{:=f} \right) = \left(1 - \frac{1}{n}\right) \left( \sum_{i=1}^n 1 - 2x_i \right)$$

and since  $n \geq 2$ , we see that this new eigenfunction has eigenvalue  $\geq 1/2$ , so the first checkbox of Wilson's method is ticked. Now we check the second condition of Wilson's method, we need to compute

$$\mathbf{E}_x[|f(X_1) - f(x)|^2]$$

A nice way to compute this is by conditioning on the events that the walk does or doesn't move:

$$\mathbf{E}_x[|f(X_1) - f(x)|^2] = \frac{1}{2} \mathbf{E}_x[(f(X_1) - f(x))^2 \mid \{\text{stays put}\}] + \frac{1}{2} \mathbf{E}_x[(f(X_1) - f(x))^2 \mid \{\text{moves}\}]$$

If the walk stays put, then  $f(X_1) = f(x)$ , and if the walk has moved, what used to be a zero in one of the coordinates has now become a one or vice-versa, which means that the total sum in the definition of  $f$  has changed by  $\pm 2$ , upon squaring this and then multiplying by the  $1/2$ , we see that

$$\mathbf{E}_x[|f(X_1) - f(x)|^2] = 2$$

which is the value of  $R$  we will take. Finally, we note that for  $x = (0, 0, \dots, 0)$  we get that  $f(x) = n$

and so finally plugging into Wilson's method we see that

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

We recall the expansion for log:

$$\log \frac{1}{1-x} = \sum_{n=1}^{\infty} \frac{x^n}{n} \quad |x| < 1$$

and so we have that

$$\log \frac{1}{1-1/n} = \frac{1}{n} + O(1/n^2)$$

Then we have that for large  $n$ ,

$$\frac{1}{\log(1/(1-1/n))} = n + O(1)$$

so putting it all together:

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

(The reason for the minus sign in front of  $C(\epsilon)$  is that for  $\epsilon > 1/2$  the  $\log \frac{1-\epsilon}{\epsilon}$  is negative)

♡





# Chapter 11

## Dirichlet Forms and Bottleneck Ratio

**Definition 11.1** (Dirichlet form ) Let  $P$  be a reversible transition matrix with respect to a stationary distribution  $\pi$ . The Dirichlet form associated to  $(P, \pi)$  is defined for functions  $f$  and  $h$  on  $\Omega$  by

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

We also refer to  $\mathcal{E}(f) := \mathcal{E}(f, f)$ .

**Lemma 11.2** (Alternative characterisation of Dirichlet form ) For all  $P$ , we have that

$$\mathcal{E}(f) = \frac{1}{2} \mathbf{E}_\pi[(f(X_1) - f(X_0))^2]$$

Moreover, if  $P$  is reversible, then

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

**Remark 11.3** Note that

$$\begin{aligned} \mathbf{E}_\pi[(f(X_1) - f(X_0))^2] &= \sum_{x, y} (f(x) - f(y))^2 \mathbf{P}(X_1 = x, X_0 = y) \\ &= \sum_{x, y} (f(x) - f(y))^2 \mathbf{P}(X_1 = x \mid X_0 = y) \mathbf{P}(X_0 = y) \\ &= \sum_{x, y} (f(x) - f(y))^2 P(y, x) \pi(y) \end{aligned}$$

*Proof of Lemma 11.2.* This is just one big computation

$$\begin{aligned}
 \mathcal{E}(f) &= \langle f, f \rangle_\pi - \langle Pf, f \rangle_\pi \\
 &= \sum_x f^2(x)\pi(x) - \sum_{x,y} P(x,y)f(y)f(x)\pi(x) \\
 &= \frac{1}{2} \left\{ \sum_x f^2(x)\pi(x) + \sum_y f^2(y)\pi(y) - 2 \sum_{x,y} f(y)f(x)\pi(x)P(x,y) \right\} \\
 &= \frac{1}{2} \left\{ \sum_{x,y} f^2(x)\pi(x)P(x,y) + \sum_{y,x} f^2(y)\pi(y)P(y,x) - 2 \sum_{x,y} f(y)f(x)\pi(x)P(x,y) \right\} \\
 &= \frac{1}{2} \left\{ \sum_{x,y} (f(x) - f(y))^2 \pi(x)P(x,y) \right\} = \frac{1}{2} \mathbf{E}_x [(f(X_1) - f(X_0))^2]
 \end{aligned}$$

To prove the second statement, we compute the value of  $\mathcal{E}(f, g)$  in two ways:

$$\begin{aligned}
 \mathcal{E}(f, g) &= \sum_x ((I - P)f)(x)g(x)\pi(x) \\
 &= \sum_{x,y} (f(x) - f(y))g(x)P(x,y)\pi(x) \quad (A) \\
 &\stackrel{(!)}{=} \sum_{x,y} (f(y) - f(x))g(y)P(x,y)\pi(x) \quad (B)
 \end{aligned}$$

Where (!) comes from reversing  $\pi(x)P(x,y)$  and then relabelling  $x \leftrightarrow y$ . Once we have calculated  $\mathcal{E}(f, g)$  to be equal to both (A) and (B), we can just declare that

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

♡

As we have seen, the information provided by the spectral gap is fantastic, but calculating the spectral gap explicitly can be hard, unless we are in some particular example where either the full spectrum can be calculated or it is easy to see which eigenvalue corresponds to  $\lambda_*$ , so we want ways to estimate the gap. First we show an alternative way to characterise it, using the Dirichlet form and variances. For notational convenience, we understand any function  $f$  to be a function  $\Omega \rightarrow \mathbf{R}$

**Theorem 11.4 (Variational characterisation of spectral gap)** Let  $P$  be reversible with respect to

$\pi$ . Then the spectral gap  $\gamma$ :

$$\begin{aligned}\gamma &:= 1 - \lambda_2 = \inf \{ \mathcal{E}(f) : \mathbf{E}_\pi[f] = 0, \|f\|_2 = 1 \} \\ &= \inf \left\{ \frac{\mathcal{E}(f)}{\|f\|_2^2} : \mathbf{E}_\pi[f] = 0 \right\} \\ &= \inf \left\{ \frac{\mathcal{E}(f)}{\text{Var}_\pi(f)} : f \text{ non-constant} \right\}\end{aligned}$$

Before proving this characterisation, let us show how this is actually intuitive: recall that  $\mathcal{E}(f)$  measures the average change in  $f$  over one step of the Markov chain when it starts at stationarity. Of course if  $f$  already has high variance, this will already be big by nature, so by dividing  $\mathcal{E}(f)/\text{Var}_\pi(f)$  you are in some sense quantifying the normalised average change of  $f$  over one step of the chain, and once we take the infimum over all  $f$  that are non-constant, we have accounted over all functions on state space, how much do the normalised versions of  $f$  change over 1 step in the Markov chain, this now seems to be more of a property of the chain itself, and a high value of this ratio, indicates that at the very start of the execution of the chain, there are really big changes over all functions on state space, intuitively indicating that the chain is going to forget its initial distribution quite quickly. This connects perfectly with the fact that  $\gamma$  controls the speed of convergence.

**Main idea:** Translation and scaling of the Dirichlet form give the last two equalities. For the first equality, use spectral decomposition and the bound  $1 - \lambda_i \geq 1 - \lambda_2$  to get one equality and then the fact that  $\mathcal{E}(f_2) = \gamma$  to get the other equality.

*Proof.* It is easy to see that since

$$\mathcal{E}(f) = \frac{1}{2} \mathbf{E}_\pi[(f(X_1) - f(X_0))^2]$$

we have that for any constant  $c \in \mathbf{R}$ :  $\mathcal{E}(f + c) = \mathcal{E}(f)$ . Thus using  $f - \mathbf{E}_\pi[f]$  instead of  $f$ , gives the last equality. Similarly, we have that  $\mathcal{E}(cf) = c^2 \mathcal{E}(f)$  gives the second equality. So all there is to show really is the first equality.

Let  $\{f_i\}$  be an orthonormal family of eigenfunctions with respect to  $P$ , so that we can write any  $f$  with  $\mathbf{E}_\pi(f) = 0$

$$f = \underbrace{\langle f, \mathbf{1} \rangle_\pi}_{\mathbf{E}_\pi(f)=0} \mathbf{1} + \sum_{j=2}^n \langle f, f_j \rangle_\pi f_j$$

And plugging inside the Dirichlet form gives

$$\begin{aligned}
 \mathcal{E}(f) &= \left\langle \sum_{j=2}^n \langle f, f_j \rangle_{\pi} (I - P)f_j, f \right\rangle_{\pi} \\
 &= \sum_{j=2}^n \langle f, f_j \rangle_{\pi} \langle (f_j - \lambda_j f_j), f \rangle_{\pi} \\
 &= \sum_{j=2}^n \langle f, f_j \rangle_{\pi} (1 - \lambda_j) \langle f_j, f \rangle_{\pi} \\
 &= \sum_{j=2}^n (1 - \lambda_j) \langle f, f_j \rangle_{\pi}^2 \\
 &\geq (1 - \lambda_2) \underbrace{\sum_{j=2}^n \langle f, f_j \rangle_{\pi}^2}_{\|f\|_2^2 = 1}
 \end{aligned}$$

Thus we have that

$$\gamma \leq \inf \{ \mathcal{E}(f) : \mathbf{E}_{\pi}[f] = 0, \|f\|_2 = 1 \}$$

But noting that  $\gamma = \mathcal{E}(f_2)$  (indeed,  $\mathcal{E}(f_2) = \langle f_2, f_2 \rangle - \langle P f_2, f_2 \rangle = (1 - \lambda_2) \|f_2\|^2 = 1 - \lambda_2$  by orthonormality of the  $\{f_i\}$ 's) and clearly

$$\mathcal{E}(f_2) \geq \inf \{ \mathcal{E}(f) : \mathbf{E}_{\pi}[f] = 0, \|f\|_2 = 1 \}$$

finishes the claim. ♡

We saw above how the Poincaré inequality was absolutely cracked out of its mind, but it only applied to reversible chains. We would like to generalise the inequality to non-reversible chains now. Of course we will not have all that spectral business, but as we've seen, we can express the spectral gap in a variational form, and so we can just extend this to all chains. The idea is to first define the Poincaré constant, which we shall see that in the case of reversible chains is nothing but the spectral gap. With some work, we will replace the spectral gap in Theorem 9.6 with the Poincaré constant and as such generalise the inequality to hopefully a broader class of chains.

**Definition 11.5** (Poincaré constant) Let  $P$  be the transition matrix of a Markov chain, the Poincaré constant  $\gamma(P)$  is defined as

$$\gamma(P) = \inf \left\{ \frac{\mathcal{E}(f)}{\text{Var}_{\pi}(f)} : f \text{ non-constant} \right\}$$

**Remark 11.6** By the above Theorem, it is clear that when  $P$  is reversible, then  $\gamma(P) = \gamma$ . Moreover

Notice that for a transition matrix  $P$ , if  $P^*$  denotes its time reversal, we have that

$$\mathcal{E}_P(f) = \langle (I - P)f, f \rangle_\pi = \|f\|_2^2 - \langle Pf, f \rangle_\pi = \|f\|_2^2 - \langle f, P^*f \rangle = \mathcal{E}_{P^*}(f)$$

And moreover

$$\mathcal{E}_{\frac{P+P^*}{2}}(f) = \|f\|_2^2 - \frac{1}{2} \langle Pf, f \rangle - \frac{1}{2} \langle P^*f, f \rangle = \mathcal{E}_P(f)$$

Therefore we have that the Dirichlet form and the consequently the Poincaré constant are invariant under time reversal and additive symmetrisation, where the additive (and multiplicative) time symmetrisation of  $P$  are

$$\frac{P + P^*}{2} \quad PP^*$$

respectively. We now state the bound on mixing time, directly analogous to the one we obtained using the relaxation time, but that doesn't require reversibility.

**Lemma 11.7 (Spectral interpretation of Poincaré constant)** For any chain  $P$  we have that

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

Where  $\lambda_2(Q)$  denotes the second largest eigenvalue of a reversible chain  $Q$ .

*Proof.* Simply combine the fact that  $\gamma(P) = \gamma\left(\frac{P+P^*}{2}\right)$  (which comes from the fact that the Dirichlet forms agree and then use the variational characterisation of the gap) and then use the fact that for a reversible chain, the Poincaré constant equals the spectral gap. ♡

**Theorem 11.8 (Poincaré bound on mixing time)** Let  $P$  be the transition matrix of a lazy Markov chain. Then for  $\epsilon \in (0, 1)$  we have that

$$t_{\text{mix}}(\epsilon) \leq \frac{2}{\gamma(P)} \log\left(\frac{1}{2\epsilon\sqrt{\pi_{\min}}}\right)$$

And for a general  $P$ :

$$t_{\text{mix}}(\epsilon) \leq \frac{2}{\gamma(PP^*)} \log\left(\frac{1}{2\epsilon\sqrt{\pi_{\min}}}\right)$$

Compare this with Theorem 9.6. To prove this we need the following Lemma obtained from Example Sheet 2.

**Lemma 11.9** Let  $f: \Omega \rightarrow \mathbf{R}$ , then we have that for a transition matrix  $P$ :

$$\text{Var}_\pi(Pf) \leq [1 - \gamma(PP^*)] \text{Var}_\pi(f)$$

Moreover, if  $P$  is lazy  $\gamma(P) \leq \gamma(PP^*)$ .

**Main idea:** The idea is to bound TV distance by an  $\mathcal{L}^2$  distance, which can then be related to the Variance of  $(P^*)^t f_x$ , where  $f_x(z) = \pi^{-1}(x) \delta_x(z)$ . From there use variational contraction property.

*Proof of Theorem 11.8.* We start by noting that due to the monotonicity of  $\ell^p$  norms:

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

And using the fact that

$$\frac{P^t(x, y)}{\pi(y)} = \frac{P^{*t}(y, x)}{\pi(x)} = \sum_z \frac{P^{*t}(y, z) \mathbf{1}_{\{x\}}(z)}{\pi(x)} = (P^{*t} f_x)(y)$$

Where we define  $f_x(z) = \pi^{-1}(x) \mathbf{1}_{\{x\}}(z)$ . Now we compute

$$\begin{aligned} \left\| \frac{P^t(x, \cdot)}{\pi(\cdot)} - 1 \right\|_2^2 &= \sum_y \left( \frac{P^t(x, y)}{\pi(y)} - 1 \right)^2 \pi(y) \\ &= \sum_y \left( \frac{P^t(x, y)}{\pi(y)} \right)^2 \pi(y) - 2 \sum_y \frac{P^t(x, y)}{\pi(y)} \pi(y) + \sum_y \pi(y) \\ &= \sum_y \left( \frac{P^t(x, y)}{\pi(y)} \right)^2 \pi(y) - 1 \stackrel{(!)}{=} \text{Var}_\pi \left( \frac{P^t(x, \cdot)}{\pi(\cdot)} \right) = \text{Var}_\pi(P^{*t} f_x) \end{aligned}$$

(Useful trick to remember to save time!). Step (!) comes from the fact that the expectation of  $P^t(x, Y)/\pi(Y)$  under  $\pi$  is one. Using the variational contraction property of the previous lemma inductively, we have that

$$\text{Var}_\pi(P^{*t} f_x) \leq [1 - \gamma(P^*P)]^t \text{Var}(f_x) = [1 - \gamma(P^*P)]^t \left( \frac{1}{\pi(x)} - 1 \right)$$

Using the general bound that  $1 + x \leq \exp(x)$ , we have that

$$\left\| \frac{P^t(x, \cdot)}{\pi(\cdot)} - 1 \right\|_2^2 \leq \frac{1}{\pi_{\min}} \exp(-t\gamma(PP^*))$$

From this we have that

$$\|P^t(x, \cdot) - \pi\|_{\text{TV}} \leq \frac{1}{2\sqrt{\pi}} \exp\left(\frac{-t}{2} \gamma(P P^*)\right)$$

This gives the bound on the general case of  $P$ . If  $P$  is lazy, we use the second part of the variational contraction Lemma along the fact that  $P^*$  is also lazy (this is easily seen as the diagonals of  $P^*$  and  $P$  agree by definition)



Here's a quick summary of what we have done in the previous chapters:

- In Chapter 8 we saw that if  $P$  is reversible, then it admits an orthonormal basis  $(f_i)_{i \leq n}$  of eigenfunctions with eigenvalues  $1 = \lambda_1 > \lambda_2 \geq \dots \lambda_n \geq -1$ , and that if the chain was aperiodic, then in fact this last inequality was strict because all eigenvalues are positive.

- More importantly, we had decompositions:

- Of the matrix:

$$\frac{P^t(x, y)}{\pi(y)} = 1 + \sum_{i \geq 2} \lambda_i^t f_i(x) f_i(y).$$

- Of functions:

$$(P^t f)(x) = \sum_{i \geq 1} \lambda_i^t \langle f_i, f \rangle f_i(x)$$

- If we let  $\lambda^*$  be the eigenvalue of greatest modulus other than the eigenvalue  $\lambda_1 = 1$ , then we defined  $\gamma^* = 1 - \lambda^*$  as the absolute spectral gap, and we also defined  $\gamma = 1 - \lambda_2$  to be the spectral gap. We defined  $t_{\text{rel}} = 1/\gamma^*$  to be the relaxation time. It turns out that the relaxation time provides powerful insights as to how the mixing time behaves, that is to say, for reversible matrices:

- A general upper bound, consequence of the Poincaré inequality 9.4 and a comparison of  $d_2$  and  $d_\infty$  distances 7.5.

$$t_{\text{mix}}(\epsilon) \leq t_{\text{rel}} \log \left( \frac{1}{\epsilon \pi_{\min}} \right)$$

- And if the chain was aperiodic, we could also get a lower bound:

$$t_{\text{mix}}(\epsilon) \geq (t_{\text{rel}} - 1) \log(1/2\epsilon).$$

- After this we moved on to provide a characterisation of the spectral gap that didn't directly involve the eigenfunctions or eigenvalues. We did so by defining the Dirichlet form  $\mathcal{E}(f)$  of a function, and we saw that

$$\gamma = \inf \{ \mathcal{E}(f) : \mathbf{E}_\pi[f] = 0, \|f\|_2 = 1 \}.$$

- Since this characterisation did not involve any sort of spectral nonsense, we could directly generalise the idea of spectral gap to non-reversible chains, defining  $\gamma(P)$  for a general chain to be the infimum above. We then saw how working directly with this notion of spectral gap and a variational Poincaré inequality, also provides upper bounds on the mixing time that were resembling of those obtained prior, namely that for a general  $P$ :

$$t_{\text{mix}}(\epsilon) \leq \frac{2}{\gamma(P P^*)} \log \left( \frac{1}{2\epsilon \sqrt{\pi_{\min}}} \right).$$



**Remark 11.10 (Where are we heading?)** So far we have seen a quite a few tools to bound mixing times, from above: coupling inequalities, strong stationary times, Markovian couplings, relaxation time... and from below: dirty bounds, relaxation time, Wilson's method, diameter bounds... One other thing of interest to us soon will be to start comparing Markov chains themselves, so that by using known bounds on mixing times of simple Markov chains, we can obtain bounds on modifications of those chains. For this purpose we start talking about the following easy Theorem.

**Theorem 11.11 (Comparison of Poincaré constants)** Let  $P$  and  $P'$  be two chains with forms  $\mathcal{E}, \mathcal{E}'$ , invariant distributions  $\pi$  and  $\pi'$ , and Poincaré constants  $\gamma, \gamma'$ . If there exists some  $A > 0$  such that

$$\mathcal{E}'(f) \leq A\mathcal{E}(f) \quad \forall f : \Omega \rightarrow \mathbf{R}$$

then we have that  $\gamma' \leq \left(\max_{x \in \Omega} \frac{\pi(x)}{\pi'(x)}\right) A\gamma$

**Main idea:** The key principle is that  $\text{Var}(X) = \mathbf{E}[(X - \mathbf{E}X)^2] = \min_{a \in \mathbf{R}} \mathbf{E}[(X - a)^2]$ .

*Proof.* We start by noting the following

$$\begin{aligned} \gamma' &= \inf \left\{ \frac{\mathcal{E}'(f)}{\text{Var}_{\pi'}(f)} : f \text{ non-constant} \right\} \\ &\leq A \inf \left\{ \frac{\mathcal{E}(f)}{\text{Var}_{\pi'}(f)} : f \text{ non-constant} \right\} \quad (*) \end{aligned}$$

So the question becomes, how can we bound  $\text{Var}_{\pi'}(f)$  using  $\text{Var}_{\pi}(f)$ . The key observation is in the main idea section. We simply note that

$$\begin{aligned} \text{Var}_{\pi}(f) &= \mathbf{E}_{\pi}[(f - \mathbf{E}_{\pi}f)^2] \\ &\leq \mathbf{E}_{\pi}[(f - \mathbf{E}_{\pi'}f)^2] \\ &= \sum_x \pi(x) (f(x) - \mathbf{E}_{\pi'}f)^2 \\ &= \sum_x \frac{\pi'(x)}{\pi'(x)} \pi(x) (f(x) - \mathbf{E}_{\pi'}f)^2 \\ &\leq \left( \max_x \frac{\pi(x)}{\pi'(x)} \right) \mathbf{E}_{\pi'}[(f - \mathbf{E}_{\pi'}f)^2] \\ &= \left( \max_x \frac{\pi(x)}{\pi'(x)} \right) \text{Var}_{\pi'}(f) \end{aligned}$$

And so  $\text{Var}_{\pi'}(f) \geq \left(\max_x \frac{\pi(x)}{\pi'(x)}\right)^{-1} \text{Var}_{\pi}(f)$  so substituting into (\*) finishes the claim.



## 11.1 Canonical Paths

We now present our first method of comparing Markov chains. Consider as an illustration the random walk on  $\mathbb{Z}_n^d$ . As we saw before, this is a particularly nice chain to analyse, since by a coupling argument we could find upper bounds relatively easily on its mixing time. Suppose now that we remove some edges from the torus, this irregularity no longer allows us use the coupling method, and since the transition probabilities of this new chain are not comparable to the transition probabilities of the original chain, we cannot hope to compare Dirichlet forms either. This new method will provide an answer:

**Definition 11.12 ( $E$ -Path)** For any states  $x, y \in \Omega$  say there is an edge  $e = (x, y)$  between them if  $P(x, y) > 0$ . We define the set of edges

$$E = \{(x, y) : P(x, y) > 0\}$$

For any pair  $x, y$  we define an  $E$ -path, denoted by  $\Gamma_{x,y}$  (if it exists) a collection of edges

$$\Gamma_{x,y} = \{(x, x_1), (x_1, x_2), \dots, (x_{k-1}, y)\}$$

We define  $|\Gamma_{x,y}| = k$  to be the length of the path. For an edge  $e = (a, b)$  we also say  $Q(e) = \pi(a)P(a, b)$

Let's introduce a preliminary result that shows how the method of considering  $E$ -paths can give us lower bounds on the spectral gap.

**Theorem 11.13 (Canonical Path Method)** For all  $x, y \in \Omega$ , fix an  $E$ -path (if it exists)  $\Gamma_{x,y}$  and let

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

Then  $\gamma(P) \geq 1/B$ .  $B$  is referred to as the congestion ratio. Of course, it depends on the choice of paths.

**Remark 11.14 (Waffling)** The quantity  $Q(e)$  should be understood as the "rate of flow in the long run through the edge  $e$ ", indeed,  $\pi(a)$  tells you the "density" of particles at vertex  $a$  in the long run, and  $P(a, b)$  tells you which fraction of the particles cross from  $a$  to  $b$  in one step, thus  $Q(e)$  is the overall flux of particles through edge  $e$ . That confusing sum can be expressed as

follows. Let  $X$  and  $Y$  be independent and distributed according to  $\pi$ . Then

$$B = \max_{e \in E} \frac{1}{Q(e)} \mathbf{E}[|\Gamma_{XY}| \mathbf{1}(e \in \Gamma_{XY})]$$

is the ratio of the average length of an  $E$ -path that contains  $e$ , with the flux of particles through the edge. If  $B$  is big, one can understand that there is a lot of congestion on average on the graph. According to this waffle, the result of the Theorem is intuitively expected: since  $\gamma \geq 1/B$ , then for the reversible case,  $t_{\text{rel}} \leq B$ , and so if  $B$  is small, meaning there is not a lot of congestion, we will have that the relaxation time is small and so the walk "diffuses quickly".

**Main idea:** Examine  $\text{Var}_{\pi}(f)$  as  $\frac{1}{2} \mathbf{E}[\{f(X) - f(Y)\}^2]$  where  $X$  and  $Y$  are iid random variables distributed according to  $\pi$ , and use Cauchy-Schwarz after decomposing  $\{f(x) - f(y)\}^2$  as  $\left(\sum_{e \in \Gamma_{xy}} \nabla f(e)\right)^2$ .

*Proof.* We start by examining

$$\text{Var}_{\pi}(f) = \frac{1}{2} \mathbf{E}[\{f(X) - f(Y)\}^2] = \frac{1}{2} \sum_x \sum_y \{f(x) - f(y)\}^2 \pi(x) \pi(y)$$

Where  $X, Y \stackrel{iid}{\sim} \pi$ . It is very easy to verify the identity for the variance. Now we have the decomposition (recall that for an edge  $e = (a, b)$ ,  $\nabla f(e) = f(a) - f(b)$ )

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

Plugging in we have that

$$\text{Var}_{\pi}(f) \leq \frac{1}{2} \sum_{x, y \in \Omega} \sum_{e \in \Gamma_{xy}} (\nabla f(e))^2 |\Gamma_{xy}| \pi(x) \pi(y)$$

The important thing to realise now is that we are summing by picking pairs of points  $(x, y)$ , and then summing over all edges in the designated path  $\Gamma_{xy}$  between them. It is clear that doing this will eventually sum over all edges in  $E$ , and it will potentially repeat some of the edges, but this sum is equivalent to first picking an edge  $e \in E$ , and then sum over the pairs  $(x, y)$  whose designated path  $\Gamma_{xy}$  contains  $e$ . (Please see diagram) In other words, we can flip the sum to be

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

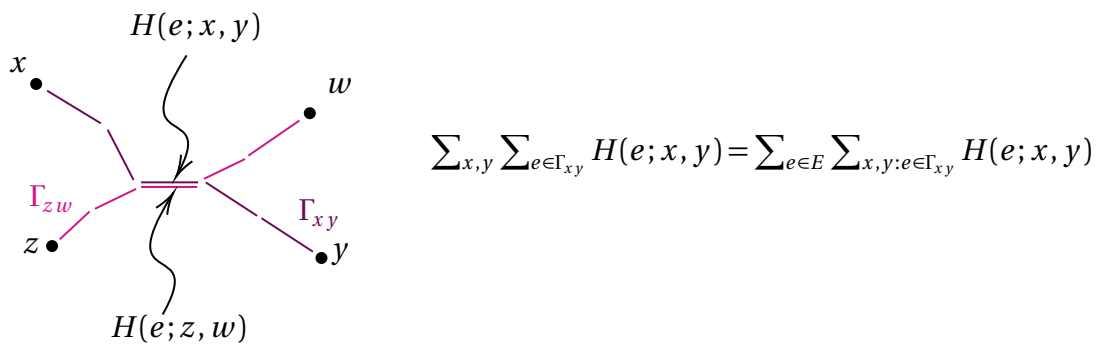


Figure 11.1: The diagram that says it all for Proof of Canonical Path Method

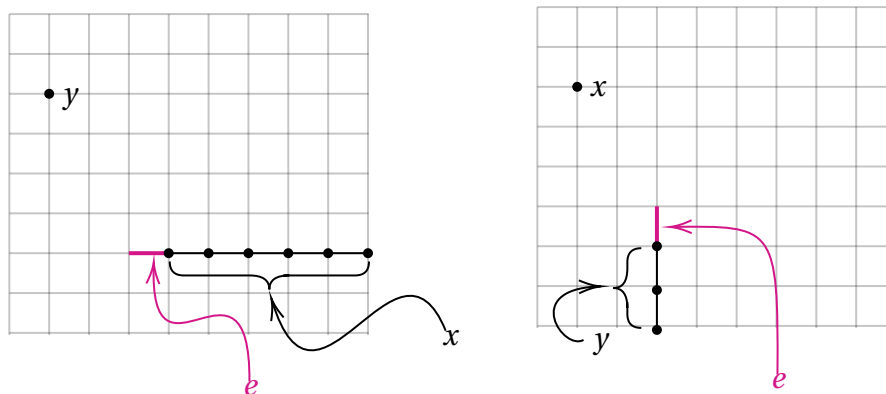


Figure 11.2: The diagram that says it all: Relaxation time for LSRW on box

Where we have used in this last step that

$$\mathcal{E}(f) = \frac{1}{2} \sum_{e \in E} Q(e) (\nabla f(e))^2$$

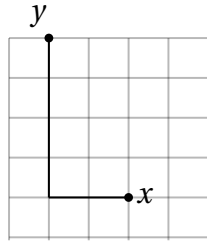
is an alternative characterisation of the Dirichlet form. Now we just observe that

$$\gamma(P) = \inf_{f \text{ n.c.}} \frac{\mathcal{E}(f)}{\text{Var}(f)} \geq \inf_{f \text{ n.c.}} \frac{\mathcal{E}(f)}{\mathcal{E}(f)B} = 1/B$$

♡

**Example 11.15 (Relaxation time of a LSRW on the box)** Let  $X$  be a LSRW on the box  $\{1, \dots, n\}^d$ . Then there exists a constant  $C > 0$  such that  $t_{\text{rel}} \leq C(dn^2)$ .

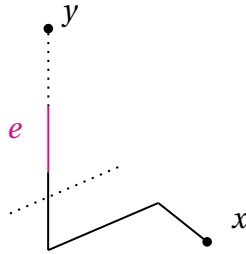
*Proof.* First we say what our canonical paths are. For a pair of points  $(x, y)$  in the box, the path  $\Gamma_{xy}$  will be to first move along the first coordinate until it becomes the first coordinate of  $y$ , and so on with all coordinates. Here's an example



Now we set off to obtain some bounds on relevant quantities. We want to estimate

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

The easiest is to note that  $\pi(x) \asymp \frac{1}{n^d}$ . (Its easy to see that if  $\mu = \frac{1}{n^d}$ , then  $\mu P \asymp \mu$  and its easy to prove that if  $\mu P \asymp \mu$ , then  $\pi \asymp \mu$ ). Now it is also easy to see that  $|\Gamma_{xy}| \leq dn$ , this is because the longest a path can be is if it runs through the edges of the box, each having length  $n$ , and in  $d$  dimensions being  $d$  of them. Next we note that  $Q(e) \asymp (dn^d)^{-1}$ , this is because  $P(y, x) \asymp \frac{1}{n^d}$ . Finally we just need to bound, for a fixed edge  $e$ , the number of pairs  $(x, y)$  such that  $\Gamma_{xy}$  contains  $e$ . Now is when I ask you to look at the diagram that says it all. An easy way to bound this quantity, is as follows. Let an edge  $e$  be given. It will always be the case, depending on the orientation of the given edge, that either the starting point or the end point can be placed wherever you want. In the drawn diagram, on the left case, the end point can be placed wherever you want, you have  $n^d$  possibilities. However, the starting point will be restricted to lie in an appropriate side of the line that contains the edge. This gives less than or equal to  $n$  possibilities, and so the total amount of possible pairs is less than or equal to  $n^{d+1}$ . If you look at the diagram on the right, its the same thing, but now the starting point can be placed wherever you want, and then the end point will have to be constrained so that the path  $\Gamma_{xy}$  contains  $e$ . (Essentially the distinction is whether the edge  $e$  points in the first direction or in a different direction. Here's another example of a case in which the starting point can be placed wherever you wish and then the end point has to be corrected:



All in all, for a fixed  $e \in E$ , one has that  $\sum_{x, y: e \in \Gamma_{xy}} 1 \leq n^{d+1}$  so putting this all together:

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

And using Theorem Canonical Paths Method, we finish the claim.



## 11.2 Comparison technique

We have the following result for reversible chains that gives an analogue to the variational characterisation of the second largest eigenvalue:

**Theorem 11.16 (Characterisation of the  $j$ th largest eigenvalue)** Let  $P$  be a reversible chain with respect to the invariant distribution  $\pi$  and let  $\lambda_j$  be its ordered eigenvalues. Then for all  $j \in [n]$  we have that

$$1 - \lambda_j = \max_{\phi_1, \dots, \phi_{j-1}} \min \{ \mathcal{E}(f) : \|f\| = 1, f \perp \phi_1, \dots, \phi_{j-1} \}$$

**Main idea:** To show  $1 - \lambda_j \geq \dots$ , choose an arbitrary collection  $\phi_1, \dots, \phi_{j-1}$  and smartly pick an  $g \perp \phi_1, \dots, \phi_{j-1}$ , and show that  $1 - \lambda_j \geq \mathcal{E}(g)$ , from this it will follow that  $1 - \lambda_j \geq \max \min \dots$  because we have kept the  $\phi$ 's general but the  $f$  is specific. For the other direction, pick a smart choice of  $\phi_1, \dots, \phi_{j-1}$ , and show that for any  $g \perp \phi_1, \dots, \phi_{j-1}$ , then  $1 - \lambda_j \leq \mathcal{E}(g)$ .

*Proof.* Following the main idea of the proof, let  $\phi_1, \dots, \phi_{j-1}$  be arbitrary functions. Let  $W = \text{span}\{\phi_1, \dots, \phi_{j-1}\}$ . Since they may or may not be linearly independent, we have that

$$\dim\{W\} \leq j - 1$$

Which means that  $\dim\{W^\perp\} \geq n - j + 1$  (because the  $\phi$ 's are elements of  $\mathbf{R}^{|\Omega|} = \mathbf{R}^n$ ). Naturally, by orthonormality we have that  $\dim\{\text{span}\{f_1, \dots, f_j\}\} = j$  (the  $f_i$ 's are the eigenfunctions) so it must be the case by comparing dimensions and the fact that for a vector space  $V$ ,  $\dim(A \cap B) \geq \dim(A) + \dim(B) - \dim(V)$ , that  $W^\perp \cap \text{span}\{f_1, \dots, f_j\}$  has dimension at least 1. In particular, pick a  $g$  in this intersection, which without loss of generality can be made to be  $\|g\|_2 = 1$ . Expressing  $g$  as being in the span of the  $(f_j)$ s gives that  $g = \sum_{i \in [j]} a_i f_i$  with  $\sum_{i \in [j]} a_i^2 = 1$ . Plugging in the Dirichlet form gives:

$$\begin{aligned} \mathcal{E}(g) &= \left\langle (I - P) \sum_{i \in [j]} a_i f_i, \sum_{i \in [j]} a_i f_i \right\rangle \\ &= \left\langle \sum_{i \in [j]} a_i (1 - \lambda_i) f_i, \sum_{i \in [j]} a_i f_i \right\rangle \\ &= \sum_{i \in [j]} a_i (1 - \lambda_i) \leq \max_{i \in [j]} (1 - \lambda_i) \sum_{i \in [j]} a_i^2 \\ &= 1 - \lambda_j \end{aligned}$$

This demonstrates that  $1 - \lambda_j \geq \max_{\phi_1, \dots, \phi_{j-1}} \min \{ \mathcal{E}(f) : \|f\|_2 = 1, f \perp \phi_1, \dots, \phi_{j-1} \}$ . Conversely, we can pick  $\phi_i = f_i$  for  $i \in [j - 1]$ , and in this case any function  $g$  perpendicular to the  $\phi_i$ s is in the

span of  $\{f_j, \dots, f_n\}$ , (because they form an orthonormal basis, so all functions are in their span!) so let  $g = \sum_{i=j}^n a_i f_i$  with  $\sum_{i=j}^n a_i^2 = 1$  and we can repeat the argument but when we get to the point that

$$\mathcal{E}(g) = \sum_{i=j}^n a_i(1 - \lambda_j) \geq \min_{j \leq i \leq n} (1 - \lambda_j) = 1 - \lambda_j$$

This show the reverse inequality.



This gives a way to compare the eigenvalues of two chains provided that we are able to compare their Dirichlet forms. The proof is immediate by noting that taking mins and max respects inequalities.

**Corollary 11.17** Let  $P$  and  $P'$  be reversible with respect to  $\pi$  and  $\pi'$  respectively. Let  $\mathcal{E}$  and  $\mathcal{E}'$  be their Dirichlet forms,  $\lambda_j$  and  $\lambda'_j$  denote the eigenvalues. Then if there is a constant  $A$  such that  $\mathcal{E}' \leq A\mathcal{E}$ , then  $1 - \lambda'_j \leq A(1 - \lambda_j)$

Of course this is only useful if we can actually compare Dirichlet forms.

**Theorem 11.18 (Comparison Theorem)** Let  $P$  and  $P'$  be two transition matrices reversible with respect to invariant distributions  $\pi$  and  $\pi'$ . Moreover, letting  $E$  and  $E'$  denote the edge-sets of the chains, assume that for any  $(x, y) \in E'$ , we fix a path  $\Gamma_{xy}$  of edges in  $E$ . Then setting

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

Gives that  $\mathcal{E}' \leq B\mathcal{E}$ . Consequently, if  $B$  is just a constant then we will have that  $\gamma' \lesssim \gamma$  and so  $t_{\text{rel}}' \gtrsim t_{\text{rel}}$

**Main idea:** Do similar things as to what the proof of Canonical Path Method did. I.e: bound  $\{f(x) - f(y)\}^2$  by  $|\Gamma_{xy}| \sum_e (\nabla f(e))^2$

*Proof.*

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

The two steps to keep an eye out for are step (1), which is the same to what happened in the proof of the Canonical Path Method,  $\{f(x) - f(y)\}^2$  was expressed as the square of a telescoping sum going through the edges in  $\Gamma_{xy}$  and then Cauchy Schwarz was applied. Step (2) used again the same idea that was used in Canonical Paths to interchange the sums  $\sum_{x,y} \sum_{e \in \Gamma_{xy}} H(e; x, y) = \sum_{e \in E} \sum_{x,y: e \in \Gamma_{xy}} H(e; x, y)$  where  $H$  is a generic function.  $\heartsuit$

**Example 11.19 (Two graphs with different edge sets)** Suppose two graphs have the same vertex set but different edge sets  $E$  and  $E'$  and we run a simple random walk on both graphs. Then we can easily compute the congestion ratio in the Comparison Method Theorem.

*Explanation.* Since a simple random walk on a graph has a degree-biased invariant distribution and a transition matrix  $P$  with

$$\pi(x) = \frac{\deg(x)}{2|E|} \quad P(x, y) = \frac{\mathbf{1}(x \sim y)}{\deg(x)}$$

It follows immediately that

$$Q(x, y) = \frac{\mathbf{1}(x \sim y)}{2|E|} \quad Q'(x, y) = \frac{\mathbf{1}(x \sim y)}{2|E'|}$$



so then the congestion ratio can be easily computed as

$$\begin{aligned} B &= \max_{e \in E} \frac{1}{Q(e)} \sum_{e \in \Gamma_{xy}} |\Gamma_{xy}| Q'(e) \\ &= \left( \max_{e \in E} \sum_{e \in \Gamma_{xy}} |\Gamma_{xy}| \right) \frac{|E|}{|E'|} \end{aligned}$$

♡

**Example 11.20 (Box with some edges removes)** Consider a 2-dimensional box  $\{1, \dots, n\}^2$ , but now remove some of the horizontal edges at even heights. Run a lazy simple random walk on this new graph. Then  $t_{\text{rel}} = O(dn^2)$

*Proof.* We will simply compare this walk to the usual walk on the box, and use the result we obtained previously that the relaxation time for the lazy simple random walk on the box was  $O(dn^2)$ . The first thing to observe is that if we set  $E'$  and  $E$  to be the edges of the old and new box respectively, if an old edge  $(x, y) \in E'$  is given, we can come up with a path  $\Gamma_{xy}$  of edges in  $E$  (see the diagram) by going up, and then traversing one of the remaining edges at odd height, and then going down (or vice-versa). Thus  $|\Gamma_{xy}| \leq 3$ . Moreover, there will be at most 3  $E$ -paths that contain  $e$  (the  $E$  path that goes down like a  $\cup$ , the one that goes up like a  $\cap$  and the one that is the edge itself). Moreover, since trivially  $|E| \leq |E'|$ , we get that  $B \lesssim 1$  and so the theorem above applies and we get that  $t_{\text{rel new}} = O(dn^2)$ .

♡

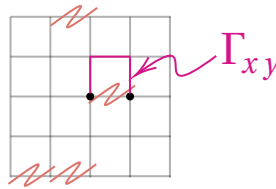


Figure 11.3: Setup for box with edges removed

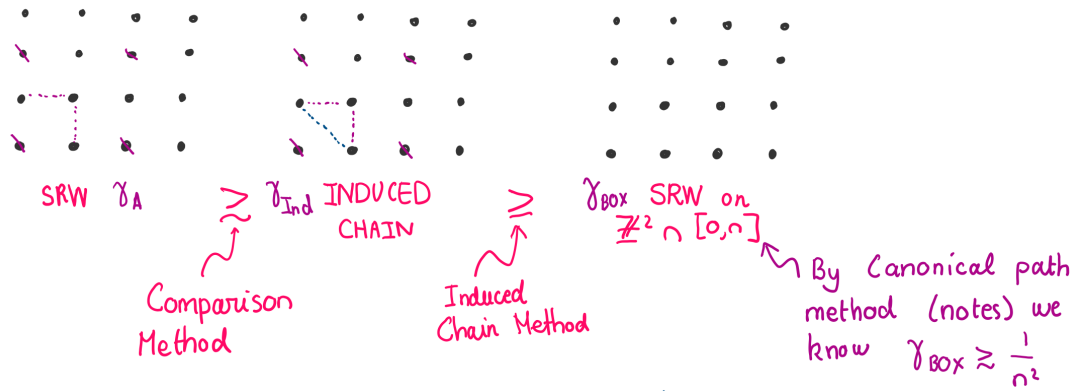
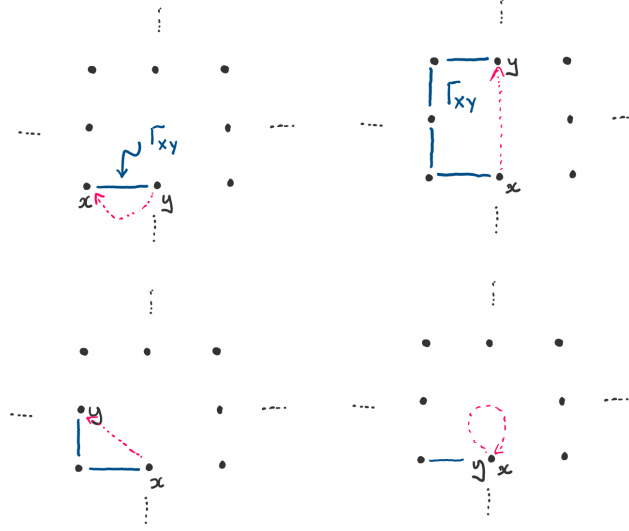


Figure 11.4: Comparison Technique and Induced Chains: The diagram that says it all (I'm sorry I didn't do this diagram in tikzcd)

Here is an example extracted from a homework problem that was too good and illustrative to be left out:

**Example 11.21 (Comparison Technique and Induced Chains)** Consider the subset  $A$  of the box  $[0, n] \cap \mathbb{Z}^2$  obtained by removing the vertices with even coordinates. Run a SRW on  $A$ , show that  $\gamma_A \gtrsim \frac{1}{n^2}$ .

*Proof.* The proof idea is to perform a two-step comparison. We will consider three chains: The SRW on  $A$ , the induced chain on  $A$ , and the original SRW on  $\text{BOX} = [0, 1] \cap \mathbb{Z}^2$ , see figure. Labelling  $\gamma_{\text{BOX}}$ ,  $\gamma_{\text{Ind}}$ , and  $\gamma_A$  the spectral gaps of the SRW on the Box, the induced chain and the SRW on  $A$  respectively, we immediately get from the Theorem on Induced Chains in the Appendix that  $\gamma_{\text{Ind}} \geq \gamma_{\text{BOX}}$ . Moreover, we found that the spectral gap of the LSRW on the box satisfied  $\gamma_{\text{LSRW}} \gtrsim \frac{1}{n^2}$ , and since the spectral gaps of the lazy and non-lazy chains are of the same order, we have that  $\gamma_{\text{BOX}} \gtrsim \frac{1}{n^2}$ . Therefore, all that remains to show is that  $\gamma_A \gtrsim \gamma_{\text{Ind}}$ , and since we are dealing with two chains with the same state space but different transition probabilities, we can apply the Comparison Method. First note that for any given edge  $e = (x, y) \in E_{\text{Ind}}$ , there is a very clear way of producing an  $E_A$ -path  $\Gamma_{xy}$ , simply let  $\Gamma_{xy}$  be the shortest path along edges of  $A$  that connects  $x$  and  $y$ . We distinguish therefore, 4 kinds of paths:



In red is denoted the path long the induced chain, and in blue the corresponding  $E_A$ -path. Recall that the induced chain represents the SRW on the original box, but only watched at the times that it spends at  $A$ , so it could be for example, that the original chain starts at a point  $x$ , leaves  $A$ , and returns to  $x$  immediately, thus giving the self-loop seen in the diagram. Now that we have established the validity to use the Comparison Theorem, we proceed to compute the congestion ratio

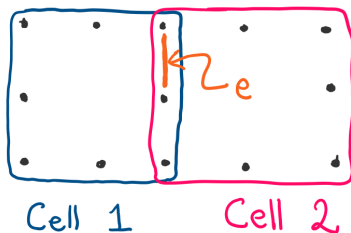
$$B = \max_{e \in E_A} \left( \frac{1}{Q_A(e)} \sum_{x, y: e \in \Gamma_{xy}} |\Gamma_{xy}| Q_{\text{Ind}}(x, y) \right)$$

And from the Comparison Theorem if we show that  $B = O(1)$ , then it will follow that  $\gamma_{\text{Ind}} \lesssim \gamma_A$ . As we need to upper bound  $B$ , we set off to compute the following tasks:

1. Bound  $|\Gamma_{xy}|$ . This is easy, by looking at the diagrams its easy to see that  $|\Gamma_{xy}| \leq 8$ .
2. Upper bound  $Q_{\text{Ind}} := \pi_{\text{Ind}}(x)P_{\text{Ind}}(x, y)$ . We know from the Theorem in the appendix that  $\pi_{\text{Ind}} = \pi(x)/\pi(A)$  for  $x \in A$ . Since  $\pi$  is the invariant distribution for the original box, its easy to see that  $\pi$  is just the degree biased distribution  $\pi(x) = \deg_{\text{BOX}}(x)/2|E_{\text{BOX}}|$ . By looking at the original box its obvious that  $2 \leq \deg_{\text{BOX}}(x) \leq 4$  for all  $x$ , and moreover,  $|E_{\text{BOX}}| \geq 2|V_{\text{BOX}}| \times 1/2$ , where the 2 comes from the minimum degree of a vertex, and the  $1/2$  comes from double counting vertices. From this it is clear that, noting that  $|V_{\text{BOX}}| = n^2$ , that  $\pi_{\text{Ind}}(x) \lesssim \frac{1}{n^2}$ . Since  $P_{\text{Ind}}(x, y) \leq 1$ , we obviously also have that  $Q_{\text{Ind}}(x, y) \lesssim \frac{1}{n^2}$  for all  $(x, y) \in E_{\text{Ind}}$ .
3. Lower bound  $Q_A$ : Since  $A$  is a SRW on a graph,  $\pi_A$  is degree biased,  $\pi_A(x) = \deg_A(x)/2|E_A|$ . Naturally  $\deg_A(x) \geq 1$  and since to obtain  $E_A$  we removed some edges from the box, we have that  $|E_A| < |E_{\text{BOX}}|$ . However, noting that the maximum degree in the box is 4, we have that  $|E_{\text{BOX}}| \leq 4|V_{\text{BOX}}| \lesssim n^2$ . Therefore  $|E_A| \lesssim n^2$ , and we also have that  $\pi_A(x) \gtrsim \frac{1}{n^2}$ . Now, since

$P_A(x, y) \geq 1/4$  for any edge, we have that  $Q_A(x, y) \gtrsim \frac{1}{n^2}$ , so the  $Q_{\text{Ind}}$  and  $Q_A$  will cancel each other out in the upper bound for  $B$ .

4. All that's left for us to determine that  $B = O(1)$  is to show that for a fixed  $e \in E_A$ , the number of  $E_A$ -paths  $\Gamma_{xy}$  that contain  $e$  is bounded above by a constant. To upper bound the number of such paths (we are going to do a brutal bound here), note that an edge  $e \in E_A$  belongs to 2 "Cells" (see diagram), in each of these cells, there are 8 vertices, each of which can be connected via a  $\Gamma$  path to no more than 8 other vertices of the same cell (actually there are less than 8 vertices, but we are just trying to get any bound here), of course, out of all of these  $\Gamma$  paths, some will contain  $e$  and some will not, but regardless, from a single cell, there will be at most  $8^2$  paths containing  $e$ . Since  $e$  is contained in 2 cells, at most  $8^2 \cdot 2 = O(1)$  paths contain  $e$ . (This is a brutal bound I know)



That's, it we have shown that  $B = O(1)$ .



Here's a summary of these last two chapters:

- As explained in the previous summary, by defining  $\gamma(P)$  through the variational characterisation, we found ways to upper bound the mixing time.
- We then wished to study chains obtained by slightly perturbing some "nice" chains. The first step was to show that if one could compare Dirichlet forms, say  $\mathcal{E}' \leq A\mathcal{E}$ , then  $\gamma' \lesssim \gamma$ . In light of the above bulletpoint, if the chains are reversible, then  $\mathcal{E}'$  and  $\mathcal{E}$  are the forms of the old and new chain respectively, then  $t_{\text{mix}}(\epsilon) \lesssim \frac{1}{\gamma'}$ , and so we are also able to obtain bounds on the new mixing time, if we know  $\gamma'$ , i.e: the spectral gap of the old chain.
- We would therefore like to know how to do two things: lower bound  $\gamma'$ , (i.e: upper bound  $1/\gamma'$ ), and be able to perform the bound  $\mathcal{E}' \leq A\mathcal{E}$  for some constant  $A$ , for this we found the following:

1. The canonical path method: which by defining the congestion ratio

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

of a chain, Theorem 11.13 gives that  $\gamma \geq 1/B$ .

2. The comparison method: which essentially says that if you have two chains on the same graph, but to obtain the new chain you remove some edges, then by computing an analogous congestion ratio:

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

one can obtain the desired bound  $\mathcal{E}' \leq B\mathcal{E}$ . Where a dash represents the original chain.

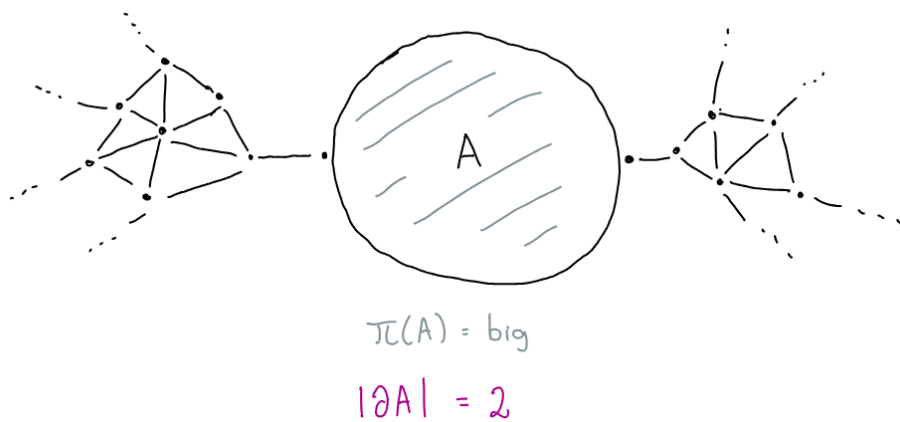


Figure 11.5: A set  $A$  with a massive  $\pi$ -measure but a very small conductance.

### 11.3 Bottleneck ratio

So far we have not seen a great deal of methods of lower bounding mixing times. We have found the (as I like to call it) "dirty approach" where you show that the possible locations of the chain after  $t$  steps do not form a significant proportion of the state space in terms of  $\pi$ -measure, i.e: formally, we found sets  $A$  such that  $\mathbf{P}[X_t \in A]$  was high but  $\pi(A)$  was small. Another way, seen in the example sheet 1, was the diameter bound, where for a transition matrix  $P$  on  $\Omega$ , we constructed a graph with vertex set  $\Omega$  and an edge  $(x, y)$  if one of  $P(x, y)$  or  $P(y, x)$  was strictly greater than zero, and if the diameter of this graph was  $L$ , we concluded that for any  $\epsilon < 1/2$ ,  $t_{\text{mix}}(\epsilon) \geq \frac{L}{2}$ . We now see yet another way to lower bound mixing time (and upper bound it too), which also in some sense captures some of the geometry of the chain.

**Definition 11.22** (Bottleneck Ratio) The bottleneck ratio (known as the Cheeger constant) is defined to be

$$\Phi_* = \min_{A \subseteq \Omega: \pi(A) \leq 1/2} \frac{Q(A, A^c)}{\pi(A)} \equiv \min_{A \subseteq \Omega: \pi(A) \leq 1/2} \frac{\sum_{x \in A, y \in A^c} \pi(x)P(x, y)}{\pi(A)}$$

We also define the conductance of a set  $A \subseteq \Omega$  to be  $\Phi(A) = \frac{Q(A, A^c)}{\pi(A)}$

**Example 11.23 (Bottleneck Ratio of a SRW on a Graph)** If we consider the simple random walk on a graph  $(V, E)$ , we have that

$$Q(x, y) = \underbrace{\frac{\deg(x)}{2|E|}}_{\pi(x)} \times \underbrace{\frac{\mathbf{1}(x \sim y)}{\deg(x)}}_{P(x, y)} = \frac{\mathbf{1}(x \sim y)}{2|E|}$$

And so, for a set  $A \subseteq V$ , we have that

$$Q(A, A^c) = \frac{1}{2|E|} \sum_{x \in A, y \in A^c} \mathbf{1}(x \sim y) = \frac{|\partial A|}{2|E|}$$

Therefore

$$\Phi(A) = \frac{|\partial A|}{\sum_{x \in A} \deg(x)} := \frac{|\partial A|}{\text{Vol}(A)}$$

(And we would have  $2\text{Vol}(A)$  if the walk were lazy) This gives a nice interpretation of the bottleneck ratio, as the ratio between surface area and volume of a subset of the graph. Imagine heat trapped in a body, if the surface is very small compared to the volume of the body, heat will have a hard time escaping.

**Main idea:**  $Q(A, B)$  measures the probability of going from  $A$  to  $B$  in one step, when starting from the invariant distribution. Therefore  $\Phi(A)$  is precisely the probability that a particle with starting distribution  $\pi_A$  exits the chain in one step:  $\mathbf{P}_{\pi_A}[X_1 \in A^c] = \frac{\mathbf{P}_{\pi}[X_0 \in A, X_1 \in A^c]}{\pi(A)}$

More formally, the relationship between  $t_{\text{mix}}$  and the bottleneck ratio is the following:

**Theorem 11.24 (Bottleneck and mixing time)**

$$t_{\text{mix}} \geq \frac{1}{4\Phi_*}$$

**Main idea:** The key is that by convexity of TV we know that  $d(t) \geq \|\mu P^t - \pi\|_{\text{TV}}$  for any distribution  $\mu$ . So we can use  $\mu = \pi_A$  where  $A$  is the set of worst conductance. Then intuitively a dirty bound will do because  $A$  is a big set that is hard to escape in one go. To extend this to  $\mathbf{P}_{\pi_A}[X_t \in A^c]$  use union bound.

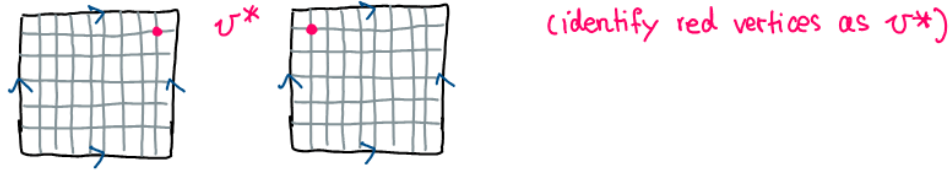


Figure 11.6: Two tori with a single vertex  $v^*$  identified, thus gluing them.

*Proof.* By convexity of TV we know that

$$\begin{aligned} \|\mu P^t(\cdot) - \pi(\cdot)\|_{\text{TV}} &= \left\| \sum_{y \in \Omega} \mu(y) P^t(y, \cdot) - \pi(\cdot) \right\|_{\text{TV}} \\ &\leq \sum_{y \in \Omega} \mu(y) \|P^t(y, \cdot) - \pi(\cdot)\|_{\text{TV}} \\ &\leq \max_{x \in \Omega} \|P^t(x, \cdot) - \pi(x)\|_{\text{TV}} \sum_y \mu(y) = d(t) \end{aligned}$$

Now we apply  $\mu = \pi_A$  and see that using the set  $A$  that attains the infimum in  $\Phi_*$

$$d(t) \geq \pi(A^c) - \mathbf{P}_{\pi_A}(X_t \in A^c) \geq \frac{1}{2} - \mathbf{P}_{\pi_A}(X_t \in A^c)$$

We now apply a union bound on this probability: if  $X_t \in A^c$ , then for some  $0 \leq s < t$ , it must be that  $X_s \in A$  and  $X_{s+1} \notin A$ . Therefore

$$\begin{aligned} \mathbf{P}_{\pi_A}(X_t \in A^c) &= \mathbf{P}_{\pi}(X_0 \in A, X_t \in A^c) \\ &= \mathbf{P}_{\pi} \left( \bigcup_{s=0}^{t-1} \{X_s \in A, X_{s+1} \in A^c\} \right) \\ &\leq \sum_{s=0}^{t-1} \mathbf{P}_{\pi}(X_s \in A, X_{s+1} \in A^c) \end{aligned}$$

Since the chain is started with  $\pi$ ,  $(X_{n+s})$  has the same distribution as  $(X_n)$  (stationarity of  $\pi$ ), we have that this sum upstairs is equal to  $t \mathbf{P}_{\pi}(X_0 \in A, X_1 \in A^c) = t Q(A, A^c)$  Therefore

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

Thus taking  $t = (4\Phi_*)^{-1}$  gives that  $d(t) \geq 1/4$  so  $t_{\text{mix}} \geq (4\Phi_*)^{-1}$ . ♡

**Example 11.25 (Glued torus)** Consider two tori glued together at a single vertex  $v^*$ . Let us lower bound its mixing time:



*Solution.* Let  $V_1$  and  $V_2$  be the left and right tori respectively, and consider the subset of this glued graph  $A$  defined by  $A = V_1 \setminus \{v^*\}$ . Since there are  $2d$  edges connecting  $v^*$  to  $V_1$ , we see that  $|\partial A| = 2d$ , and since  $A$  has  $n^d - 1$  vertices, each with degree  $2d$ ,  $\pi(A) = 2d(n^d - 1)$ , therefore

$$\Phi^* \leq \Phi(A) = \frac{2d}{2d(n^d - 1)},$$

and so  $t_{\text{mix}} \gtrsim \frac{1}{n^d - 1}$



**Example 11.26 (Binary tree)** In previous examples we talked about the mixing time of the lazy simple random walk on the binary tree of  $n$  vertices. We now give an alternative proof for the lower bound that  $t_{\text{mix}} \gtrsim n$  using Isoperimetry:

*Proof.* Let  $S$  be the right-hand side of the tree. Since there are  $n - 1$  edges, and  $\deg(x) \leq 3$

$$\pi(x) = \frac{\deg(x)}{2(n-1)} \leq \frac{3}{2(n-1)}.$$

Moreover,  $P(x, y) = \frac{1}{\deg(x)} \mathbf{1}\{x \sim y\} \leq \mathbf{1}\{x \sim y\}$ , from which we gather that in fact

$$Q(S, S^c) = \sum_{x \in S, y \in S^c} \pi(x) P(x, y) \leq \frac{3}{2(n-1)} |\partial S| = \frac{3}{2(n-1)}.$$

Finally, we note that

$$\pi(S) = \sum_{x \in S} \pi(x) \leq \frac{3}{2(n-1)} |S| = \frac{3(n+1)}{4(n-1)},$$

from which it now follows clearly that  $\Phi(S) \lesssim 1/n$ , and so  $t_{\text{mix}} \gtrsim n$ .



We now get an inequality that bounds  $\gamma$  in terms of  $\Phi_*$ . This following inequality is very powerful and qualitatively tells us that bottlenecks are the only major obstruction to fast mixings. For this we will need to invoke the power of the Dirichlet form.

**Theorem 11.27 (Cheeger's inequality)** Let  $P$  be reversible with respect to  $\pi$ . Let  $\gamma$  be the spectral gap. Then

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

Observe that if the chain exhibits a big bottleneck ( $\Phi_*$  is small), then  $\gamma$  is also small (from the upper bound), which means that the mixing will be slow. Conversely, if there are no bottlenecks ( $\Phi_*$  is big), then  $\gamma$  will also be big (lower bound), which means the mixing will be quick.

**Main idea:** For the upper bound, use variational characterisation of  $\gamma$  and consider the function  $f(x) = \mathbf{1}_A(x)$  for some set with  $\pi(A) \leq 1/2$ .

*Upper bound.* • Upper bound: consider the function  $f(x) = \mathbf{1}_A(x)$ . Plugging in the variational characterisation of the spectral gap, which is allowed because  $f(x)$  is non-constant, gives that

$$\gamma \leq \frac{\mathcal{E}(f)}{\text{Var}_\pi[f]} = \frac{\sum_{x,y \in \Omega} \pi(x)P(x,y)[f(x)-f(y)]^2}{\sum_{x,y \in \Omega} \pi(x)\pi(y)[f(x)-f(y)]^2}$$

Let us manually compute, say the denominator, and then it will be clear how the numerator also follows:

$$\sum_{x,y} \pi(x)\pi(y)[\mathbf{1}_A(x) - \mathbf{1}_A(y)]^2 = \sum_{x \in A, y \in A^c} \% + \sum_{x \in A^c, y \in A} \% + \sum_{x \in A, y \in A} \% + \sum_{x \in A^c, y \in A^c} \% = 2 \sum_{x \in A, y \in A^c} \%$$

In here we have used the fact that in the last two sums, the indicators will both be the same and so will cancel always each other out, and then to make the first two sums equal, we just observed that the summands are symmetric in  $x \leftrightarrow y$ . This last sum will be equal to

$$\sum_{x \in A^c, y \in A} \pi(x)\pi(y) = \sum_{x \in A^c} \pi(x) \sum_{y \in A} \pi(y) = \pi(A^c)\pi(A)$$

Similarly the numerator will be  $2Q(A, A^c)$ . I.e:

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

This worked for any  $A$ , so let us choose any  $A$  with  $\pi(A) \leq 1/2$ , so that  $\pi(A^c) \geq 1/2$ , and as

such

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

Since this holds for all  $A$  with  $\pi(A) \leq 1/2$ , in particular we have that  $\gamma \leq 2\Phi_*$ .



**Main idea:** For the lower bound there are two steps

1. Find a non-negative  $f$  with  $\pi(\{f > 0\}) \leq 1/2$  and  $\gamma \geq \frac{\mathcal{E}(f)}{\|f\|_2^2}$ . For this consider  $f = f_2 \vee 0$  where  $f_2$  is the second eigenvector.
2. Show that for any function  $\psi \geq 0$  we have that  $\frac{\mathcal{E}(\psi)}{\|\psi\|_2^2} \geq \frac{h(\psi)^2}{2}$  for some suitable  $h$ .

*Proof of lower bound.* We prove two steps

- Step 1.

We show there exists a non-negative function  $f$  such that  $\pi(\{f > 0\}) \leq 1/2$ , and  $\gamma \geq \frac{\mathcal{E}(f)}{\|f\|_2^2}$ . The function in question will be the eigenfunction  $f_2$  corresponding to eigenvalue  $\lambda_2$ . Suppose without loss of generality that  $\pi(\{f_2 > 0\}) \leq 1/2$ , otherwise consider the function  $-f_2$ . Now let  $f = f_2 \vee 0$ . Then we claim that  $[(I - P)f](x) \leq \gamma f(x)$ . Indeed: If  $f(x) = 0$  then there's nothing to prove. If  $f(x) > 0$ , then  $f(x) = f_2(x)$  and so

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

This inequality holds because generally  $f \geq f_2$ , and

$$(Pf)(x) = \sum_y P(x, y)f(y) \geq \sum_y P(x, y)f_2(y) = (Pf_2)(x).$$

Now we can plug in the Dirichlet form:

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

- Step 2.

Fix a non-negative function  $\psi \geq 0$ , and for every  $t > 0$ , define  $S_t = \{x : \psi(x) > t\}$ . Let

$$h(\psi) = \inf_{\emptyset \neq A \subseteq \{x : \psi(x) > 0\}} \frac{Q(A, A^c)}{\pi(A)}$$

Then if for some  $t$  you have that  $S_t \neq \emptyset$ , then it also follows that (using reversibility) that

$$h(\psi) \leq \frac{Q(S_t, S_t^c)}{\pi(S_t)} = \frac{Q(S_t^c, S_t)}{\pi(S_t)} = \frac{1}{\pi(S_t)} \sum_{\psi(x) \leq t, \psi(y) > t} Q(x, y)$$

Then we have that

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

Where in step (1) we used Cauchy-Schwarz and in step (2) we used the fact that  $(a+b)^2 \leq 2(a^2+b^2)$ . The proof of the lower bound now follows immediately. Indeed: We have just shown that for a function  $\psi \geq 0$ ,

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

and in section one we showed that there is a function  $f \geq 0$  with  $\pi\{f > 0\} < 1/2$  such that  $\gamma \geq \frac{\mathcal{E}(f)}{\|f\|_2^2}$ , thus in combination:

$$\gamma \geq \frac{1}{2} \inf_{A \subseteq \{f > 0\}} \frac{Q(A, A^c)}{\pi(A)} \geq \frac{1}{2} \inf_{A: \pi(A) \leq 1/2} \frac{Q(A, A^c)}{\pi(A)} = \frac{1}{2} \Phi^*$$



**Example 11.28** (Agreement with previous examples) Recall that we proved that for the LSRW on  $\mathbb{Z}_n$  we had that  $t_{\text{rel}} \asymp n^2$ . Recalling that for the Lazy Simple Random Walk, we have that

$$\Phi(A) = \frac{|\partial A|}{2\text{Vol}(A)}$$

it is easy to check that the smallest  $\Phi(A)$  can be over all  $A$  with  $\pi(A) \leq 1/2$ , corresponds to the line segment of length  $n/2$ , which has a boundary of two vertices, and has a volume of  $n/2$  so we get that

$$\Phi_* \asymp \frac{1}{n}$$

Thus we see that the lower bound of Cheeger's inequality is achieved up to constants here. To see an example where the upper bound is attained, consider the LSRW on the hypercube  $\{0,1\}^n$  and let the set  $A = \{(x_1, \dots, x_n) : x_1 = 0\}$ . Clearly this is just half of the hypercube and so  $\pi(A) = 1/2$ . Moreover, the probability of exiting it, is the probability of  $x_1$  changing to a 1, which is of  $1/2n$ . Therefore  $\Phi_* \leq 1/2n$ . We showed that  $\gamma = 1/n$  and so it must be that  $\Phi_* = 1/2n$  and we conclude that the upper bound is sharp.

## 11.4 Expander graphs

We have shown that when a graph has a narrow bottleneck, i.e:  $\Phi_*(G) \ll 1$ , the walk mixes slowly. Now we study the following question: how efficiently can a family of graphs avoid bottlenecks?

**Definition 11.29** (Bottleneck ratio of a SRW on a graph) Let  $G$  be a graph. We denote by  $\Phi_*(G)$  the bottleneck ratio of the simple random walk on  $G$ .

**Definition 11.30** (Expander graph) A sequence  $G_n = (V_n, E_n)$  of graphs, is called  $(d, \alpha)$ -regular expander family if:

- $\lim_{n \rightarrow \infty} |V_n| = \infty$ .
- $G_n$  is  $d$ -regular, i.e: if for each  $n \in \mathbb{N}$ , and each  $v \in V_n$ , we have that  $\deg(v) = d$
- The bottleneck ratios satisfy  $\Phi_*(G_n) \geq \alpha$  for all  $n$ .

We call the family  $(\Delta, \alpha)$  expander if instead of requiring regularity on the degrees we just require the degrees to be bounded by  $\Delta$ , and we say the family is  $\alpha$  expander if there exists some  $\Delta$  so that the graph family is  $(\Delta, \alpha)$  expander.

**Proposition 11.31** (Mixing time of expanders) Let  $G_n$  be an  $\alpha$ -expander family. Then the mixing time of the lazy simple random walk on  $G_n$  satisfies  $t_{\text{mix}} = O(\log |V(G_n)|)$

**Main idea:** From definition of Expander Graph, we have that  $\gamma_{SRW} \geq \alpha^2/2$ , now use the fact that  $\gamma_{LSRW} = \frac{1}{2}\gamma_{SRW}$  and the inequality that bounds  $t_{\text{mix}}$  above by  $t_{\text{rel}}$  and then estimate  $\pi_{\min}$ .

*Proof.* From the lower bound in Cheeger's inequality and the assumption of  $\alpha$ -expander, we have that

$$\gamma_{SRW} \geq \frac{\Phi_*^2(G_n)}{2} \geq \frac{\alpha^2}{2}$$

Now we note that (labelling  $P'$  for the lazy matrix)  $\gamma_{LSRW} = \inf_f$  with some regularity  $\mathcal{E}_{P'}(f)$  and  $\mathcal{E}_{P'}(f) = \langle (I - \frac{P+I}{2})f, f \rangle = \frac{1}{2} \langle (I + P)f, f \rangle = \frac{1}{2} \mathcal{E}_P(f)$ , we have that  $\gamma_{SRW} = 2\gamma_{LSRW}$ . Moreover, note that for a lazy chain we have that  $\gamma^* = \gamma$ , so by the bound on the mixing time with relaxation time we see that

$$t_{\text{mix}}' \leq t_{\text{rel}}' \log \frac{4}{\pi_{\min}} \leq \frac{4}{\alpha^2} \log \frac{4}{\pi_{\min}}$$

We know that the invariant distribution  $\pi_n$  satisfies

$$\pi_n(x) = \frac{\deg(x)}{2|E_n|}$$

but we can estimate  $|E_n| \leq \Delta|V_n|/2$ . Indeed: the total number of edges is at most the number of vertices times the degree of each vertex, but since doing this we double count each edge, we can divide by 2. This gives

$$\pi_{\min} \geq \frac{1}{\Delta|V_n|}$$

so plugging back in we get that

$$t_{\text{mix}} \leq \frac{4}{\alpha^2} \log(4\Delta|V_n|) = O(\log|V_n|)$$



In fact we can show that expanders have the fastest mixing time among all regular graphs

**Proposition 11.32** Let  $G$  be a bounded degree graph, then  $t_{\text{mix}} \gtrsim \log|V_n|$

*Proof.* We recall the result from ES1 Q7, which we have cited already at the start of a previous section: let  $G = (V, E)$  be a finite connected graph with diameter  $D$ . Let  $X$  be a (lazy) simple random walk on  $G$ . Then for all  $\epsilon < 1/2$ , we have that

$$t_{\text{mix}}(\epsilon) \geq D/2$$

Now we just aim to show that if  $G$  is a graph with bounded degrees, its diameter  $D$  satisfies  $D \gtrsim \log|V(G_n)|$ . This will finish the claim. We will actually prove the following:

**Claim:** Let  $G$  be a graph whose degrees are bounded above by  $\Delta$ . Fix  $x \in V$ , then

$$\#\{y \in V : \text{dist}(x, y) = r\} \leq \Delta^r.$$

**Proof of Claim:** the proof is by induction. At  $r = 0$ , it is clear, because the only vertex  $y \in V$  with  $\text{dist}(x, y) = 0$  is  $x$  itself. Now assume the claim holds for some  $r = k$ . To show the claim holds for  $r = k + 1$ , we note that since there are at most  $\Delta^k$  vertices at a distance  $k$  and each of these vertices can produce at most  $\Delta$  vertices going out to a distance  $k + 1$ , the total number of vertices at distance  $k + 1$  must be at most  $\Delta^{k+1}$ .

Now we can finish the proof of the Proposition, because using our claim, the number of vertices at a distance at most  $r$  from  $x$  will be at most  $1 + \Delta + \Delta^2 + \dots + \Delta^r \leq \Delta^{r+1}$ . However, by definition

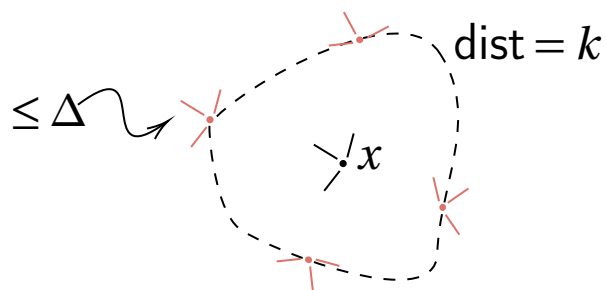


Figure 11.7: The diagram that says it all of the Proof of Claim of Theorem Expanders Have Fastest Mixing Time

of diameter, the the number of vertices at a distance  $D$  from  $x$  is exactly  $|V(G_n)|$ , and so we have that

$$|V(G_n)| \leq \Delta^{D+1}$$

and rearranging we have that  $D \gtrsim \log |V(G_n)|$





**Theorem 11.33** There exists a random graph model  $G_n$  with  $|V_n| \geq n$ ,  $G_n$  is 3-regular and for some  $\alpha > 0$ ,

$$\mathbf{P}(\Phi_*(G_n) \geq \alpha) \rightarrow 1$$

**Main idea:** The proof is by construction. A bipartite graph family with  $V_n = A_n \cup B_n$  is specified, and then edges are set so as to make the graph 3 regular (this might yield a multigraph). We then show that with this construction, with a very high probability, every small enough subset of  $A$  will have a large enough number of neighbours in  $B$ , this will show that the bottleneck ration cannot be too small.

*Proof.* Let our vertex set  $V_n = A_n \cup B_n$  where  $A_n = \{a_1, \dots, a_n\}$ , and  $B_n = \{b_1, \dots, b_n\}$ . We construct our edge set

$$E_n = \{(a_i, b_i), (a_i, b_{\sigma_1(i)}), (a_i, b_{\sigma_2(i)}) : 1 \leq i \leq n\}$$

Where  $\sigma_1, \sigma_2$  are drawn uniformly at random from the symmetric group  $\mathcal{S}_n$ . It is clear that this graph will be 3-regular and that  $|V_n| \rightarrow \infty$ , so we will now show that

$$\mathbf{P}(\Phi_*(G_n) > 0.01) \rightarrow 1$$

We approach this by first showing that *for  $n$  large enough, with high probability, every subset of  $A$  that has size at most  $k = n/2$ , must have more than  $(1 + \delta)k$  neighbours.* Indeed, let  $S \subseteq A$  be any subset of  $A$  with  $|S| = k \leq n/2$ . Recall that

$$N(S) = \{v \in V \setminus S : v \sim w \text{ for some } w \in S\}$$

By the pairing of vertices we have done, since  $a_i \sim b_i$  for all  $i$ , we automatically have that  $N(S) \geq k$ , so the probability

$$\mathbf{P}(|N(S)| \leq (1 + \delta)k)$$

is less than or equal to probability that if we pick any set of surplus  $\delta k$  vertices in  $B$ , the edges specified by  $\sigma_1$  and  $\sigma_2$  all fall within that set or within the  $k$  vertices that are already linked. There are  $\binom{n}{\delta k}$  ways to specify a set of  $\delta k$  surplus edges from  $B$ , and now let's calculate the probability that the edges specified by  $\sigma_1$  all fall within either this new set of size  $\delta k$ , or back in the vertices that are already linked. This is easy to compute, simply

$$\frac{\binom{k + \delta k}{k}}{\binom{n}{k}}$$

(Remember that  $\sigma_1$  is specifying  $k$  edges of  $S$ ). Therefore, since  $\sigma_2$  is an iid copy, we conclude

that

$$\mathbf{P}(|N(S)| \leq (1 + \delta)k) \leq \binom{n}{\delta k} \left( \frac{\binom{k+\delta k}{k}}{\binom{n}{k}} \right)^2$$

Now by union bound:

$$\mathbf{P}(\text{for some } S, |S| \leq n/2 \text{ and } |N(S)| \leq (1 + \delta)k) \leq \sum_{k=1}^{n/2} \binom{n}{k} \binom{n}{\delta k} \left( \frac{\binom{k+\delta k}{k}}{\binom{n}{k}} \right)^2$$

By some magic (ES3), it can be shown that this probability tends to zero as  $n \rightarrow \infty$  for  $\delta$  small enough, so

$$\mathbf{P}(\text{any } S \subseteq A, |S| \leq n/2, \text{ has that } |N(S)| > (1 + \delta)k) \rightarrow 1$$

We now finish by showing that on the event

$$\{\text{any } S \subseteq A, |S| \leq n/2, \text{ has that } |N(S)| > (1 + \delta)k\}$$

we have that  $\Phi_*(G_n) > \delta/2$ . Recall that (since  $|V_n| = 2n$ ):

$$\Phi_*(G_n) = \min_{S \subseteq V_n, |S| \leq n} \frac{|N(S)|}{|S|}$$

so we are going to start by picking any  $S \subseteq V_n$  with size  $|S| \leq n$ . Let  $A' = A \cap S$  and  $B' = B \cap S$ , without loss of generality we can assume  $|A'| \geq |B'|$ . We distinguish between two cases

- If  $|A'| \leq n/2$ : then we can apply the above argument, and see that  $|N(A')| > (1 + \delta)|A'|$ . Therefore  $A'$  has more than

$$(1 + \delta)|A'| - |B'|$$

neighbours in  $B \setminus B'$ , and since  $|A'| \geq |S|/2$  (because  $|A'| \geq |B'|$  wlog), then

$$(1 + \delta)|A'| - |B'| \geq \delta|S|/2$$

in other words, there are more than  $\delta|S|/2$  edges going from  $S$  to  $S^c$ , so  $\Phi(S) \geq \delta/2$ .

- If  $|A'| > n/2$ , then choose a subset  $A'' \subseteq A'$  of size  $\lfloor n/2 \rfloor$ , so we can apply the hypothesis again, and deduce that  $A''$  has more than  $\delta|S|/2$  neighbours in  $B \setminus B'$ , which in turn obviously implies that  $A$  has more than that amount, and as such the amount of edges going from  $S$  to  $S^c$  is once again more than  $\delta|S|/2$ , so we reach the same conclusion.



**Remark 11.34** We have shown a family of expander graphs exists, but the graphs we have constructed are multigraphs, it can be shown that a family of regular graphs can also be obtained by performing some minor modifications to our construction. But I shall not include this proof.



# Chapter 12

## Spectral profile and isoperimetric profile

### 12.1 Spectral Profile

In this section we will define the spectral profile, a generalisation of the Poincaré constant. The idea is that the definition of the Poincaré constant takes into account all (non-constant) functions on the state space to  $\mathbf{R}$ , but perhaps a more detail inspection can be obtained by restricting to functions which are supported on some set of  $\pi$  measure  $\pi_{\min} < r \leq 1$ . We will see how we will obtain analogous inequalities for the mixing times, which will be tighter. In particular, we have the corresponding inequalities

Poincaré Constant	Theme	Spectral Profile
$(\text{Var}_{\pi}[P^*f] \leq (1 - \gamma(P P^*)) \text{Var}_{\pi}[f])$	Variational Contraction	$(\text{Var}_{\pi}[P^*f] \leq (1 - \frac{1}{2} \Lambda_{P P^*} \left( \frac{4(\mathbb{E}_{\pi}[f])^2}{\text{Var}_{\pi}[f]} \right)) \text{Var}_{\pi}[f])$
$((d_2(t))^2 \leq (1 - \gamma(P P^*))^2 \frac{1}{\pi_{\min}})$	$(\mathcal{L}^2) Distance$	$(d_2(t) \leq \frac{4}{\sqrt{P P^*}(t/2)})$
$(t_{\text{mix}}(\epsilon) \leq \frac{2}{\gamma(P P^*)} \log \left( \frac{1}{2\epsilon \sqrt{\pi_{\min}}} \right))$	Mixing Time	$(t_{\text{mix}}^{(\infty)}(\epsilon) \leq 2 \left[ \int_{4\pi_{\min}}^{4/\epsilon} \frac{du}{au\Lambda(u)} \right])$

Table 12.1: Comparison of Poincaré Constant and Spectral Profile Equations by Theme

**Definition 12.1** Let  $P$  be a transition matrix, and let  $r \geq \pi_{\min}$ , the spectral profile  $\Lambda_P(r)$  is given by

$$\Lambda_P(r) \equiv \Lambda(r) = \inf_{\pi_{\min} \leq \pi(A) \leq r} \lambda_P(A)$$

Where  $\lambda_P(A)$  is defined by

$$\lambda_P(A) = \inf_{f \in C_0^+(A)} \frac{\mathcal{E}(f)}{\text{Var}_{\pi}(f)}$$

and  $C_0^+(A) = \{f \geq 0 : \text{supp}(f) \subseteq A, f \text{ non-constant}\}$

**Remark 12.2** We have the following two observations

- $\Lambda_P(r) \geq \gamma(P)$ : this is because in the Poincaré constant, the infimum is taken over all non-constant functions, whereas when computing  $\Lambda_P(r)$  we take the infimum of an analogous quantity to the Poincaré constant that takes into account a smaller subset of functions.
- $\Lambda_P(r)$  is non-increasing in  $r$ : this is obvious because if  $r$  is larger, then when taking the infimum to compute  $\Lambda_P(r)$  there are more options to include and so the infimum cannot be larger.

**Lemma 12.3** (Variational contraction: Spectral profile, ♪) For a non-negative function  $f : \Omega \rightarrow \mathbf{R}_+$ , we have that

$$\frac{\mathcal{E}(f)}{\text{Var}_\pi(f)} \geq \frac{1}{2} \Lambda_P \left( \frac{4\mathbf{E}_\pi[f]^2}{\text{Var}_\pi[f]} \right)$$

which in turn implies a variational contraction result:

$$\text{Var}_\pi[f] - \text{Var}_\pi[P^*f] \geq \frac{1}{2} \text{Var}_\pi[f] \Lambda_{PP^*} \left( \frac{4\mathbf{E}_\pi[f]^2}{\text{Var}_\pi[f]} \right)$$

**Main idea:** Show that  $\mathcal{E}(f) \geq \text{Var}_\pi[(f-c)^+] \Lambda(\pi\{f > c\})$  and then show that  $\text{Var}_\pi[(f-c)^+] \geq \text{Var}_\pi[f] - 2c\mathbf{E}_\pi[f]$ . Then use Markov's inequality with  $c = \frac{\text{Var}_\pi[f]}{4\mathbf{E}_\pi[f]}$

*Proof.* Let us show the first inequality. First we have that for any  $0 \leq c \leq \max f$ ,

$$\begin{aligned} \mathcal{E}(f) &\stackrel{(1)}{=} \mathcal{E}(f-c) \\ &\stackrel{(2)}{\geq} \mathcal{E}((f-c)^+) \\ &\stackrel{(3)}{\geq} \text{Var}_\pi[(f-c)^+] \inf_{g \in c_0^+ \{x: f(x) > c\}} \frac{\mathcal{E}(g)}{\text{Var}_\pi[g]} \\ &\stackrel{(4)}{\geq} \text{Var}_\pi[(f-c)^+] \Lambda(\pi\{f > c\}). \end{aligned}$$

Where (1) comes from the fact that  $\mathcal{E}(f)$  is translation invariant (since  $\mathcal{E}(f) = \sum_{x,y} \{f(x) - f(y)\}^2 \pi(x)P(x,y)$ ). Step (2) comes from the fact that  $(a-b)^2 \geq (a^+ - b^+)^2$ , (and so plugging in the expression of  $\mathcal{E}(f)$  in the previous gray comment gives step (2)). Step (3) comes from the fact that  $(f-c)^+$  is a non-negative function supported on the set where  $f(x) > c$ . Step (4) comes from the fact that  $\lambda(A) \geq \Lambda(\pi(A))$  (as  $\Lambda(r)$  takes the infimum over all sets of mass at most  $r$ ).

Now if  $a, b \geq 0$ , then naturally we have that  $((a-b)^+)^2 \geq a^2 - 2ba$  and  $(a-b)^+ \leq a$ . (Just check

cases  $a \geq b$  or  $b > a$ ) Now we compute

$$\begin{aligned}\text{Var}_\pi[(f-c)^+] &:= \mathbf{E}_\pi[((f-c)^+)^2] - (\mathbf{E}_\pi[(f-c)^+])^2 \\ &\geq \mathbf{E}_\pi[f^2] - 2c\mathbf{E}_\pi[f] - (\mathbf{E}_\pi[f])^2 \\ &\geq \text{Var}_\pi[f] - 2c\mathbf{E}_\pi[f]\end{aligned}$$

And so by choosing  $c = \frac{\text{Var}_\pi[f]}{4\mathbf{E}_\pi[f]}$  (Which can be easily checked - by expanding it - that it is no larger than  $f_{\max}$  . Therefore, putting this all together:

$$\begin{aligned}\mathcal{E}(f, f) &\stackrel{(1)}{\geq} (\text{Var}_\pi[f] - 2c\mathbf{E}_\pi[f])\Lambda(\pi\{f > c\}) \\ &\stackrel{(2)}{\geq} (\text{Var}_\pi[f] - 2c\mathbf{E}_\pi[f])\Lambda\left(\frac{\mathbf{E}_\pi[f]}{c}\right) \\ &\stackrel{(3)}{=} \frac{1}{2}\text{Var}_\pi[f]\Lambda\left(\frac{4(\mathbf{E}_\pi[f])^2}{\text{Var}_\pi[f]}\right)\end{aligned}$$

Where step (1) comes from mashing all the inequalities together, step (2) comes from the fact that  $\pi\{f > c\} \leq \frac{\mathbf{E}_\pi[f]}{c}$  by Markov's inequality, and the fact that  $\Lambda(r)$  is non-increasing (taking infimum over more things). Step (3) comes from the definition of  $c$ . Now the second statement follows immediately:

$$\begin{aligned}\text{Var}_\pi[f] - \text{Var}_\pi[P^*f] &\stackrel{(1)}{=} \langle f, f \rangle - \langle P^*f, P^*f \rangle \\ &\stackrel{(2)}{=} \langle (I - PP^*)f, f \rangle \\ &= \mathcal{E}_{PP^*}(f)\end{aligned}$$

Where step (1) comes from the fact that  $\mathbf{E}_\pi[f] = \mathbf{E}_\pi[P^*f]$ , and step (2) comes from the fact that  $P^*$  is the adjunct operator of  $P$ . Now applying the inequality obtained above on  $\mathcal{E}$  finishes the claim. ♥

**Theorem 12.4 (Bound on  $\mathcal{L}^2$  distance via spectral profile)** For a chain  $Q$  with invariant distribution  $\mu$ , define  $V_Q(t)$  to be the quantity defined by

$$\int_{4\mu_{\min}}^{V_Q(t)} \frac{du}{u\Lambda_Q(u)} = t$$

Then for a Markov chain  $P$  with  $P^*P$  and  $PP^*$  both irreducible, we have that

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

**Main idea:** We know that we can express  $d_2(t)^2$  as  $\text{Var}_\pi(P^{*t} f_x)$ . This latter quantity can be thought of as a function of  $t$ . Extending linearly allows us to look at the value of the derivative. From this we can rearrange and integrate over time to obtain a  $t$ , and remove this derivative. Careful manipulation gives the result.

*Proof.* To bound  $d_2(t)$  we study  $\left\| \frac{P^t(x, \cdot)}{\pi(\cdot)} - 1 \right\|_2^2$  which can be expressed as a variance, let us go over the computation again for the sake of clarity, recall that  $f_x = \frac{\mathbf{1}(\{x\})}{\pi(x)}$ :

$$\begin{aligned}
 \left\| \frac{P^t(x, \cdot)}{\pi(\cdot)} - 1 \right\|_2^2 &= \sum_{y \in \Omega} \left( \frac{P^t(x, y)}{\pi(y)} - 1 \right)^2 \pi(y) \\
 &= \sum_{y \in \Omega} \left( \frac{P^t(x, y)}{\pi(y)} \right)^2 \pi(y) - 1 && \text{(Expanding the bracket)} \\
 &= \sum_{y \in \Omega} \left( \frac{P^{*t}(y, x)}{\pi(x)} \right)^2 \pi(y) - 1 && \text{(Time reversing)} \\
 &= \sum_{y \in \Omega} (P^{*t} f_x)^2 \pi(y) - 1 && \text{(Previous computation)} \\
 &= \text{Var}_\pi[P^{*t} f_x] && \text{(Since } \mathbf{E}_\pi[P^{*t} f_x] = 1)
 \end{aligned}$$

Then in view of Lemma 12.3 (Variational contraction for spectral profile) we have that

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

Now we let  $I_x(m) = \text{Var}_\pi[P^{*m} f_x]$ , and we note that  $I_x$  is **non-increasing** (due to variational contraction). We can extend  $I_x$  linearly so that for  $t \in (m, m+1)$

$$I_x(t) = I_x(m) + (t - m)(I_x(m+1) - I_x(m))$$

Therefore for  $t \in (m, m+1)$ , using the variational contraction Lemma again we have that

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

Now since  $t > m$ , we have that  $I_x(t) \leq I_x(m)$ , and similarly  $\Lambda_{PP^*}(4/I_x(m)) \geq \Lambda_{PP^*}(4/I_x(t))$ , so putting it all together we have that

$$I'_x(t) \leq -\frac{1}{2} I_x(t) \Lambda_{PP^*} \left( \frac{4}{I_x(t)} \right)$$



Rearranging gives

$$\int_0^t \frac{I'_x(t)}{I_x(t) \Lambda_{PP^*}\left(\frac{4}{I_x(t)}\right)} dt \leq -\frac{1}{2}t$$

Performing the substitution  $u = 4/I_x(t)$  to clear the argument of  $\Lambda_{PP^*}$  gives that

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

Now it is easy to verify that

$$I_x(0) = \text{Var}_\pi(f_x) = \frac{1}{\pi(x)} - 1 \leq \frac{1}{\pi_{\min}} - 1 \leq \frac{1}{\pi_{\min}}$$

So we have that

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

But recall that by definition of  $V_{PP^*}$

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

Now comparing the integrals, and using the fact that the integrand is non-negative, we immediately get that

$$V(t/2) \leq 4/I_x(t)$$

Rearranging and using the definition of  $I_x(t)$  gives:

$$V_{PP^*}(t/2) \leq \frac{4}{I_x(t)} = \frac{4}{\text{Var}_\pi(P^{*t} f_x)} = \frac{4}{(d_2(t)^2)}$$

This gives the desired result.



**Remark 12.5** This computation was long, but in summary, we know that in time

$$\int_{4\pi_{\min}}^{4/\epsilon} \frac{du}{u \Lambda_{PP^*}(u)}$$

The  $\mathcal{L}^2$  distance drops down to  $\epsilon$ . Since  $\mathcal{L}^2$  distance upper bounds  $\mathcal{L}^1$  (and hence total variation) distance, we also get a bound on the total variation distance, however this is only useful if we can find the spectral profile of  $PP^*$  and we don't know whether this is easier to do than finding the spectral profile of  $P$ . The following result gives a way to bound mixing time if we can only compute

the spectral profile of  $P$ .

**Corollary 12.6 (Mixing time and spectral profile)** Suppose that there exists a constant  $\alpha > 0$  for which  $P(x, x) \geq \alpha$  for all  $x$  (for example, when  $P$  is  $\alpha$ -lazy), then for  $\epsilon > 0$ , the  $\mathcal{L}^\infty$  mixing time satisfies

$$t_{\text{mix}}^{(\infty)} \leq 2 \left[ \int_{4\pi_{\min}}^{4/\epsilon} \frac{du}{\alpha u \Lambda_P(u)} \right] \leq 2 \left[ \int_{4\pi_{\min}}^{4/M} \frac{du}{\alpha u \Lambda_P(u)} \right] + 2 \left[ \frac{1}{\alpha \gamma(P)} \log(M/\epsilon) \right]$$

**Main idea:** First we note a relation between  $d_\infty$  and  $d_2$ . Since we know by the previous theorem that  $d_2$  depends on  $V_{PP^*}$ , we wish to find a comparison between  $V_P$  and  $V_{PP^*}$ . This can be done because the simple observation that  $P^*(x, x) = P(x, x)$  gives a comparison between  $\mathcal{E}_{PP^*}$  and  $\mathcal{E}_P$ .

*Proof.* Recall that  $d_\infty(2t)$  can be expressed as

$$\begin{aligned} \left| \frac{P^{2t}(x, y)}{\pi(y)} - 1 \right| &= \left| \sum_{z \in \Omega} \frac{P(x, z)P(z, y)}{\pi(y)} - 1 \right| \\ &= \left| \sum_{z \in \Omega} \left( \frac{P(x, z)}{\pi(z)} - 1 \right) \left( \frac{P(z, y)}{\pi(y)} - 1 \right) \pi(z) \right| && \text{(Standard trick)} \\ &= \left| \sum_{z \in \Omega} \left( \frac{P(x, z)}{\pi(z)} - 1 \right) \left( \frac{P^*(y, z)}{\pi(z)} - 1 \right) \pi(z) \right| && \text{(Time reversal)} \\ &\leq \|P^t(x, \cdot) - \pi\|_2 \|P^{*t}(y, \cdot) - \pi\|_2 && \text{(Cauchy-Schwarz)} \end{aligned}$$

Using the previous Theorem, which allows us to characterise  $\mathcal{L}^2$  distance in terms of  $V_{PP^*}$  we get that

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

With this in mind, we wish to relate  $V_{PP^*}$  (and  $V_{P^*P}$ ) to  $V_P$ . This is done by comparing the Dirichlet forms, and in turn, this can be done first noting that since  $P^*(x, x) = P(x, x) \geq \alpha$ , we see that (This is a trick that appears on ES2)

$$\begin{aligned} PP^*(x, y)\pi(x) &= \sum_{z \in \Omega} P(x, z)P^*(z, y)\pi(x) \\ &\geq (P(x, x)P^*(x, y) + P(x, y)P^*(y, y))\pi(x) \\ &\geq \alpha(P(y, x)\pi(y) + P(x, y)\pi(x)) \end{aligned}$$

and so we can plug into the Dirichlet form:

$$\begin{aligned}\mathcal{E}_{PP^*}(f) &\geq \alpha \frac{1}{2} \sum_{x,y} (f(x) - f(y))^2 (P(x,y)\pi(x) + P(y,x)\pi(y)) \\ &= 2\alpha \mathcal{E}_P(f)\end{aligned}\quad (\text{Splitting and relabelling indices})$$

This comparison of the Dirichlet form automatically gives a comparison on the Spectral profile:  $\Lambda_{PP^*}(r) \geq 2\alpha \Lambda_P(r)$ . Now using the definition of  $V_P(\alpha t)$  we see that

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

i.e: we have that

$$\frac{t}{2} \geq \int_{4\pi_{\min}}^{V_P(\alpha t)} \frac{du}{u\Lambda_{PP^*}(u)}$$

which means (since the integrand is non-negative) that

$$V_{PP^*}(t/2) \geq V_P(\alpha t)$$

Of course a similar argument implies the same for  $V_{P^*P}$ . Now plugging back into the expression that related  $d_\infty(2t)$  to  $V_{PP^*}$  and  $V_{P^*P}$  gives that

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

And so if we require that

$$\left| \frac{P^{2t}(x,y)}{\pi(y)} - 1 \right| \leq \epsilon$$

then we can require that

$$\frac{4}{V_P(\alpha t)} \leq \epsilon$$

and this is equivalent to asking that

$$t \geq \left\lceil \int_{4\pi_{\min}}^{4/\epsilon} \frac{du}{\alpha u \Lambda(u)} \right\rceil$$

as required. ♡

We now have found a general upper bound on the mixing time, but the question remains as to whether this will give any better bounds on the mixing time compared to anything we had before, and we also have the problem of computing an infimum over all non-negative functions supported on a certain set, this clearly doesn't seem very computationally friendly. It turns out that we have a spectral representation of the spectral profile, which much in a similar manner to how the spectral gap, which

could easily be computed by finding eigenvalues, was the same as the Poincaré constant for reversible chains, will give us a way to estimate  $\Lambda(u)$  by looking at eigenvalues of a certain matrix.

**Lemma 12.7** (Spectral representation of the spectral profile ) For any  $A \subseteq \Omega$  and transition matrix  $P$ , we define

$$\lambda_0(A) = \inf_{f \in c_0(A)} \frac{\mathcal{E}_{P_A}(f)}{\|f\|_2^2} \quad P_A(i, j) = P(i, j) \text{ if } i, j \in A, 0 \text{ otherwise.}$$

Where  $c_0 = \{f : \text{supp}(f) \subseteq A, f \text{ non-constant}\}$ . For every chain we have that

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

Moreover, for  $r \geq \pi_{\min}$ , define  $\Lambda_0(r) = \inf_{\pi(A) \leq r} \lambda_0(A)$ . Then for a reversible chain we have that

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

and in the reversible case, we also have that  $\lambda_0(A)$  is the smallest eigenvalue of the matrix  $I - P_A$

*Proof.* ES3



We have another useful result for computing the spectral profile.

**Lemma 12.8** (Spectral profile is the infimum over functions with connected support) Suppose that for a set  $A \subseteq \Omega$  we have the decomposition into connected components

$$A = \bigcup_{i=1}^k A_i$$

Then  $\lambda(A) = \min_{i \leq k} \lambda(A_i)$ .

*Proof.* From definition of  $\lambda$ , it is clear that  $\lambda(A) \leq \lambda(A_i)$  and as such  $\lambda(A) \leq \min_{i \leq k} \lambda(A_i)$ . We now need to show that  $\lambda(A) \geq \min_{i \leq k} \lambda(A_i)$ . Since by assumption the  $(A_i)$ s are disjoint, we have that for any function  $f$  defined on  $A$ ,  $f = \sum_{i=1}^k f \mathbf{1}_{(A_i)} \equiv \sum_{i=1}^k f_i$ . From this we can compute the ingredients needed to compute  $\lambda(A)$ :

- Variance:

$$\begin{aligned}
 \text{Var}_\pi(f) &= \text{Var}_\pi\left(\sum_{i=1}^k f_i\right) \\
 &= \mathbf{E}\left[\left(\sum_{i=1}^k f \mathbf{1}(A_i)\right)^2\right] - \left(\sum_{i=1}^k \mathbf{E}[f \mathbf{1}(A_i)]\right)^2 \\
 &= \mathbf{E}\left[\sum_{i=1}^k f_i^2 + \sum_{i \neq j} f^2 \underbrace{\mathbf{1}(A_i) \mathbf{1}(A_j)}_{\delta_{i,j}}\right] - \left(\sum_{i=1}^k \mathbf{E}[f \mathbf{1}(A_i)]\right)^2 \\
 &= \sum_{i=1}^k \mathbf{E}[f_i^2] - \sum_{i=1}^k \mathbf{E}[f_i]^2 - \underbrace{\sum_{i \neq j} \mathbf{E}[f_i] \mathbf{E}[f_j]}_{\geq 0} \\
 &\leq \sum_{i=1}^k \text{Var}_\pi(f_i)
 \end{aligned}$$

- Dirichlet Form: it is clear that by linearity of inner product on first entry:  $\mathcal{E}(f) = \sum_{i=1}^k \mathcal{E}(f_i)$

Then putting it all together

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

Using the fact that for any  $f_i$   $\lambda(A_i) \leq \frac{\mathcal{E}(f_i)}{\text{Var}_\pi(f_i)}$ , we have that

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



We will now revisit the example of the lazy random walk on  $\mathbf{Z}/n\mathbf{Z}$  but through the lens of the Spectral Profile.

**Example 12.9 (LSRW on  $\mathbf{Z}/n\mathbf{Z}$  revisited)** The Lazy Simple Random Walk on  $\mathbf{Z}/n\mathbf{Z}$  has a mixing time of

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

*Proof.* We are going to use the previous results. In particular, by Lemma 12.8, we only need to bound  $\lambda(A)$  for connected sets  $A$ , and by Lemma 12.7, it suffices to bound  $\lambda_0(A)$  for such sets.

Since connected sets of the cycle are simply lines, the matrix  $P_A$  looks like

$$P_A = \begin{pmatrix} 1/2 & 1/4 & 0 & \cdots \\ 1/4 & 1/2 & 1/4 & \cdots \\ 0 & 1/4 & 1/2 & \cdots \\ 0 & 0 & 1/4 & \ddots \end{pmatrix}$$

It is an exercise in linear algebra that the smallest eigenvalue of the matrix  $I - P_A$  is given by

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

Therefore

$$\Lambda_0(r) = \inf_{\pi(A) \leq r} \cdots = \inf_{|A| \leq rn} \frac{1}{2} \left( 1 - \cos \frac{\pi}{|A|+1} \right) = \frac{1}{2} \left( 1 - \cos \frac{\pi}{[rn]+1} \right) \asymp \frac{1}{(rn)^2}$$

And using Lemma 12.7, we see  $\Lambda(r) \gtrsim \frac{1}{(rn)^2}$ , so using Corollary 12.6 we can bound the mixing time, using  $M = 8$  and  $\alpha = 1/2$  (lazy chain) by

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

♡

Of course, we already knew from other techniques that for the LSRW on  $\mathbf{Z}/n\mathbf{Z}$ , the mixing time was  $O(n^2)$ , now we see a nice application of this, which is that we can actually use the Spectral profile to bound the mixing time of a LSRW on a modified cycle, i.e. a cycle with more edges added.

**Example 12.10 (Mixing time of modified cycle)** Consider a LSRW on  $\mathbf{Z}/n\mathbf{Z}$ , to which we add any edges we want, but such that any vertex has degree at most  $\Delta > 2$ . Then we also have that  $t_{\text{mix}}(\epsilon) \lesssim n^2 \log(1/\epsilon)$

**Main idea:** Here is the overview: let  $\Lambda$  and  $\tilde{\Lambda}$  be the spectral profiles of the LSRW on  $G_n = \mathbf{Z}/n\mathbf{Z}$  and the modified graph respectively, and define  $\gamma$  and  $\tilde{\gamma}$  for the spectral gaps respectively. We are going to show that  $\tilde{\Lambda}(r) \gtrsim \Lambda(r)$  and that  $\tilde{\gamma} \gtrsim \gamma$ . From this, Theorem 12.6 and the computation of the previous example, will tell us that  $\widetilde{t_{\text{mix}}}^{(\infty)} = O(n^2)$ . To compare the quantities we are interested in, we will exploit the fact that we can give bounds on  $\tilde{\pi}(x)$  and will allow us conclude  $\tilde{\mathcal{E}} \gtrsim \mathcal{E}$  and  $\text{Var}_{\tilde{\pi}}[f] \lesssim \text{Var}_{\pi}[f]$ , then we will compare the spectral profiles and the spectral gaps.

*Proof.* Let  $P$  be the LSRW on the normal cycle, and  $\tilde{P}$  be the LSRW on the modified cycle. We are going to compare the spectral profiles of these two random walks by comparing variances and Dirichlet forms. First of all, we note the following bounds on the invariant distribution  $\tilde{\pi}$ :

$$\frac{1}{\Delta n} \leq \tilde{\pi}(x) \leq \frac{\Delta}{n}$$

Indeed: since  $\tilde{\pi}(x) = \frac{\deg(x)}{2|E|}$  we can on the one hand say that  $\deg(x) \leq \Delta$  and  $|E| \geq n$  (this is an obviously brutal bound, but its all we need), and from this the upper bound follows. For the lower bound, we note that  $\deg(x) \geq 2$  and  $|E| \leq \Delta n$ . From this the lower bound follows. It is also easy to see that  $\tilde{P}(x, y) \geq 1/(2\Delta)$  (since  $x$  has at most  $\Delta$  neighbors. With this in mind, we note that

$$\Delta^2 \tilde{\pi}(x) \tilde{P}(x, y) \geq \frac{1}{2n} \geq \pi(x) P(x, y)$$

and so we can compare Dirichlet forms:

$$2\mathcal{E}(f) = \sum_{x,y} \pi(x) P(x, y) (f(x) - f(y))^2 \leq \Delta^2 \sum_{x,y} \tilde{\pi}(x) \tilde{P}(x, y) (f(x) - f(y))^2 = 2\Delta^2 \tilde{\mathcal{E}}(f)$$

Now we can compare variances, we recall that we have a technique for comparing variances with respect to two measures  $\pi$  and  $\tilde{\pi}$ , but let me repeat the argument because these calculations need to be really nailed down

$$\begin{aligned} \text{Var}_{\tilde{\pi}}(f) &= \sum_x \tilde{\pi}(x) (f(x) - \mathbf{E}_{\tilde{\pi}}[f])^2 \\ &\leq \sum_x \tilde{\pi}(x) (f(x) - \mathbf{E}_{\pi}[f])^2 \quad (\text{Minimising argument, see 11.11}) \\ &= \sum_x \frac{\tilde{\pi}(x)}{\pi(x)} \pi(x) (f(x) - \mathbf{E}_{\pi}[f])^2 \\ &\leq \max_x \left( \frac{\tilde{\pi}(x)}{\pi(x)} \right) \text{Var}_{\pi}(f) \end{aligned}$$

And from the upper bound on  $\tilde{\pi}(x)$ , we see that

$$\max_x \frac{\tilde{\pi}(x)}{\pi(x)} \leq \Delta$$

So putting into the definition of  $\tilde{\lambda}(A)$ , we get that

$$\tilde{\lambda}(A) = \inf_{f \in A} \frac{\tilde{\mathcal{E}}(f)}{\text{Var}_{\tilde{\pi}}(f)} \gtrsim \lambda(A)$$

Now we can compare spectral profiles and since  $\tilde{\pi}(A) \geq \frac{1}{\Delta} \pi(A)$ , we have that

$$\{\tilde{\pi}(A) < r\} \subseteq \{\pi(A) < \Delta r\}$$

we see that

$$\tilde{\Lambda}(r) = \inf_{\tilde{\pi}(A) < r} \tilde{\lambda}(A) \gtrsim \inf_{\pi(A) < \Delta r} \lambda(A) = \Lambda(\Delta r) \gtrsim \frac{1}{(nr^2)}$$

Now all left for us to be able to apply Theorem 12.6 is to compare Dirichlet constants  $\tilde{\gamma}$  and  $\gamma$ , but using the definition of the Poincaré constant and the bounds we have derived on the Dirichlet form and the variance it follows that  $\tilde{\gamma} \gtrsim \gamma$ , so using now finally Theorem 12.6, it follows that the mixing time of this modified chain is also  $O(n^2)$



**Remark 12.11** This is finally an example of where the spectral profile provides a better bound than the previous techniques, because if we relied only on comparing Poincaré constants, we would have had to use Theorem 11.8 and we would have had to pay an extra price of  $\log(1/\sqrt{\pi_{\min}})$

Before moving on, we include a summary of this chapter so far:



1. We define for a subset  $A \subseteq \Omega$ ,  $\lambda(A) = \inf_{f \in C_0^+(A)} \frac{\mathcal{E}(f)}{\text{Var}_\pi(f)}$ . Where  $C_0^+(A)$  is the set of non-negative non-constant functions supported on  $A$ . We then define

$$\Lambda(u) = \inf_{\pi(A) \leq u} \lambda(A),$$

this should be seen as a generalisation of the Poincaré constant.

2. In Lemma 12.3 we see an inequality that is analogous to the Poincaré inequality of the Spectral techniques section:

$$\text{Var}_\pi(f) - \text{Var}_\pi(P^*f) \geq \frac{1}{2} \text{Var}_\pi(f) \Lambda\left(\frac{4\mathbf{E}_\pi(f)^2}{\text{Var}_\pi(f)}\right)$$

3. By now considering a piece-wise linear extension of  $\text{Var}_\pi(P^{*t}f_x)$ , where  $f_x(\cdot) = \mathbf{1}\{x\}/\pi(\cdot)$ , and using this variational contraction, we reach the following conclusion: if we let  $V(t)$  be the number for which

$$t = \int_{4\pi_{\min}}^{V(t)} \frac{du}{u\Lambda(u)},$$

then  $d_2(t)^2 \leq \frac{4}{V_{PP^*}(t/2)}$ .

4. By now inspecting  $d_\infty(2t)$ , we see that if the chain is lazy (we use the laziness to compare Dirichlet forms of  $PP^*$  and  $P$ , we have that

$$t_{\text{mix}}(\epsilon) \lesssim \int_{4\pi_{\min}}^{4/\epsilon} \frac{du}{u\Lambda(u)} \lesssim \int_{4\pi_{\min}}^{4/M} \frac{du}{u\Lambda(u)} + \frac{1}{\gamma_P} \log(M/\epsilon).$$

5. For this to be useful we need to be able to control  $\Lambda(u)$ , but taking infimum over sets of functions is not easily done, so we have spectral characterisation: if we let  $\lambda_0(A) = \inf_{f \in C_0(A)} \frac{\mathcal{E}_{P_A}(f)}{\|f\|_2^2}$ , where  $P_A(i, j) = P(i, j) \mathbf{1}\{i, j \in A\}$ , and define  $\Lambda_0(r) = \inf_{\pi(A) \leq r} \lambda_0(A)$ , we have that

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

Moreover,  $\lambda_0(A)$  is also the smallest eigenvalue of the transition matrix  $I - P_A$ .

6. Finally, it is easy to show that to compute the infimum  $\Lambda_0(r) = \inf_{\pi(A) \leq r} \lambda_0(A)$ , it suffices to consider connected sets, this was the content of Lemma 12.8.

We now learn about another way of bounding the spectral profile, that instead of relying on comparing Dirichlet forms and variances, relies on the underlying geometry of the chains. This is quite similar to how the Bottleneck ratio bounded the Poincaré constant through Cheeger's inequality, and in fact, this new technique, the Isoperimetric Profile, is nothing but a direct generalisation of the Bottleneck

ratio, and it will also lead to a somewhat general version of Cheeger's inequality.

## 12.2 Isoperimetric Profile

Recall how the Bottleneck ratio was defined as the worst possible conductance of any non-empty set in the chain, this of course admits an immediate generalisation, very similar in spirit to the definition of the Spectral profile:

**Definition 12.12 (Isoperimetric profile)** Let  $P$  be a transition matrix, the Isoperimetric profile is the function  $\Phi_* : [\pi_{\min}, \infty) \rightarrow \mathbf{R}$  given by

$$\Phi_*(r) = \begin{cases} \inf_{\pi_{\min} \leq \pi(A) \leq r} \Phi(A) & r \leq 1/2 \\ \Phi_*(1/2) & r > 1/2 \end{cases}$$

Here we see the corresponding result to Cheeger's inequality for the Isoperimetric profile

**Theorem 12.13 (Generalised Cheeger's Inequality)** For  $r \in [\pi_{\min}, 1/2)$ , we have that

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

*Proof.* [Finish this section once ES3 falls]



The following corollary, which follows from the lower bound of  $\Lambda(r)$  given by the Generalised Cheeger's Inequality and the bound of the Poincaré constant by the Spectral profile and then bounding this by the Isoperimetric profile, gives us a useful way to bound mixing times

**Corollary 12.14 (Mixing time and Isoperimetric Profile)** Suppose that for some  $\alpha > 0$  we have that  $P(x, x) \geq \alpha$  (as is the case for  $\alpha$ -lazy chains), then for  $\epsilon > 0$ , the  $\mathcal{L}^\infty$  mixing time (and hence the rest of mixing times) satisfy

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

And in a similar spirit to Theorem 12.6, we also have that

$$t_{\text{mix}}^{(\infty)} \leq 2 \left[ \int_{4\pi_{\min}}^{4/M} \frac{2du}{\alpha u \Phi_*^2(u)} \right] + 2 \left[ \frac{2}{\alpha \Phi_*^2} \log(M/\epsilon) \right]$$

*Proof.* The proof follows by lower bounding the spectral profile and Poincaré constant in Corollary 12.6 by the isoperimetric profile and the bottleneck ratio as per Generalised Cheeger's inequality and Cheeger's inequality.

♡

An example of an application of this Theorem is a bound on the mixing time of a special class of graphs

**Example 12.15 (Small set expander mixing time)** A family of graphs  $G_n$  is called a  $(\Delta, \alpha, c)$ -small set expander family if for all  $n$  we have that the maximal degree in  $G_n$  is  $\Delta$  and that  $\Phi_*(c) \geq \alpha$ . Consider a LSRW on a small set expander family. The mixing time satisfies

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

*Proof.* We apply the Mixing-time and Isoperimetric Profile result, which states that for any  $M > 0$ ,

$$t_{\text{mix}}^{(n)}(\epsilon) \lesssim \int_{4\pi_{\min}}^{4/M} \frac{du}{u \Phi_*^2(u)} + \frac{1}{\gamma(P)} \log(M/\epsilon)$$

Therefore, by the small set expander property, we have some  $c > 0$  such that  $\Phi_*(c) \gtrsim 1$ , and as the chain is lazy and reversible,  $1/\gamma(P)$  is actually  $t_{\text{rel}}$ , so by combining all of this, we have that setting  $M = 4/c$ :

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

Where we have used that  $\pi(x) = \frac{\deg(x)}{2|E|} \asymp \frac{1}{n}$

♡

As a final application of the techniques involving the Isoperimetric profile, we give bounds on the transition probabilities of LSRW on bounded degree graphs

**Proposition 12.16** (Transition probabilities for random walks) Let  $G$  be a connected graph on  $n$  vertices with degree bounded by  $\Delta$  and let  $P$  be the transition matrix of a LSRW on  $G$ . Then for all  $t \in \mathbb{N}$ , all vertices  $x$  and  $y$  we have that

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

*Proof.* We first show a bound on the  $\mathcal{L}^\infty$  mixing time, and then apply it to get the desired bound. To bound the  $\mathcal{L}^\infty$  mixing time, we will employ the Isoperimetric Profile bound, which will become useful for the simple structure of this graph will make it feasible to obtain appropriate bounds for  $\Phi_*(r)$ . We start by proving a lower bound on  $\Phi_*(A)$ , the bottleneck ratio, for a set  $A$ . For this we will employ a brutal bound.

$$\Phi_*(A) = \frac{Q(A, A^c)}{\pi(A)} \geq Q(A, A^c) = \frac{|\partial A|}{2E(A)} \geq \frac{1}{2E(A)}$$

Where the first inequality comes from brutally noting that  $\pi(A) \leq 1$ , the middle equality is by a computation done in a previous example, and the final inequality comes from the fact that since the graph is connected, any set  $A$  will have at least one edge coming out of it. Also  $E(A) := \sum_{x \in A} \deg(x)$ . We can now relate  $\Phi(A)$  to  $\pi(A)$ . Indeed:  $\pi(A) = \frac{E(A)}{2|E|}$ , so by rearranging we have that

$$\Phi_*(u) = \inf_{\pi(A) \leq u} \Phi(A) \geq \frac{1}{4\pi(A)|E|} = \frac{1}{4u|E|}$$

With this in mind, we can plug this into the Isoperimetric Profile Integral, and see that for  $M \geq 8$  (This is so that what's going into  $\Phi_*$  is at most  $1/2$ )

$$t_{\text{mix}}^{(\infty)}(M) \lesssim \int_{4/\pi_{\min}}^{4/M} \frac{dx}{x\Phi_*(x)^2} \lesssim \int_{4/\pi_{\min}}^{4/M} \frac{dx}{x\left(\frac{1}{\Delta n x}\right)^2} \lesssim \frac{\Delta^2 n^2}{M^2}$$

This last inequality is because  $\pi_{\min} \asymp \frac{1}{n}$  so it doesn't really matter. From this, we have that there is some constant  $C$  with

$$t_{\text{mix}}^{(\infty)}(M) \leq C \frac{\Delta^2 n^2}{M^2}$$

So for a fixed time  $t$ , we have that choosing

$$M = \frac{2\Delta n \sqrt{C}}{\sqrt{t}}$$

Gives, without loss of generality (Just choose  $C$  larger if needed), that  $M \geq 8$ , and that

$$t = 4C \frac{\Delta^2 n^2}{M^2} \frac{\Delta^2 n^2}{M^2}$$

Which means that

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

From this, we can employ the definition of  $\mathcal{L}^\infty$  distance, and see that for any  $x, y$ , we have that

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

Now, noting that this being a random walk on a graph we have that for all  $x$ :

$$\pi(x) = \frac{\deg(x)}{2|E|} \leq \frac{\Delta}{n}$$

We can rewrite the above as

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



The summary for this second part of the chapter is shorter.

1. If we define  $\Phi_*(u) = \inf_{\pi(A) \leq u} \Phi(A)$ , where  $\Phi(A) = Q(A, A^c)/\pi(A)$  is the usual bottleneck ratio, we then achieve a generalised Cheeger Inequality:

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

This, and in particular the lower bound, gives us a way, in combination with the earlier part of this chapter, to bound mixing times using the geometry of the chain:

$$t_{\text{mix}}(\epsilon) \lesssim \int_{4\pi_{\min}}^{4/\epsilon} \frac{du}{u\Phi_*(u)}$$



# Chapter 13

## Geometric Techniques

### 13.1 Varopoulos-Carne bound

**Definition 13.1** (Distance of states ) Let  $x, y \in \Omega$  be two states of a Markov chain  $P$ . We define the distance as

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

**Remark 13.2** (Distance of states is a metric) The function  $\text{dist}(x, y)$  is a metric if the chain is reversible.

*Proof.* Non-negativity is obvious, if the chain is reversible, then

$$P^t(x, y) = \frac{\pi(y)}{\pi(x)} P^t(y, x)$$

and since  $\pi$  is always positive, we have that  $P^t(x, y) > 0$  if and only if  $P^t(y, x) > 0$  which means that  $\text{dist}(x, y)$  and  $\text{dist}(y, x)$  take infimum over the same set. As for the triangle inequality, all we need to show is that  $t := \text{dist}(x, y) + \text{dist}(y, z)$  has that  $P^t(x, z) > 0$ , then it will follow that  $t \geq \text{dist}(x, z)$ . For this:

$$\begin{aligned} P^t(x, z) &= \sum_w P^{\text{dist}(x, y)}(x, w) P^{\text{dist}(y, z)}(w, z) \\ &= \underbrace{P^{\text{dist}(x, y)}(x, y) P^{\text{dist}(y, z)}(y, z)}_{>0} + \underbrace{\sum_{w \neq y} P^{\text{dist}(x, y)}(x, w) P^{\text{dist}(y, z)}(w, z)}_{\geq 0} > 0 \end{aligned}$$



We would like some control on how far the walk gets after  $t$  steps, and for this we will show a bound on

$P^t(x, y)$  in terms of  $\text{dist}(x, y)$ . For proving a general result it will be useful to first study the following bound for a SRW on  $\mathbf{Z}$ .

**Proposition 13.3** (Distance of a SRW on  $\mathbf{Z}$  from origin) Let  $X_t$  be a SRW on  $\mathbf{Z}$  from the origin. Then for all  $t, d \in \mathbf{N}$ , we have that

$$\mathbf{P}(|X_t| \geq d) \leq 2 \exp\left(-\frac{d^2}{2t}\right)$$

**Main idea:** Use a Chernoff bound

*Proof.* We first note that since  $X_t$  is **symmetric**, we have that  $\mathbf{P}(|X_t| \geq d) = \mathbf{P}[X_t \geq d] + \mathbf{P}[X_t \leq -d] = 2\mathbf{P}[X_t \geq d]$ , and so to compute this quantity we can use a Chernoff Bound:

$$\begin{aligned} \mathbf{P}[X_t \geq d] &= \mathbf{P}[\exp(\lambda X_t) \geq \exp(\lambda d)] \\ &\leq \exp(-\lambda d) \mathbf{E}[\exp(\lambda X_t)] && \text{(Markov's Inequality)} \\ &= \exp(-\lambda d) \mathbf{E}\left[\exp\left(\lambda \sum_{i=1}^t \eta_i\right)\right] \\ &= \exp(-\lambda d) \prod_{i=1}^t \mathbf{E}[\exp(\lambda \eta_i)] \\ &= \exp(-\lambda d) \left(\frac{\exp(\lambda) + \exp(-\lambda)}{2}\right)^t \\ &\stackrel{(!)}{\leq} \exp(-\lambda d) \exp\left(\frac{\lambda^2 t}{2}\right) \\ &\stackrel{(!!)}{\leq} \exp\left(-\frac{d^2}{2t}\right) \end{aligned}$$

Where step (!) comes from the following useful found:

$$\frac{\exp(\lambda) + \exp(-\lambda)}{2} = \sum_{k=0}^{\infty} \frac{\lambda^{2k}}{(2k)!} \leq \sum_{k=0}^{\infty} \frac{\lambda^{2k}}{2^k k!} = \exp\left(\frac{\lambda^2}{2}\right)$$

where the inequality comes from the fact that  $(2k)! = (2k)(2k-1)\cdots k! \geq 2 \times 2 \times \cdots \times k!$ . Step (!!)

simply comes from minimising  $-\lambda d + \lambda^2 t/2$ , which is minimised for  $\lambda = d/t$ . ♡

We now have a general result for transition probabilities of reversible chains



**Theorem 13.4 (Varopoulos-Carne bound)** Let  $P$  be reversible with invariant distribution  $\pi$ . Then for all states  $x, y$  and all  $t \in \mathbf{N}$ , we have that

$$P^t(x, y) \leq \sqrt{\frac{\pi(y)}{\pi(x)}} \mathbf{P}(|Z_t| \geq \text{dist}(x, y)) \leq 2 \sqrt{\frac{\pi(y)}{\pi(x)}} \exp\left(-\frac{\text{dist}(x, y)^2}{2t}\right)$$

Where  $Z_t$  is a SRW on  $\mathbf{Z}$  started at the origin.

**Main idea:** We use Chebyshev's polynomials, which give the key result that

$$z^t = \sum_{k=0}^t \mathbf{P}[|Z_t| = k] q_k(z)$$

Then we can use this polynomial expression for  $P$  instead of  $z$ , and then finally show that

$$q_k(P)(x, y) \leq \sqrt{\frac{\pi(y)}{\pi(x)}}$$

This comes from the fact that whenever a polynomial  $h$  has that  $h[-1, 1] \subseteq [-1, 1]$ , we have that

$$\langle h(P)f, g \rangle \leq \|f\|_2 \|g\|_2$$

Then we can use  $h = q_k$ ,  $f = \mathbf{1}_x$  and  $g = \mathbf{1}_y$  and the claim follows.

*Proof.* Define the Chebyshev polynomials  $\{q_k\}_{k \geq 0}$ :

$$q_{k+1}(z) := 2z q_k(z) - q_{k-1}(z) \quad k \geq 1,$$

With initial conditions  $q_0(z) = 1$  and  $q_1(z) = z$ . We are going to show the following: if  $\{Z_t\}$  is a simple random walk on  $\mathbf{Z}$ , then as polynomials, we have the equality

$$z^t = \sum_{k=0}^t \mathbf{P}[|Z_t| = k] q_k(z).$$

We recall the trigonometric identity

$$2 \cos(k\theta) \cos(\theta) = \cos((k+1)\theta) + \cos((k-1)\theta)$$

which shows that  $q_k(\cos \theta) = \cos(k\theta)$  for all  $\theta \in \mathbf{R}$  and  $k \in \mathbf{N}$ . Now we observe that

$$\begin{aligned}
 (\cos \theta)^t &= \left( \frac{\exp(i\theta) + \exp(-i\theta)}{2} \right)^t \\
 &= \mathbf{E}[\exp(i\theta Z_t)] = \mathbf{E}[\cos(\theta Z_t)] + i \mathbf{E}[\sin(\theta Z_t)] \\
 &= \sum_{k=-t}^t \mathbf{P}[Z_t = k] (\cos(\theta k) + i \sin(\theta k)) \\
 &\stackrel{(!)}{=} \sum_{k=0}^t \mathbf{P}[|Z_t| = k] \cos(\theta k) = \sum_{k=0}^t \mathbf{P}[|Z_t| = k] q_k(\cos \theta)
 \end{aligned}$$

For step (!), we note that since  $\sin(x) = -\sin(-x)$ , all the  $\sin(k\theta)$ 's in the sum will disappear, similarly, since  $\cos(-x) = \cos(x)$ , the terms  $\mathbf{P}[Z_t = k] \cos(\theta k)$  and  $\mathbf{P}[Z_t = -k] \cos(-\theta k)$  pair up to be  $\mathbf{P}[|Z_t| = k] \cos(\theta k)$ . Therefore, we have verified that for infinitely many values of  $z$  (as the equalities held for all  $\theta \in \mathbf{R}$ ), we have that

$$z^t = \sum_{k=0}^t \mathbf{P}[|Z_t| = k] q_k(z)$$

and if two polynomials agree on infinitely many values, they must be the same polynomial. Moving on, we apply this identity to our matrix  $P$ , to see that

$$P^t(x, y) = \sum_{k=\text{dist}(x, y)}^t \mathbf{P}[|Z_t| = k] q_k(P)(x, y)$$

(The reason for this new lower bound on the sum is that since  $q_k$  has degree  $k$  and by definition of  $\text{dist}(x, y)$ , whenever  $k \leq \text{dist}(x, y)$ ,  $q_k(P)(x, y) = 0$ ). It now remains to show that for all  $k$ ,

$$q_k(P)(x, y) \leq \sqrt{\frac{\pi(y)}{\pi(x)}}$$

The first observation we need is that  $|\langle q_k(P)f, g \rangle_\pi| \leq \|f\|_2 \|g\|_2$ , this we will prove at the end of this proof to not get messy calculations here now. Then in particular we can take  $f = \mathbf{1}_y$  and  $g = \mathbf{1}_x$ , and have that  $\pi(x) q_k(P)(x, y) \leq \sqrt{\pi(x) \pi(y)}$ , which gives the desired inequality and thus finishes the claim. Let us now show [the remaining](#): let  $\{\phi_i\}$  be an orthonormal basis of eigenvectors

of  $P$  and define  $a_i = \langle f, \phi_i \rangle$  and  $b_i = \langle g, \phi_i \rangle$

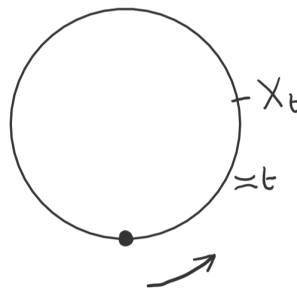
$$\begin{aligned}
 |\langle q_k(P)f, g \rangle| &= \left| \left\langle q_k(P) \sum_i a_i \phi_i, \sum_j b_j \phi_j \right\rangle \right| \\
 &= \left| \sum_{i,j} a_i b_j \langle q_k(P) \phi_i, \phi_j \rangle \right| \\
 &= \left| \sum_{i,j} a_i b_j q_k(\lambda_i) \langle \phi_i, \phi_j \rangle \right| \\
 &= \left| \sum_i a_i b_i q_k(\lambda_i) \right| \\
 &\stackrel{(1)}{\leq} \sum_i |a_i b_i| \\
 &\stackrel{(2)}{\leq} \|f\|_2 \|g\|_2
 \end{aligned}$$

Where (1) is from the fact that since  $\lambda_i \in [-1, 1]$  and we know that  $q_k(\cos \theta) = \cos(k\theta)$ , it follows that  $q_k([-1, 1]) \subseteq [-1, 1]$ , and so  $|q_k(\lambda_i)| \leq 1$ . Step (2) is Cauchy-Schwarz.  $\heartsuit$

**Remark 13.5** Let us admire for a second the VC bound, because it quantifies the following statement: If two states  $x$  and  $y$  are geometrically very far away, the probability to transition from  $x$  to  $y$  in  $t$  steps will be very low for small values of  $t$  but as  $t$  increases, the upper bound will start to relax.

**Remark 13.6 (On the assumption of reversibility)** Reversibility really is crucial in the VC bound:

*Sketch.* Indeed, consider the biased random walk on  $\mathbf{Z}/n\mathbf{Z}$ . By symmetry of the chain,  $\pi$  is uniform, so by checking detail balance equations it is easy to see that the chain is not reversible. Suppose that the chain has a drift to the right, as in the diagram, and is started at 0.



Then for large  $t$  (if  $n$  is still larger than  $t$ ), the Strong Law of Large Numbers tells us that with

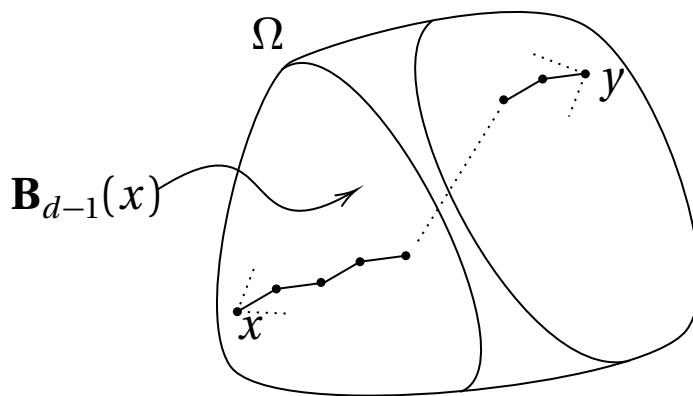


Figure 13.1: Diffusive bound: the diagram that says it all

probability 1,  $X_t$  will be very close to  $ct$  for some constant  $c$ . However, choosing a  $c' < c$  and plugging into the VC bound, given that time is very large, would give a very small probability of the walk being past the  $c't$  mark, contradicting the fact that the SLLN has prescribed that the  $X_t$  will be around the  $ct$  mark. ♡

Now we have a straightforward bound on the mixing time from the VC bound:

**Corollary 13.7 (Diffusive bound)** For any  $\epsilon \in (0, 1/2)$  we have that for large enough  $n$ , if  $P$  is a lazy simple random walk on a simple  $n$  vertex graph, then

$$t_{\text{mix}}(\epsilon) \geq \frac{\text{diam}(P)^2}{16 \log n}$$

**Main idea:** We are going to use a "dirty bound", letting  $x$  and  $y$  be the states that attain the diameter of the state-space, the key idea is to split the state-space by considering the balls around  $x$  and  $y$  of radius approximately half the diameter, then we will show that these balls are disjoint, hence they can't both have mass more than  $1/2$ , so we can control the mass of one of the balls, say the one centred at  $x$ . Similarly, since the balls give us a good control on the distance from  $x$ , using the VC bound, we can control the probability of exiting the ball around  $x$  in  $t$  steps, which gives us the lower bound on distance to stationarity.

*Proof.* Let  $x$  and  $y$  be two states such that  $\text{dist}(x, y) = \text{diam}(P)$  (need not be unique but we don't care). Then set  $d = \lceil \text{diam}(P)/2 \rceil$ . It is clear that  $d < \text{diam}(P)/2 + 1$  (convince yourself that  $\lceil x \rceil < x + 1$ ), then rearranging gives that  $\text{diam}(P) > 2(d - 1)$ . using this, we can show that  $\mathbf{B}_{d-1}(x) := \{z : \text{dist}(x, z) \leq d - 1\}$  and  $\mathbf{B}_{d-1}(y)$  are disjoint. Indeed: if  $z \in \mathbf{B}_{d-1}(x) \cap \mathbf{B}_{d-1}(y)$ , then  $\text{dist}(x, y) \leq \text{dist}(x, z) + \text{dist}(z, y) \leq (d - 1) + (d - 1) = 2(d - 1) < \text{diam}(P)$ , but we chose  $x$  and  $y$  to attain the diameter, so we get a contradiction. Since these two balls are disjoint, it must follow

than not both of them can have  $\pi$ -measure of strictly more than  $1/2$ , so without loss of generality, we just say that  $\pi(\mathbf{B}_{d-1}(x)) \leq 1/2$ , i.e:  $\pi(\mathbf{B}_{d-1}^c(x)) \geq 1/2$ . Now we try to control the probability that the walk started at  $x$  exits  $\mathbf{B}_{d-1}(x)$  in time  $t$ :

$$\begin{aligned}
 P^t(x, \mathbf{B}_{d-1}^c(x)) &= \sum_{w \in \mathbf{B}_{d-1}^c(x)} P^t(x, w) \\
 &\stackrel{(1)}{\leq} 2 \sum_{w \in \mathbf{B}_{d-1}^c(x)} \sqrt{\frac{\pi(w)}{\pi(x)}} \exp\left(-\frac{\text{dist}(x, w)^2}{2t}\right) \\
 &\stackrel{(2)}{\leq} 2 \sum_{w \in \mathbf{B}_{d-1}^c(x)} \sqrt{n} \exp\left(-\frac{\text{diam}(P)^2}{8t}\right) \\
 &\stackrel{(3)}{\leq} 2n^{3/2} \exp\left(-\frac{\text{diam}(P)^2}{8t}\right)
 \end{aligned}$$

Where (1) is the VC bound, (2) comes from the fact that  $\pi(z) = \deg(z)/n$ , and  $\deg(w)/\deg(x)$  can be bounded above by  $n$  because  $1 \leq \deg(z) \leq n$  (the upper bound on  $\deg(z)$  comes from the hypothesis of simple graph, i.e: no double edges, the lower bound comes from the fact that we always work with irreducible chains). Moreover, since  $w \in \mathbf{B}_{d-1}^c(x)$ , it follows that  $\text{dist}(x, w) \geq d = \lceil \text{diam}(P)/2 \rceil \geq \text{diam}(P)/2$ . Then step (3) comes from the fact that  $|\mathbf{B}_{d-1}^c(x)| \leq n$  (this is a brutal bound). Combining all this together:

$$\begin{aligned}
 d(t) &\geq \|P^t(x, \cdot) - \pi\|_{\text{TV}} \\
 &\geq |P^t(x, \mathbf{B}_{d-1}^c(x)) - \pi(\mathbf{B}_{d-1}^c(x))| \geq \frac{1}{2} - 2n^{3/2} \exp\left(-\frac{\text{diam}(P)^2}{8t}\right)
 \end{aligned}$$

Now if we choose  $t \leq \text{diam}(P)^2/(16 \log n)$ . we have that  $d(t) \geq 1/2 - 2\sqrt{n}$ , which for a fixed  $\epsilon < 1/2$ , is eventually greater than  $\epsilon$ , and so the claim follows.  $\heartsuit$

## 13.2 Path Coupling

Notice how the TV distance is blind to the geometric distances between states. Indeed: suppose that you have two states  $x$  and  $y$  that are reachable from each other in one step. Then in some sense, these two states are geometrically close, but if we consider the point masses  $\delta_x$  and  $\delta_y$ , these two distributions are at maximal distance in the eyes of TV. This brings us to the following definition:

**Definition 13.8 (Transportation metric)** Let  $\rho$  be a metric on a state space  $\Omega$ . We define the transportation metric between two distributions  $\mu, \nu \in \mathcal{P}(\Omega)$  as:

$$\rho_K(\mu, \nu) = \inf\{\mathbf{E}[\rho(X, Y)] : (X, Y) \text{ a coupling of } \mu, \nu\}$$

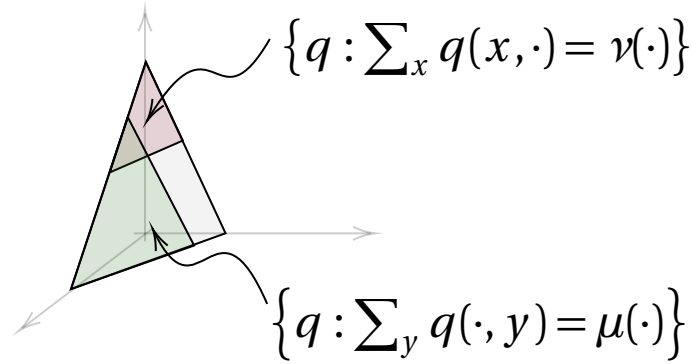


Figure 13.2: The picture that says it all:  $\rho$ -optimal coupling

**Remark 13.9** If  $\mu = \delta_x$  and  $\nu = \delta_y$ , then  $\rho_K(\mu, \nu) = \rho(x, y)$  because a coupling of  $\mu$  and  $\nu$  must be the constants  $x$  and  $y$ . Moreover, if  $\rho(x, y) = \mathbf{1}(x \neq y)$ , then  $\rho_K(\mu, \nu) = \|\mu - \nu\|_{TV}$ .

**Remark 13.10** (Is a metric required?) If instead of having  $\rho$  be a metric we just require it to be a non-negative function that satisfies the triangle inequality, then all of the results that are to follow would also hold, just without symmetry, this more general version could be more useful in cases where we want to consider a distance defined by  $\rho(x, y) = \min\{t : P^t(x, y) > 0\}$ , which need not be symmetric, if the chain is not reversible.

We begin with a Lemma that is analogous to the existence of a TV-optimal coupling.

**Lemma 13.11** (Optimal  $\rho$ -coupling) There exists a coupling  $q_*$  of  $\mu$  and  $\nu$  such that

$$\rho_K(\mu, \nu) = \sum_{(x,y) \in \Omega^2} q_*(x, y) \rho(x, y) = \mathbb{E}[\rho(X, Y)]$$

Where  $(X, Y) \sim q_*$ , i.e: that the infimum in the definition of  $\rho_K$  is attained.

*Proof.* The proof is by compactness. Indeed, a coupling  $q$  is nothing than a vector in the subspace of  $[0, 1]^{\Omega \times \Omega}$ . The fact that it is a probability measure, constrains  $q$  to live in the simplex of  $[0, 1]^{\Omega \times \Omega} - 1$  dimensions. This is a compact set. Indeed: it is clearly bounded, and moreover, the conditions of having the corresponding marginals make it closed. Indeed: consider the following subset  $A$  of the simplex:

$$A = \{q : \sum_x q(\cdot, x) = \mu(\cdot)\}$$

I claim this is closed. Indeed, suppose that  $\{q_n\}$  is a sequence in  $A$  with limit  $q$  in the simplex.

Let us show that  $q \in A$ . This is simple, we just need to check the marginal conditions:

$$\sum_x q(\cdot, x) = \sum_x \lim_{n \rightarrow \infty} q_n(\cdot, x) = \lim_{n \rightarrow \infty} \sum_x q_n(\cdot, x) = \mu(\cdot)$$

Where the limit was swapped due to finiteness. We have just established then, that the set of couplings of  $\mu$  and  $\nu$  form a compact subset of  $[0, 1]^{\Omega \times \Omega}$ . Therefore, since the map:

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

is continuous in  $q$ , it follows that its infimum is attained by our desired  $q_*$ . Of course we should also say that the set we are working in is non-empty, but an example of such a coupling is just the product measure  $\mu \otimes \nu$ . ♡

Of course there is something crucial that we need to justify now:

**Lemma 13.12 (Transportation metric is a metric)** The function  $\rho_K$  turns  $(\mathcal{P}(\Omega), \rho_K)$  into a metric space.

**Main idea:** Non-negativity and symmetry are obvious. To show that  $\rho_K(\mu, \nu) = 0$  if and only if  $\mu = \nu$ , show that if  $\rho_K(\mu, \nu) = 0$ , then their optimal coupling must be supported on the diagonal, and then show that this implies that  $\mu = \nu$ . Conversely, if  $\mu = \nu$ , just take the coupling  $(X, X)$  of  $(\mu, \nu)$ , it is clear that  $\mathbb{E}[\rho(X, X)] = 0$ . For the triangle inequality. Let  $p(x, y)$  and  $q(y, z)$  be a  $\rho$ -optimal coupling of  $(\mu, \nu)$  and  $(\nu, \eta)$ , then

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

is a coupling of  $(\mu, \nu, \eta)$ . From this one finishes the claim using the triangle inequality of the distance  $\rho$ .

*Proof.* We check each thing separately.

- Non-negativity: obvious. Moreover, if  $\rho_K(\mu, \nu) = 0$ , then for the optimal coupling, we have that

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

so if  $\rho(x, y) > 0$ , then  $q_*(x, y) = 0$ , which means that  $q_*$  has support on the diagonal  $\{(x, x) : x \in \Omega\}$ , which means that  $\mu = \nu$ . Indeed:

$$\mu(x) = \sum_y q_*(x, y) = q_*(x, x) = \sum_y q_*(y, x) = \nu(x)$$

Conversely, suppose that  $\mu = \nu$ . Then a coupling of  $\mu$  and  $\nu$  is simply to take the diagonal random variable  $(X, X)$  where  $X \sim \mu$ . Then since  $\rho_K$  is the infimum over all couplings, it follows that

$$0 \leq \rho_K(\mu, \nu) \leq \mathbf{E}[\rho(X, X)] = 0$$

Therefore  $\rho_K(\mu, \nu) = 0$

- Symmetry: follows from symmetry of  $\rho$ . (This is the only place where symmetry will be used, hence why morally we could drop symmetry from  $\rho$ )
- Triangle inequality: let  $\mu, \nu, \eta \in \mathcal{P}(\Omega)$ , then let  $p(x, y)$  be an optimal  $\rho$ -coupling of  $\mu$  and  $\nu$ , and let  $q(y, z)$  be an optimal  $\rho$ -coupling of  $\nu$  and  $\eta$ . Define

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

Then this is easily seen by summing on the marginals that  $r$  is a coupling of  $\mu, \nu, \eta$  and if we let  $(X, Y, Z) \sim r$ , then  $(X, Z)$  is a coupling of  $\mu$  and  $\eta$ , indeed: let's maybe for completeness check the one for  $\mu$ :

$$\sum_{z, y} \frac{p(x, y)q(y, z)}{\nu(y)} = \sum_y \frac{p(x, y)\nu(y)}{\nu(y)} = \mu(x)$$

and by the triangle inequality on  $\rho$ , we have that

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

since  $(X, Z)$  is a coupling of  $\mu$  and  $\eta$ , it follows that  $\rho_K(\mu, \eta) \leq \mathbf{E}[\rho(X, Z)]$ , which by combining with the expression above gives the triangle inequality.

♡

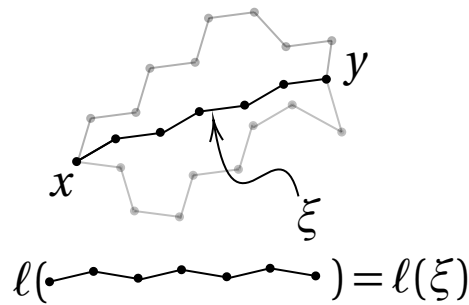


Figure 13.3: A silly figure to take some space



We now define a class of metrics on graphs:

**Definition 13.13 (Path metric)** For a connected graph  $G = (V, E)$ , and a *length function*  $\ell : E \rightarrow \mathbf{R}_+$  satisfying that for all  $(x, y) \in E$ ,  $\ell(x, y) \geq 1$ , we define a path metric corresponding to  $\ell$  to be

$$\rho(x, y) = \min \left\{ \sum_{i=0}^{r-1} \ell(x_i, x_{i+1}) : x_0 = x, x_1, x_2, \dots, x_r = y \text{ is a path } x \rightarrow y \right\}$$

where of course a path from  $x$  to  $y$  is understood to be a sequence of vertices that are connected.

We now use this new path metric in conjunction with the transportation metric defined before to bound the total variation distance:

**Proposition 13.14 (Transportation metric of a path metric bounds the TV distance)** Let  $\mu, \nu \in \mathcal{P}(V)$  and let  $G = (V, E)$  be a connected graph and let  $\rho$  be a path metric corresponding to some length function  $\ell$  on  $G$ . Then

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

Where  $\rho_K$  is the transportation metric induced by  $\rho$ .

*Proof.* Let  $(X, Y)$  be a coupling for  $\mu, \nu$ , then

$$\|\mu - \nu\|_{\text{TV}} \leq \mathbf{P}(X \neq Y) = \mathbf{E}[\mathbf{1}(X \neq Y)]$$

But since  $\ell \geq 1$ , and the graph is connected, any two vertices  $x$  and  $y$  will have a path between it of length at least 1. It follows that  $\mathbf{E}[\mathbf{1}(X \neq Y)] \leq \mathbf{E}[\rho(X, Y)]$ . Then minimising over all couplings  $(X, Y)$  gives the expression.  $\heartsuit$

Now we present a Theorem first discovered by Bubley and Dyer in 1997, which can be used to bound the mixing time under some regularity conditions:

**Theorem 13.15 (Mixing time bound in terms of the transportation metric)** Let  $G = (V, E)$  be a graph with distance function  $\ell$ . Suppose that  $X$  takes values on  $V$ . Let  $\rho$  be the corresponding path metric. Suppose that for all edges  $(x, y) \in E$ , we have that there exists some coupling  $(X_1, Y_1)$  of  $P(x, \cdot)$  and  $P(y, \cdot)$  such that for some  $\alpha$ :

$$\mathbf{E}_{x,y}[\rho(X_1, Y_1)] \leq \exp(-\alpha)\rho(x, y) \quad (*)$$

Then for any  $\mu, \nu \in \mathcal{P}(V)$ , we have that

$$\rho_K(\mu P, \nu P) \leq \exp(-\alpha)\rho_K(\mu, \nu) \quad (**)$$

In particular,

$$d(t) \leq \exp(-\alpha t) \text{diam}(V) \quad (***)$$

and if  $\alpha > 0$ , then

$$t_{\text{mix}}(\epsilon) \leq \frac{1}{\alpha} (\log \text{diam } V + \log(1/\epsilon))$$

**Main idea:** First show the result for  $\mu$  and  $\nu$  being point masses (of not necessarily adjacent points). To do so simply consider the path of shortest length from them and apply the hypothesis along with the triangle inequality. Then technically inequality (\*\*\*) already follows but, to prove inequality (\*\*), combine the optimal coupling  $\theta_{xy}$  for  $P(x, \cdot)$  and  $P(y, \cdot)$  found in part (\*) with an optimal coupling  $\eta$  of  $(\mu, \nu)$ . Then setting

$$\theta(w, z) = \sum_{x, y} \eta(x, y) \theta_{xy}(w, z)$$

gives a coupling of  $\mu P, \nu P$ . And so using the result of part (\*) and optimality of  $\eta$  finishes the claim.

*Proof.* We first establish the inequality (\*\*) for point masses  $\mu = \delta_x$  and  $\nu = \delta_y$ . Let  $x = x_0 \sim x_1 \sim \dots \sim x_k = y$  be the path from  $x$  to  $y$  of shortest length. Then by the triangle inequality:

$$\rho_K(\mu P, \nu P) = \rho_K(P(x, \cdot), P(y, \cdot)) \leq \sum_{i=0}^{k-1} \rho_K(P(x_i, \cdot), P(x_{i+1}, \cdot))$$

By hypothesis, for each  $x_i \sim x_{i+1}$ , there exists a coupling satisfying (\*), and since  $\rho_K(P(x_i, \cdot), P(x_{i+1}, \cdot))$  is the infimum over such couplings, we deduce that

$$\rho_K(P(x_i, \cdot), P(x_{i+1}, \cdot)) \leq \exp(-\alpha) \rho(x_i, x_{i+1})$$

By definition of  $\rho(x_i, x_{i+1})$  being the infimum over path lengths, we have that  $\rho(x_i, x_{i+1}) \leq \ell(x_i, x_{i+1})$  (it is not necessary that the direct edge from  $x_i$  to  $x_{i+1}$  minimises path distance, it could be that some more long winded route in terms of edges is actually "cheaper"), so putting it all together

$$\rho_K(P(x, \cdot), P(y, \cdot)) \leq \exp(-\alpha) \sum_{i=0}^{k-1} \ell(x_i, x_{i+1}) = \exp(-\alpha) \rho(x, y)$$

Where this last equality comes from the assumption that the path  $x_0 \sim x_1 \sim \dots \sim x_k$  is the one attaining the infimum in the path metric, thus proving (\*\*) for point masses. We now extend to general measures. The goal is to show that

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

So we will start by trying to find a coupling for  $\mu P$  and  $\nu P$ . Note that for any  $(x, y)$  we have the

existence of a coupling  $\theta_{x,y}$  of  $P(x, \cdot), P(y, \cdot)$  such that  $\mathbf{E}_{\theta_{x,y}}[\rho(X, Y)] \leq e^{-\alpha} \rho(x, y)$  (Indeed, by the Point Mass result, we have shown that for any  $x, y$ , we have that  $\rho_K(P(x, \cdot), P(y, \cdot)) \leq e^{-\alpha} \rho(x, y)$ , but since there always is a  $\rho_K$  optimal coupling, we have our desired  $\theta_{x,y}$ ), and since we know there exists some optimal coupling  $\eta$  of  $\mu, \nu$ , we can try combining these couplings to obtain a coupling of  $\mu P, \nu P$ . The probability distribution we will consider is

$$\theta := \sum_{x,y} \eta(x, y) \theta_{x,y}$$

Let us show this is indeed a coupling of  $\mu P$  and  $\nu P$ . (We will only show it for  $\mu P$  for simplicity.

$$\begin{aligned} \sum_{z \in \Omega} \theta(\omega, z) &= \sum_{z \in \Omega} \sum_{x,y \in \Omega} \eta(x, y) \theta_{x,y}(\omega, z) \\ &= \sum_{x,y} \eta(x, y) \sum_z \theta_{x,y}(\omega, z) \\ &= \sum_{x,y} \eta(x, y) P(x, \omega) = \sum_x \mu(x) P(x, \omega) \\ &= (\mu P)(\omega) \end{aligned}$$

Therefore, putting this all together:

$$\begin{aligned} \rho_K(\mu P, \nu P) &\leq \sum_{u,v} \rho(u, v) \theta(u, v) && (\theta \text{ is a coupling of } \mu P, \nu P) \\ &= \sum_{u,v} \sum_{x,y} \eta(x, y) \theta_{x,y}(u, v) \rho(u, v) \\ &= \sum_{x,y} \eta(x, y) \mathbf{E}_{\theta_{x,y}}[\rho(X, Y)] \\ &\leq e^{-\alpha} \sum_{x,y} \eta(x, y) \rho(x, y) && (\text{point mass result}) \\ &= e^{-\alpha} \rho_K(\mu, \nu) && (\eta \text{ is optimal}) \end{aligned}$$

To prove  $(\star\star\star)$  we simply note that

$$\begin{aligned} d(t) &\leq \bar{d}(t) = \max_{x,y} \|P^t(x, \cdot) - P^t(y, \cdot)\|_{\text{TV}} \\ &\leq \max_{x,y} \rho_K(P^t(x, \cdot), P^t(y, \cdot)) \\ &= \max_{x,y} \rho_K(\mathbf{1}_x P^t, \mathbf{1}_y P^t) \\ &\leq \max_{x,y} e^{-\alpha t} \rho(x, y) = e^{-\alpha t} \text{diam}(V) \end{aligned}$$



We now present a specific application of the above theory related to graph colourings.

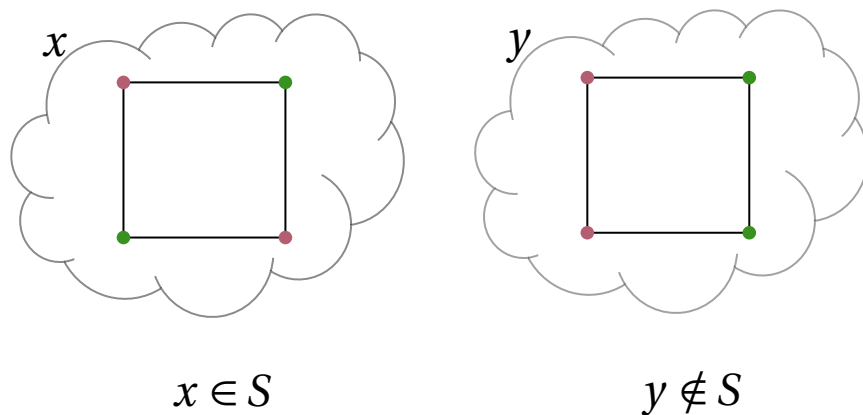


Figure 13.4: A proper colouring and a non-proper colouring of a graph with 4 vertices.

**Definition 13.16** (Proper colourings, Glauber dynamics) Let  $G = (V, E)$  be a graph. A proper vertex colouring of  $G$  with  $q$  colours is a function  $x \in \{1, \dots, q\}^V$  such that whenever two vertices  $v \sim w$  are neighbours, we require  $x(v) \neq x(w)$ . We let  $S$  be the set of proper colourings:

$$S = \{x \in \{1, \dots, q\}^V : x(u) \neq x(v), \text{ if } u \sim v\}$$

We define a Markov chain on  $S$ , the set of all proper colourings, using the Glauber dynamics, starting from a colouring  $x \in S$ :

1. Pick uniformly at random a vertex  $w \in V$
2. Update its colour to be, uniformly at random, from the allowed colours, i.e: from the colours not taken by any of its neighbours.

**Example 13.17** (Star graph) Consider the star graph of size  $n$ , i.e: one central node  $v^*$  with  $n - 1$  vertices coming out of it. As an initial example to get comfortable with Glauber dynamics and proper colourings, we will recap the tools of lower bounding mixing times with the Bottleneck ratio. Since these kinds of examples have two layers of abstraction, it is quite complicated to visualise the isoperimetry.

*Explanation.* Consider the subset  $A \subseteq S$  of proper  $q$ -colourings such that the root vertex has the label 1, i.e:  $A = \{x \in S : x(v^*) = 1\}$ . We now attempt to compute

$$Q(A, A^c) = \sum_{x \in A, y \in A^c} \pi(x) P(x, y).$$

Note that  $\pi(x)$  is naturally  $1/|S|$ , since the invariant distribution is uniform. Moreover, we have

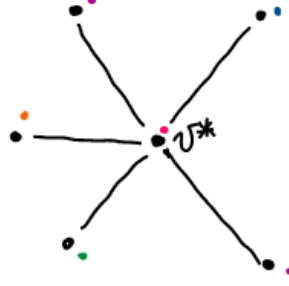


Figure 13.5: The star graph

that  $P(x, y) \leq 1/n$ , since  $y \notin A$  means that in particular, vertex  $v^*$  was chosen to be updated to go from  $x$  to  $y$  (we have an inequality because we would also have to guarantee that  $y(v^*) \neq 1$ ). Finally, the size of  $A \times A^c$  is at most  $(q-1)(q-2)^{n-1}$ , since the value of  $y(v^*)$  could take at most  $q-1$  values, and the remaining leaves, of which there are  $n-1$  of them, each can have  $q-2$  different colours (neither 1, nor whatever  $y(v^*)$  is). Hence we have that

$$Q(A, A^c) \leq \frac{1}{|S|n} (q-1)(q-2)^{n-1},$$

and since  $\pi(A) = (q-1)^{n-1}$ , since each of the  $n-1$  leaves can take  $q-1$  choices, we see that

$$\Phi(A) \leq \frac{(q-1)(q-2)^{n-1}}{n(q-1)^{n-1}},$$

from which now by rearranging and using the lower bound on mixing time using the Bottleneck ratio, we see that

$$t_{\text{mix}} \gtrsim \frac{n(q-2)}{(q-1)^2} \exp(n/q - 1)$$

♡

After this initial example, we now use the theory of Geometric Techniques to bound mixing times of Glauber dynamics under certain regularity conditions.

**Theorem 13.18 (Mixing time of Glauber dynamics on proper colourings)** Let  $G$  be a graph of  $n$  vertices with maximum degree  $\Delta$ . Let  $q > 2\Delta$ , then the mixing time has

$$t_{\text{mix}}(\epsilon) \leq \left\lceil \frac{q\Delta}{q-2\Delta} n \log(n - \log \epsilon) \right\rceil.$$

*Proof.* We consider the graph of all colourings and extend Glauber dynamics to this graph. Given colourings  $x, y$ , we use the metric  $\rho(x, y) = \sum_{v \in V} \mathbf{1}(\{x(v) \neq y(v)\})$ . Two colourings are deemed neighbours if and only if  $\rho(x, y) = 1$ , i.e: they differ in exactly one vertex. Note that this neigh-

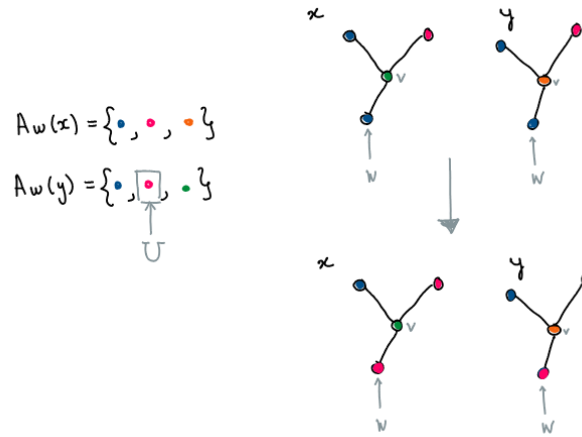
bouring rule defines a graph different from the graph defined by the transition of the chain, but Theorem 13.15 did not require the chain to be irreducible, and since we start on the set of proper colourings, we will remain in the set of proper colourings  $S$ .

We wish to define a coupling  $(X_1, Y_1)$  for when  $X_0 = x, Y_0 = y$  are two proper colourings with  $\rho(x, y) = 1$ , and show that  $\mathbb{E}_{xy}[\rho(X_1, Y_1)] \leq e^{-\alpha} \rho(x, y) = e^{-\alpha}$ , for some  $\alpha$ , then we will be able to apply Theorem 13.15. We describe how the coupling evolves:

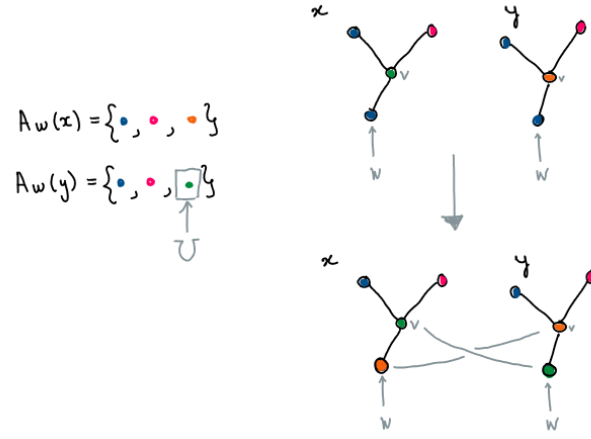
Let  $v$  be the vertex where  $x$  and  $y$  differ. We pick, uniformly at random, the same vertex  $w$  in both configurations. If  $w$  is not a neighbour of  $v$ , then we can update both chains to the same colour, and this would be a correct coupling since  $A_w(x) = A_w(y)$ , and the distance would remain at 1. (Recall that  $A_w(f)$  is the allowed colours that  $f$  can take at vertex  $w$ ).

Suppose on the other hand, that  $v$  is a neighbour of  $w$ . Suppose without loss of generality that  $|A_w(x)| \leq |A_w(y)|$ . We pick a colour  $U$  uniformly at random from  $A_w(y)$ .

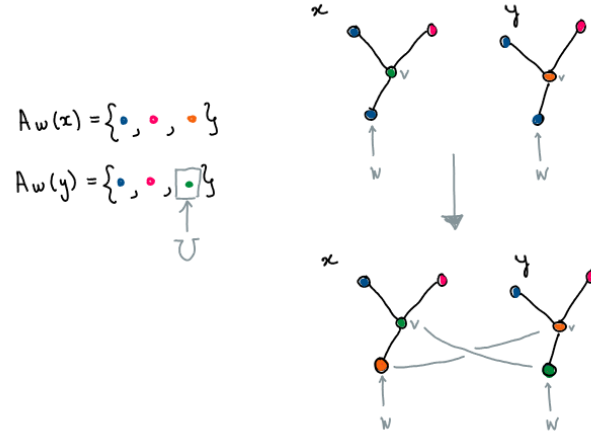
- If  $U \neq x(v)$ : then we update both  $x(w)$  and  $y(w)$  to  $U$ , and see that the distance remains at 1:



- If  $U = x(v)$ , then we distinguish two cases, in both of which the distance will increase to 2.
  1. If  $|A_w(x)| = |A_w(y)|$ , then set  $x(w) = y(v)$ , and  $y(w) = U$ :



Item If  $|A_w(x)| < |A_w(y)|$ , then set  $x(w)$  at random from  $A_w(x)$ , and  $y(w) = U$ .



Finally we note that if  $w = v$ , then we can update both to be the same, and the distance will fall to zero. Now notice that the only cases where  $x$  and  $y$  differ at  $w$  is when the colour  $U$  that was chosen uniformly at random from  $A_w(y)$  was equal to  $w(v)$ . This happens with probability  $1/|A_w(y)|$ , and since  $A_w(y) = q - \deg(w) \geq q - \Delta$ , we have that  $1/|A_w(y)| \leq 1/(q - \Delta)$

Now we notice that the distance  $\rho(x, y)$  increases to 2 when we pick a neighbour of  $v$ , this happens with probability  $\deg(v)/n$ , and we update it differently in both configurations, which happens with probability  $1/|A_w(y)|$ . The distance goes to zero if we pick  $v$ , which happens with probability  $1/n$ . Therefore the probability that we remain at distance 1, is going to be 1 minus the probability that it goes to zero minus the probability that it goes to two, i.e:  $1 - \frac{1}{n} - \frac{\deg(v)}{n} \frac{1}{|A_w(y)|}$ . Putting this all together we have that

$$\mathbf{E}_{x,y}[\rho(X_1, Y_1)] \leq 1 - \frac{1}{n} + \frac{\deg(v)}{n} \frac{1}{|A_w(y)|}$$

Now we use the bound that  $|A_w(y)| \geq q - \Delta$ , and  $\deg(v) \leq \Delta$ , to get that

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

where  $\alpha = \frac{q-2\Delta}{q-\Delta}$ , which is in  $(0, 1)$  by assumption of  $q > 2\Delta$ . This establishes the claim.



**Example 13.19 (Exclusion process)** Consider the following Interacting Particle System: on the complete graph  $K_{2n}$ , there are  $n$  white and  $n$  black particles. At each time, an edge is chosen at random and the particles swap. The mixing time is of order  $n \log n$ .

*Proof.* Maybe one day when I have time. For now find on written notes.





# Chapter 14

## Hit-Mix

In this section we will use hitting times to bound mixing times. Recall the definition that the hitting time for a set  $A$  is given by

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

**Definition 14.1 (Hit time)** Let  $P$  be a finite irreducible chain with invariant distribution  $\pi$ . For any  $\alpha, \epsilon \in (0, 1)$  and  $t \geq 0$ , we define the maximal probability of not hitting a set of size at least  $\alpha$  by time  $t$  starting from  $x$  as:

$$p_x(\alpha, t) = \max\{\mathbf{P}_x(T_A > t) : A \subseteq S, \pi(A) \geq \alpha\}$$

Then we also set  $p(\alpha, t) = \max_x p_x(\alpha, t)$ , and define the  $\epsilon$ -hit $_\alpha$  time starting from  $x$  and from the worst starting point as

$$\text{hit}_{\alpha,x}(\epsilon) = \min\{t : p_x(\alpha, t) \leq \epsilon\} \quad \text{hit}_\alpha(\epsilon) = \min\{t : p(\alpha, t) \leq \epsilon\}$$

In plain English:  $\text{hit}_\alpha(\epsilon)$  is the least time, for which the chain has hit all sets of size at least  $\alpha$  with probability at least  $1 - \epsilon$ . Alternatively, it is the least time for which the probability of not having hit some set of size at least  $\alpha$  is at most  $\epsilon$ . The intuition now is clear: if the chain hasn't hit some large set by time  $t$ , it can't be well mixed.

**Proposition 14.2 (Lower bound on mixing time by hit time)** For any chain, we have that for any  $\epsilon \in (0, 1)$ ,  $\delta \in (0, 1 - \epsilon)$ :

$$t_{\text{mix}}(\epsilon) \geq \text{hit}_{\epsilon+\delta}(1 - \delta)$$

**Main idea:** The intuition for the connection with mixing times is the following: if the chain is  $\epsilon$ -mixed by time  $t$ , then starting from any  $x$ ,  $P^t(x, \cdot)$  and  $\pi(\cdot)$  are very close in the TV sense, so if  $A$  is a set with

$\pi(A) \gg \epsilon$ , then it can't be that  $P^t(x, A) \ll 1$ . Indeed: this is because  $P^t(x, A) \geq \pi(A) - \epsilon$  by definition of TV.

*Proof.* Let  $t = t_{\text{mix}}(\epsilon)$ . Then by definition of TV distance, we have that for any set  $A$  with  $\pi$ -measure at least  $\delta + \epsilon$ , and for any  $x \in \Omega$ , we have that  $\pi(A) - P^t(x, A) \leq \epsilon$ , i.e:

$$P^t(x, A) \geq \pi(A) - \epsilon = \delta$$

But also,  $\mathbf{P}_x(T_A \leq t) \geq P^t(x, A) \geq \delta$ , so we have that

$$\mathbf{P}_x(T_A > t) < 1 - \delta$$

and since this holds for any  $x$ , it follows that  $\text{hit}_{\epsilon+\delta}(1-\delta) \leq t$ . ♡

## 14.1 Properties of hit-time

Now we have some result that will be useful in computations:

**Proposition 14.3 (Submultiplicativity of hit-times)** For any  $\alpha, \epsilon, \delta \in (0, 1)$  we have that

$$\text{hit}_\alpha(\epsilon\delta) \leq \text{hit}_\alpha(\epsilon) + \text{hit}_\alpha(\delta)$$

*Proof.* Let  $A \subseteq S$  be of  $\pi$ -measure at least  $\alpha$ . Let  $x \in \Omega$ , and  $t, s \geq 0$ . Then if starting from  $x$ ,  $A$  hasn't been hit by time  $t + s$ , then it definitely hasn't been hit by time  $t$ , and wherever the chain was at time  $t$ , say  $z$ , using the Markov Property, the chain started from  $z$  did not hit  $A$  by time  $s$ , i.e:

$$\mathbf{P}_x(T_A > t + s) \leq \mathbf{P}_x(T_A > t) \max_{z \in \Omega} \mathbf{P}_z(T_A > s) \leq p(\alpha, t)p(\alpha, s)$$

Therefore  $p(\alpha, t + s) \leq p(\alpha, t)p(\alpha, s)$ . Now take  $t = \text{hit}_\alpha(\epsilon)$  and  $s = \text{hit}_\alpha(\delta)$ , then by definition  $p(\alpha, t) < \epsilon$ , and  $p(\alpha, s) < \delta$ , which means that  $p(\alpha, t + s) < \epsilon\delta$  which means that  $\text{hit}_\alpha(\epsilon\delta) \leq t + s = \text{hit}_\alpha(\epsilon) + \text{hit}_\alpha(\delta)$ . ♡

**Proposition 14.4 (Hit-time is decreasing in  $\alpha$ )** Let  $\delta \in (0, 1)$  be given, any  $0 < \alpha \leq \beta \leq 1$ . Then

$$\text{hit}_\beta(\delta) \leq \text{hit}_\alpha(\delta)$$

**Main idea:** What its saying is that waiting to be  $1 - \delta$  sure that you have visited all small sets, is going to take much longer than waiting to be  $1 - \delta$  sure that you have visited all large sets. The proof is once again a formal restatement of this.

*Proof.* It is easy to see that

$$\max_{\pi(A) \geq \beta} \{\dots\} \leq \max_{\pi(A) \geq \alpha} \{\dots\}$$

(There are less sets of size at least  $\beta$  than there are sets of size at least  $\alpha$ ), so this means that  $p(\beta, t) \leq p(\alpha, t)$ , and so when  $p(\alpha, t)$  drops below  $\delta$ , so does  $p(\beta, t)$ .  $\heartsuit$

Above we have shown the rather trivial fact that hitting times  $\text{hit}_\alpha(\epsilon)$  are decreasing in  $\alpha$ , i.e: if  $\alpha \leq \beta$ , we can lower bound  $\text{hit}_\alpha$  with  $\text{hit}_\beta$ . It turns out, that in the case of reversible chains, even if we are looking at smaller sets (i.e.  $\alpha \leq \beta$ ), we can still upper bound  $\text{hit}_\alpha$  by only looking at the larger sets with  $\pi(A) \geq \beta$ , but we must pay a price: this price will be that you have to be even more certain that you have hit all sets of size  $\beta$ , as well as paying an extra factor of  $t_{\text{rel}}$ . These two factors of course increase waiting time and this is what gives the upper bound. We now state the result formally.

**Proposition 14.5 (Hit time bound with smaller sets)** For a reversible chain with relaxation time  $t_{\text{rel}}$ , and any  $0 < \epsilon < \delta < 1$ , and  $0 < \alpha \leq \beta < 1$ , we have the following comparison:

$$\text{hit}_\alpha(\delta) \leq \underbrace{\text{hit}_\beta(\delta - \epsilon)}_{\text{more certain}} + \underbrace{\left[ \alpha^{-1} t_{\text{rel}} \log \left( \frac{1 - \alpha}{(1 - \beta)\epsilon} \right) \right]}_{\text{extra price of } t_{\text{rel}}}$$

To prove this we first need the following Lemma:

**Lemma 14.6** For a reversible transition matrix  $P$  with invariant distribution  $\pi$  and relaxation time  $t_{\text{rel}}$ , we have that for any non-empty  $A$  and all times  $t \geq 0$ :

$$\mathbf{P}_\pi(T_A > t) \leq \pi(A^c) \exp \left( -\frac{t \pi(A)}{t_{\text{rel}}} \right) \quad (\star)$$

In particular, for any  $c, w > 0$ , let  $B = B(A, w, c) = \{y \in \Omega : \mathbf{P}_y(T_A > \lceil \frac{t_{\text{rel}} w}{\pi(A)} \rceil) \geq c\}$ . Then:

$$\pi(B) \leq \frac{\pi(A^c)}{e^{wc}} \quad (\star\star) \quad \text{and} \quad \pi(A) \mathbf{E}_\pi[T_A] \leq t_{\text{rel}} \pi(A^c) \quad (\star\star\star)$$

*Proof.* We first recall the result from ES2 Q10 that states that if  $B = A^c$  is a connected set, then

$$\mathbf{P}_{\pi_B}(T_A > t) \leq \left( 1 - \frac{\pi(A)}{t_{\text{rel}}} \right)^t$$

Where for a set  $X$ ,  $\pi_X$  is  $\pi$  conditioned on being in  $X$ , i.e:  $\pi_X(x) = \frac{\pi(x) \mathbf{1}(X)}{\pi(X)}$ . This pretty much

already gives  $(\star)$  but we need to fix the case when  $A^c$  is not connected. To do this, we express

$$A^c = \bigcup_{j=1}^k C_j$$

where each  $C_j$  is one of the connected components of  $A^c$ . Then since each  $C_i$  is connected, it is obvious that  $(C_i^c)^c$  is connected, so we may apply the Example Sheet result and see that

$$\begin{aligned} \mathbf{P}_\pi(T_A > t) &= \underbrace{\mathbf{P}_{\pi_A}(T_A > t)}_{=0} \pi_A(A) + \sum_{i=1}^k \pi(C_i) \mathbf{P}_{\pi_{C_i}}(T_A > t) \\ &= \sum_{i=1}^k \pi(C_i) \mathbf{P}_{\pi_{C_i}}(T_{C_i^c} > t) \end{aligned}$$

Where step  $(\dagger)$  comes from the fact that since  $\geq 0$ , the probability that starting from  $A$ , not having hit  $A$  by a time strictly greater than 0 clearly has probability zero. The last equality comes from the fact that if you start in  $C_i$ , one of the connected components of  $A^c$ , not having hit  $A$  by time  $t$  is equivalent to not having exited  $C_i$  by time  $t$ . So now using the Example Sheet question gives:

$$\begin{aligned} \mathbf{P}_\pi(T_A > t) &\leq \sum_{i=1}^k \pi(C_i) \left(1 - \frac{\pi(C_i^c)}{t_{\text{rel}}}\right)^t \\ &\leq \sum_{i=1}^k \pi(C_i) \left(1 - \frac{\pi(A)}{t_{\text{rel}}}\right)^t && (\text{Since } \pi(C_i^c) \geq \pi(A)) \\ &= \pi(A^c) \left(1 - \frac{\pi(A)}{t_{\text{rel}}}\right)^t \\ &\leq \pi(A^c) \exp\left(-\frac{t \pi(A)}{t_{\text{rel}}}\right). \end{aligned}$$

This shows  $(\star)$ . To show the further implications, we note that when  $t = t(A, w) = \left\lceil \frac{t_{\text{rel}} w}{\pi(A)} \right\rceil$ ,

$$\begin{aligned} \pi(B)^c &\leq \pi(B) \mathbf{P}_{\pi_B}(T_A > t) && (\text{definition of } B) \\ &\leq \mathbf{P}_\pi(T_A > t) && (\text{Since } A = B \cup (A \setminus B)) \\ &\leq \pi(A^c) \exp\left(-\frac{t \pi(A)}{t_{\text{rel}}}\right) && (\text{by } (\star)) \\ &\leq \pi(A^c) e^{-w} \end{aligned}$$

Where this last inequality was obtained by the fact that  $\left\lceil \frac{t_{\text{rel}} w}{\pi(A)} \right\rceil \geq \frac{t_{\text{rel}} w}{\pi(A)}$ . This proves the first

inequality of the "moreover" part. For the second inequality we simply note that

$$\begin{aligned} \mathbf{E}_\pi[T_A] &= \sum_{t=0}^{\infty} \mathbf{P}_\pi(T_A > t) \\ &\leq \sum_{t=0}^{\infty} \pi(A^c) \left(1 - \frac{\pi(A)}{t_{\text{rel}}}\right)^t \\ &\leq \pi(A^c) \frac{t_{\text{rel}}}{\pi(A)} \end{aligned}$$

Where the last inequality comes from the geometric series. ♡

We can now prove Proposition 14.5.

*Proof of Proposition 14.5.* Take  $A$  to be with  $\pi(A) \geq \alpha$ . Define:

$$\hat{A} = \{x \in \Omega : \mathbf{P}_x(T_A > s) \leq \epsilon\}$$

Where  $s = \lceil \frac{t_{\text{rel}} w}{\pi(A)} \rceil$ . Quite naturally,  $A \subseteq \hat{A}$ , because if you start at  $A$ , the probability that you haven't hit  $A$  by time  $s$  is zero which is less than or equal to  $\epsilon$ . Note from the bound

$$\pi(B) \leq \frac{\pi(A^c)}{e^{w c}}$$

that when choosing  $B = \hat{A}^c$ , we can take complements and note that

$$\pi(\hat{A}) \geq 1 - \frac{\pi(A^c)}{e^{w c}}$$

and by making  $s$  large enough (i.e:  $w$  large enough), we can get this as close as we want to 1. Therefore there exists some  $s$  such that  $\pi(\hat{A}) \geq \beta$ . Moreover, by definition of  $\hat{A}$ , whenever we start from  $\hat{A}$ , we reach  $A$  by time  $s$  with probability  $1 - \epsilon$ . With this choice of  $s$ , pick  $t = \text{hit}_\beta(\delta - \epsilon)$ , and we see that

$$\begin{aligned} \max_{x \in \Omega} \mathbf{P}_x(T_A > s + t) &\leq \max_{x \in \Omega} \mathbf{P}_x \left( T_{\hat{A}} > t \cup T_A^{X_{T_{\hat{A}}}} > s \right) \\ &\leq \max_{x \in \Omega} \mathbf{P}_x(T_{\hat{A}} > t) + \max_{x \in \hat{A}} \mathbf{P}_x(T_A > s) \\ &\leq (\delta - \epsilon) + \epsilon = \delta \end{aligned}$$

Where in the first inequality we have used the fact that if  $A$  is not reached by time  $s + t$  then either  $\hat{A}$  hasn't been reached by time  $t$  or the chain started from wherever it was at time  $T_{\hat{A}}$  takes more than  $s$  to hit  $A$ . For the second inequality we have used a union bound and the Strong Markov Property, and for the third inequality, we have used the fact that the definition of  $t = \text{hit}_\beta(\delta - \epsilon)$  implies that for any set of size at least  $\beta$  (like  $\hat{A}$ ), the probability that by time  $t$  we haven't reached

the set is at most  $\delta - \epsilon$ . For the second summand, we have used the observation that when started from  $\hat{A}$ , the chain hits  $A$  in time at most  $s$  with probability  $1 - \epsilon$ . This immediately implies that, since  $\pi(A) \geq \alpha$  by assumption, then  $\text{hit}_\alpha(\delta) \leq t + s = \text{hit}_\beta(\delta - \epsilon) + s$ . We are almost done, now we just need to bound  $s$ . In fact, we can just show that taking  $s = \left\lceil \alpha^{-1} t_{\text{rel}} \log\left(\frac{1-\alpha}{(1-\beta)\epsilon}\right) \right\rceil$ , we will still have  $\pi(\hat{A}) \geq \beta$ , and then the whole proof above will still work and we will get the desired upper bound. This is just a slightly tedious computation in the inequality

$$\pi(\hat{A}) \geq 1 - \frac{\pi(A^c)}{e^{w c}}$$

where  $w = \frac{s\pi(A)}{t_{\text{rel}}}$ .



Here's a quick summary of what we've seen so far:

- If we have an irreducible Markov chain  $X$  on a state space  $\Omega$ , we can talk about the  $\text{hit}_\alpha(\epsilon)$  time, formally defined as

$$\text{hit}_\alpha(\epsilon) = \min \left\{ t : \max_x \{ \mathbf{P}_x(T_A > t) : \pi(A) \geq \alpha \} < \epsilon \right\}.$$

In plain English: the first time  $t$  at which we are  $1-\epsilon$  certain that we have hit any set of size at least  $\alpha$  by time  $t$  when starting from the worst possible point.

- Intuitively it is clear that there should be a connection between mixing and hitting times: if  $t < t_{\text{hit}_{\epsilon+\delta}}(1-\delta)$ , then there is some starting point  $x$  and some set of mass at least  $\epsilon+\delta$  such that the probability of having it is by time  $t$  from  $x$  is at most  $\delta$ . Combining these two things it easily follows that the distance to stationarity at time  $t$  is at least  $\epsilon$ , or in other words:  $t_{\text{mix}}(\epsilon) \geq t_{\text{hit}_{\epsilon+\delta}}(1-\delta)$ .
- We also have some properties of hit times, some more immediate than others:

1. The hit times are monotone decreasing in size: this is not a surprise, since increasing the size reduces the set over which the maximum is being taken. Formally: if  $\alpha \leq \beta$ , then

$$t_{\text{hit}_\alpha}(\epsilon) \geq t_{\text{hit}_\beta}(\epsilon).$$

2. An easy application of the Markov Property is that if we haven't hit a set  $A$  by time  $t+s$ , then in particular, we haven't hit it by time  $t$ , and then the chain started afresh hasn't hit  $A$  by time  $s$ , i.e:  $\max_x \mathbf{P}_x(T_A > t+s) \leq (\max_x \mathbf{P}_x(T_A > t))(\max_z \mathbf{P}_z(T_A > s))$ . This immediately gives that

$$t_{\text{hit}_\alpha}(\epsilon\delta) \leq t_{\text{hit}_\alpha}(\epsilon) + t_{\text{hit}_\alpha}(\delta).$$

3. A more technical argument can also be used to show that in contrast to bullet point 1. we can also have an "increasing inequality" on the sizes of the sets, but only if we reduce our certainty on the hitting probability and pay a price of  $t_{\text{rel}}$ :

$$t_{\text{hit}_\alpha}(\delta) \leq t_{\text{hit}_\delta}(\delta - \epsilon) + C(\alpha, \beta, \epsilon) t_{\text{rel}}.$$

## 14.2 An upper bound on mixing time

In the previous section we saw a lower bound for the mixing time in terms of the hitting times. In this section we present a rather long proof that gives an upper bound. In particular, we will see that when  $P$  is a reversible irreducible transition matrix, then there is some  $C(\alpha, \epsilon, \delta)$  such that:

$$t_{\text{mix}}((\epsilon + \delta) \wedge 1) \leq \text{hit}_{1-\alpha}(\epsilon) + C(\alpha, \epsilon, \delta) t_{\text{rel}},$$

and in particular, when  $\alpha \downarrow 0$ , then  $C(\alpha, \epsilon, \delta) \uparrow$ , and vice-versa: when  $\alpha \uparrow 1$ , then  $C(\alpha, \epsilon, \delta) \downarrow$ .

Let us start by stating two results:

**Lemma 14.7 (Generalised Poincaré Inequality)** For a reversible transition matrix  $P$  with invariant distribution  $\pi$  and any  $f : S \rightarrow \mathbf{R}$  we have that for all  $t \geq 0$ :

$$\text{Var}_{\pi}(P^t f) \leq e^{-2t/t_{\text{rel}}} \text{Var}_{\pi}(f).$$

*Proof.* Example Sheet 2 Question 5 (a).



**Lemma 14.8 (Starr's maximal inequality)** For a reversible irreducible and finite transition matrix  $P$  with invariant distribution  $\pi$ ,  $p \in (1, \infty)$  and  $p^*$  its conjugate exponent, we have that for all  $f : \Omega \rightarrow \mathbf{R}$ ,

$$\|f^*\|_p \leq p^* \|f\|_+$$

Where  $f^* : \Omega \rightarrow \mathbf{R}$  is the maximal function at even times, i.e:

$$f^*(x) = \sup_{t \geq 0} |(P^{2t} f)(x)| = \sup_{t \geq 0} |\mathbf{E}_x[f(X_{2t})]|$$

The key to this proof are the so-called Good sets. In plain English, a good set for  $A \subseteq \Omega$  is a set that satisfies the property that when the chain is started from the Good set, the chain will hit  $A$  by some time  $s$  with probability close to  $\pi(A)$ . If we can show that the Good sets are large, e.g: say that the good set for  $A$  has  $\pi$ -mass  $1 - \delta$ , then by time  $t = t_{\text{hit}1-\delta} + s$  we will be in  $A$  with probability close to  $\pi(A)$  plus some small error, from which we will see that  $|P^t(x, A) - \pi(A)|$  will be small.

**Definition 14.9 (Good set)** Let  $A \subseteq \Omega$  be a set. Then for a fixed  $s > 0$  define  $\sigma_s = \exp\left(-\frac{s}{t_{\text{rel}}}\right) \sqrt{\text{Var}_{\pi}(\mathbf{1}_A)}$ .



Then for  $m > 0$ , we define the Good set for  $A$ :

$$G_s(A, m) = \{y \in \Omega : |P^t(y, A) - \pi(A)| < m\sigma_s \text{ for } t \geq s\}$$

We will now state the proposition which will be key in establishing the hit-mix bound.

**Proposition 14.10 (Good sets are large)** Let  $P$  be irreducible and reversible. Then for all  $A \subseteq \Omega$ ,  $s \geq 0$  and  $m > 0$ , we have that

$$\pi(G_s(A, m)) \geq 1 - 8m^{-2}$$

*Proof.* Start by defining  $f_s(x) = P^s(x, A) - \pi(A) = P^s(\mathbf{1}(A) - \pi(A))(x)$ . Then using the notation of Starr's inequality:

$$f_s^*(x) = \sup_{t \geq 0} |P^{2t} f_s(x)| = \sup_{t \geq 0} |P^{2t+s}(x, A) - \pi(A)|$$

and

$$(P f_s)^*(x) = \sup_{t \geq 0} |P^{2t+1+s}(x, A) - \pi(A)|.$$

And so using this notation, we see that

$$G_s(A, m) = \{y : f_s^*(y) < m\sigma_s \text{ and } (P f_s)^*(x) < m\sigma_s\}$$

this is because taking the supremum of  $t \geq 0$ , will cover all even times after  $s$  (this will be done by  $f_s^*$ ) as well as all odd times (done by  $(P f_s)^*$ ), and so by a union bound:

$$1 - \pi(G_s(A, m)) \leq \pi(\{f_s^*(x) \geq m\sigma_s\}) + \pi(\{(P f_s)^*(x) \geq m\sigma_s\})$$

This gives us an upper bound on the  $\pi$ -measure of  $G_s(A, m)$ . All left to do is understand the  $\pi$ -measure of the sets on the right hand side of the inequality, and here is where we will use the Generalised Poincaré Inequality (GPI) as well as Starr's inequality. First we need to note that

$$\mathbf{E}_\pi[f_s] = \sum_x \pi(x) \{P^s(x, A) - \pi(A)\} = (\pi P^s)(A) - \pi(A) = 0$$

which also means by standard arguments that  $\mathbf{E}_\pi[P f_s] = 0$ . This allows us to express the variances of both  $f_s$  and  $P f_s$  directly as  $\ell^2$  norms, and so we have the following:

$$\begin{aligned} \|P f_s\|_2^2 &\stackrel{(1)}{\leq} \|f_s\|_2^2 \\ &\stackrel{(2)}{=} \text{Var}_\pi(P^s f_0) \\ &\stackrel{(3)}{\leq} e^{-2s/t_{\text{rel}}} \text{Var}_\pi(f_0) \\ &\stackrel{(4)}{=} e^{-2s/t_{\text{rel}}} \pi(A)(1 - \pi(A)) =: \sigma_s^2 \end{aligned}$$

Where (1) comes from GPI, (2) comes from the definition  $f_0 := \mathbf{1}(A) - \pi(A)$  as well as the fact that  $\mathbf{E}_\pi[f_0]$  is trivially checked to be zero, (3) comes again from GPI, (4) comes from direct computation of  $\mathbf{E}_\pi[f_0^2] = \mathbf{E}_\pi[\mathbf{1}(A) - 2\mathbf{1}(A)\pi(A) + \pi(A)]$ . Hence we compute the  $\pi$ -measures of the sets we need:

$$\begin{aligned} \pi(\{f_s^*(x) \geq m\sigma_s\}) &\stackrel{(1)}{=} \pi(\{(f_s^*(x))^2 \geq m^2\sigma_s^2\}) \\ &\stackrel{(2)}{\leq} \frac{\mathbf{E}_\pi[(f_s^*)^2]}{m^2\sigma_s^2} \\ &\stackrel{(3)}{=} \frac{\|f_s^*\|_2^2}{m^2\sigma_s^2} \\ &\stackrel{(4)}{\leq} \frac{(2\|f_s\|_2^2)}{m^2\sigma_s^2} \\ &\stackrel{(5)}{\leq} \frac{4}{m^2} \end{aligned}$$

Where (1) comes from  $f_s^*$  being non-negative, (2) comes from Markov's inequality, (3) comes from definition of  $\ell^2$  norm, (4) comes from Starr's inequality using the fact that the conjugate exponent of 2 is 2, and (5) comes from the bound we obtained above. We could repeat the argument above for  $\pi(\{(Pf_s)^*(x) \geq m\sigma_s\})$  and use the fact that above we have shown (in step (1) of the previous argument) that  $\|Pf_s\|_2^2 \leq \|f_s\|_2^2$ , and so we can insert this in step (3) of this argument. This shows that  $\pi(\{(Pf_s)^*(x) \geq m\sigma_s\}) \leq 4/m^2$  and so the proof follows.  $\heartsuit$

We are finally ready to prove the hit-mix bound:

**Theorem 14.11** Let  $P$  be a reversible irreducible transition matrix, then there is some  $C(\alpha, \epsilon, \delta)$  such that:

$$t_{\text{mix}}((\epsilon + \delta) \wedge 1) \leq \text{hit}_{1-\alpha}(\epsilon) + C(\alpha, \epsilon, \delta) t_{\text{rel}}$$

*Proof.* First, fix the value of  $x$  (it exists because everything's finite), such that  $d(t+s) = \|P^{t+s}(x, \cdot) - \pi(\cdot)\|_{\text{TV}}$ . Then the goal is going to be to find the values of  $t, s$  such that  $\pi(A) - P^{t+s}(x, A) \leq \epsilon + \delta$ . We will do so by considering a good set for  $A$ , given by

$$G = G_s(A, \sqrt{8/\alpha})$$

From Proposition Good sets are large, we have that  $\pi(G) \geq 1 - \alpha$ , and we will now proceed to lower bound  $P^{t+s}(x, A)$ , as this will produce the desired upper bound. We will do this by conditioning on having hit the good set by time  $t$ . Naturally:

$$P^{t+s}(x, A) \geq \mathbf{P}_x(T_G \leq t) \mathbf{P}_x(X_{t+s} \in A \mid T_G \leq t)$$

We now make our choice of  $t$ . If we set  $t = \text{hit}_{1-\alpha}(\epsilon)$ , then since  $G$  has  $\pi$ -measure  $1 - \alpha$ , it follows

that  $\mathbf{P}_x(T_G \leq t) \geq 1 - \epsilon$ . Now we continue by lower bounding  $\mathbf{P}_x(X_{t+s} \in A \mid T_G \leq t)$ . This will be the heart of the proof:

$$\begin{aligned}
 \mathbf{P}_x(X_{t+s} \in A \mid T_G \leq t) &\stackrel{(1)}{=} \sum_{u=0}^t \sum_{y \in G} \mathbf{P}_x(X_{t+s} \in A, T_G = u, X_u = y \mid T_G \leq t) \\
 &\stackrel{(2)}{=} \sum_{u=0}^t \sum_{y \in G} \mathbf{P}_x(X_{t+s} \in A \mid T_G = u, X_u = y, T_G \leq t) \mathbf{P}_x(T_G = u, X_u = y \mid T_G \leq t) \\
 &\stackrel{(3)}{=} \sum_{u=0}^t \sum_{y \in G} \mathbf{P}_y(X_{t+s-u} \in A) \mathbf{P}_x(T_G = u, X_u = y \mid T_G \leq t) \\
 &\stackrel{(4)}{\geq} \sum_{u=0}^t \sum_{y \in G} \left( \pi(A) - \sqrt{\frac{8}{\alpha} \sigma_s} \right) \mathbf{P}_x(T_G = u, X_u = y \mid T_G \leq t) \\
 &\stackrel{(5)}{=} \pi(A) - \sqrt{\frac{8}{\alpha} \sigma_s}
 \end{aligned}$$

Where (1) comes from the Law of Total Probability, (2) comes from definition of conditional probability, (3) comes from the Markov Property, more precisely, if  $u$  is the present, given the present, we can shift the Markov chain and forget of everything that has happened up until now, which includes the event  $T_G = u$  (which is an event determined by the present). Then (4) comes from the fact that  $y \in G$ , and since  $G$  is a good set, the probability that we hit  $A$  at some time later than or equal to  $s$  (note that  $t - u \geq 0$ ) is bonded below by the measure of  $A$  minus  $m\sigma_s$ , and in the definition of  $G$  we chose  $m$  to be that square root. Step (5) comes from pulling out the constant factor out of the sum and noting that the sums account for all possible things so it will be equal to 1. Putting all this together we get that

$$P^{t+s}(x, A) \geq (1 - \epsilon) \left( \pi(A) - \sqrt{\frac{8}{\alpha} \sigma_s} \right)$$

Now we make the smart choice of  $s$ , in particular, for

$$s = \left\lceil \frac{1}{2} t_{\text{rel}} \left\{ \log \left( \frac{2(1-\epsilon)^2}{\alpha \epsilon \delta} \right) \vee 0 \right\} \right\rceil$$

we get that

$$\pi(A) - P^{t+s}(x, A) \leq \epsilon + \delta$$

there is no substance in the calculation above, hence why I skip it. This finishes showing that

$$t_{\text{mix}}(\{\epsilon + \delta\} \wedge 1) \leq \text{hit}_{1-\alpha}(\epsilon) + \left\lceil \frac{1}{2} t_{\text{rel}} \left\{ \log \left( \frac{2(1-\epsilon)^2}{\alpha \epsilon \delta} \right) \vee 0 \right\} \right\rceil$$

as required (note for completeness that we take the minimum with one because the maximum TV distance could be is 1). ♡

### 14.3 Hit cutoff

Recall from our discussion on cutoff that chains without the product condition, i.e: without  $t_{\text{rel}} \ll t_{\text{mix}}$ , one cannot have mixing time cutoff. Given that we have seen how hit times and mixing times are almost of the same order (up to paying an extra factor of  $t_{\text{rel}}$ ), it makes sense that there is an parallel discussion of cutoff that one can make with hit times.

**Definition 14.12 (Hit cutoff)** For  $\alpha \in (0, 1)$ , a sequence of chains exhibits  $\text{hit}_\alpha$  cutoff if for all  $\epsilon \in (0, 1/4)$ , we have that

$$\text{hit}_\alpha(\epsilon) - \text{hit}_\alpha(1 - \epsilon) \ll \text{hit}_\alpha(1/4)$$

If the chain exhibits  $\text{hit}_\alpha$  cutoff for all  $\alpha \in (0, 1)$  we say that the chain exhibits hit cutoff.

**Remark 14.13** In the definition above, of course we should have written  $\text{hit}_\alpha^{(n)}$  instead of  $\text{hit}_\alpha$  to indicate that this is a property of the sequence of chains as  $n \rightarrow \infty$ .

We now see that if a sequence of chains has the product condition, then  $\text{hit}_\alpha$  cutoff and hit cutoff are the same.

**Proposition 14.14 (hit $_\alpha$  cutoff gives hit cutoff on reversible product condition chains)** For a sequence of reversible Markov chains which satisfy the product condition, i.e:  $t_{\text{rel}} \ll t_{\text{mix}}$ , we have that the chain exhibits  $\text{hit}_\alpha$  cutoff for some  $\alpha$  if and only if it exhibits hit cutoff.

Moreover,  $\text{hit}_\alpha(1/4) \asymp t_{\text{mix}}$ . And if there is hit cutoff, then

$$\lim_{n \rightarrow \infty} \frac{\text{hit}_\alpha^{(n)}(1/4)}{\text{hit}_{1/2}^{(n)}(1/4)} = 1$$

for any  $\alpha \in (0, 1)$

*Proof.* Let us first show that  $\text{hit}_\alpha(1/4) \asymp t_{\text{mix}}$ . Fix  $\alpha \in (0, 1)$ . Before diving into a proof, we note that this is “morally obvious” because we have seen how to upper bound hit times by mixing times at the start of the chapter, and we later saw a way of lower bounding hit times by mixing times with an extra factor of  $t_{\text{rel}}$ . Since we are assuming the product condition holds, this factor of  $t_{\text{rel}}$

is negligible. We have the following trivial observation:

$$(1 - 3\alpha/4)^{4\alpha^{-1}} \leq e^{-3} \leq 1/4$$

Recall that if  $p < q$ , then  $\text{hit}_\alpha(p) \geq \text{hit}_\alpha(q)$ . So

$$\begin{aligned} \text{hit}_\alpha(1/4) &\leq \text{hit}_\alpha(\{1 - 3\alpha/4\}^{4\alpha^{-1}}) \\ &\stackrel{(2)}{\leq} 4\alpha^{-1} \text{hit}_\alpha(1 - 3\alpha/4) \\ &\stackrel{(3)}{\leq} 4\alpha^{-1} t_{\text{mix}}(\alpha/4) \\ &\stackrel{(4)}{\leq} C(\alpha) t_{\text{mix}} \end{aligned}$$

Where in (2) we have used the fact that hit times are submultiplicative, to lower the exponent as a sum of hit times, in (3) we have used the fact that  $\text{hit}_{\epsilon+\delta}(1-\delta) \leq t_{\text{mix}}(\epsilon)$ , and so writing  $\epsilon + \delta = \alpha$  and  $\delta = 3\alpha/4$  gives  $\epsilon = \alpha/4$ . For (4) we have used the fact that  $t_{\text{mix}}(\epsilon) \leq \lceil -\log_2(\epsilon) \rceil t_{\text{mix}}$ . To show the other bound we have that by the Hit-Mix bound (Theorem 14.11),

$$t_{\text{mix}}(1/4) \leq \text{hit}_\alpha(1/8) + C(1 - \alpha, 1/8, 1/8) t_{\text{rel}}$$

But we can bound  $\text{hit}_\alpha(1/8) \leq \text{hit}_\alpha(1/16) \leq 2 \text{hit}_\alpha(1/4)$ , and so we have that

$$t_{\text{mix}} \leq 2 \text{hit}_\alpha(1/4) + C t_{\text{rel}}$$

So by the product condition we have the desired result that  $\text{hit}_\alpha(1/4) \asymp t_{\text{mix}}$ . We now show how having  $\text{hit}_\alpha$  cutoff implies having cutoff for any  $\alpha$ . To do so fix  $0 < \alpha < \beta < 1$ , we shall show that  $\text{hit}_\alpha$  cutoff occurs if and only if  $\text{hit}_\beta$  cutoff occurs.

Recall that we know how to compare  $\text{hit}_\alpha$  and  $\text{hit}_\beta$  with the following:

$$\text{hit}_\alpha(\delta) \leq \text{hit}_\beta(\delta - \epsilon) + C(\alpha, \beta, \epsilon) t_{\text{rel}}$$

so if we fix  $\epsilon \in (0, 1/8)$ , we have that

- $\text{hit}_\alpha(1 - \epsilon) \leq \text{hit}_\beta(1 - 2\epsilon) + C(\alpha, \beta, \epsilon) t_{\text{rel}}$
- $\text{hit}_\alpha(2\epsilon) \leq \text{hit}_\beta(\epsilon) + C(\alpha, \beta, \epsilon) t_{\text{rel}}$

Moreover, since  $\text{hit}_\alpha$  is decreasing in  $\alpha$  (as its going to take less to have hit all large sets than it takes to hit small sets), we have that

- $\text{hit}_\beta(2\epsilon) \leq \text{hit}_\alpha(2\epsilon) \leq \text{hit}_\alpha(\epsilon)$ .

- $\text{hit}_\beta(1-\epsilon) \leq \text{hit}_\beta(1-2\epsilon) \leq \text{hit}_\alpha(1-2\epsilon)$ .

Putting this all together you have that

$$\text{hit}_\beta(2\epsilon) - \text{hit}_\beta(1-2\epsilon) \leq \text{hit}_\alpha(\epsilon) - \text{hit}_\alpha(1-\epsilon) - C(\alpha, \beta, \epsilon) t_{\text{rel}}.$$

and

$$\text{hit}_\alpha(2\epsilon) - \text{hit}_\alpha(1-2\epsilon) \leq \text{hit}_\beta(\epsilon) - \text{hit}_\beta(1-\epsilon) + C(\alpha, \beta, \epsilon) t_{\text{rel}}.$$

Now its easy to check that say, if there is  $\beta$  cutoff, then the right hand side of the second inequality is  $o(\text{hit}_\beta(1/4))$  but since by the first part we know that  $\text{hit}_\beta(1/4) \asymp \text{hit}_\alpha(1/4) \asymp t_{\text{mix}}$ , then the right hand side of the second inequality is  $o(\text{hit}_\alpha(1/4))$  which means there is  $\alpha$  cutoff. One can do the same argument to show the other direction. Now finally to show that

$$\lim_{n \rightarrow \infty} \frac{\text{hit}_\alpha^{(n)}(1/4)}{\text{hit}_{1/2}^{(n)}(1/4)} = 1$$

for any  $\alpha \in (0, 1)$ , we start by assuming  $\alpha < 1/2$ , then once again since we can compare  $\text{hit}_\alpha$  and  $\text{hit}_{1/2}$  times by paying a multiple of  $t_{\text{rel}}$ , we have that

$$\text{hit}_\alpha(1/4 + \epsilon) - C t_{\text{rel}} \leq \text{hit}_{1/2}(1/4) \leq \text{hit}_\alpha(1/4)$$

Now we observe the following: since  $\epsilon \leq 1/4$ , we automatically get that

$$\frac{1}{4} + \epsilon \in (\epsilon, 1 - \epsilon),$$

whence it follows that  $\text{hit}_\alpha(\frac{1}{4} + \epsilon) \in [\text{hit}_\alpha(1 - \epsilon), \text{hit}_\alpha(\epsilon)]$ . But in a similar spirit,  $\frac{1}{4} \in (\epsilon, 1 - \epsilon)$  as well, and so  $\text{hit}_\alpha(\frac{1}{4}) \in [\text{hit}_\alpha(1 - \epsilon), \text{hit}_\alpha(\epsilon)]$  too. Since we are assuming cutoff holds this window has order  $o(\text{hit}_\alpha(1/4))$ , which means that

$$|\text{hit}_\alpha(1/4) - \text{hit}_\alpha(1/4 + \epsilon)| \leq o(\text{hit}_\alpha(1/4)),$$

and in particular  $\text{hit}_\alpha(1/4 + \epsilon) \geq (1 - o(1))\text{hit}_\alpha(1/4)$ , thus showing the limit. The case for  $\alpha \geq 1/2$  is similar. ♥

With this result one could now show the following, which is the main result of this section. We skip the proof due to a massive lack of energy.

**Theorem 14.15** (Hit cutoff and cutoff are equivalent under the product condition in reversible chains) Consider a reversible chain with product condition. Then for any  $\alpha$ , the chain exhibits  $\text{hit}_\alpha$  cutoff if and only if it exhibits cutoff.

Let us summarise this section:

- We start by making an analogous definition of cutoff for hitting times: we say that a sequence of chains exhibits  $\text{hit}_\alpha$  cutoff if for any  $\epsilon \in (0, 1/2)$ :

$$t_{\text{hit}_\alpha}(\epsilon) - t_{\text{hit}_\alpha}(1 - \epsilon) \ll t_{\text{hit}_\alpha}(1/4).$$

In plain English, this just says that the window of time during which the chain goes from very unlikely to have hit  $\geq \alpha$  sets, until it is very likely to have hit them, has a width that grows negligibly when compared to the growth of the  $t_{\text{hit}_\alpha}(1/4)$  time.

- Motivated by the fact that we saw how to compare mixing and hit times up to a factor of  $t_{\text{rel}}$ , we set off to show that under the hypothesis of product condition ( $t_{\text{rel}} \ll t_{\text{mix}}$ ), we have an equivalence of cutoff and  $t_{\text{hit}}$  cutoff: doing so involved a few steps:

1. For any  $\alpha$ :  $\text{hit}_\alpha(1/4) \asymp t_{\text{mix}}(1/4)$ . This essentially comes from using the fact that we can compare  $\text{hit}_\alpha$  and  $t_{\text{mix}}$  up to relaxation time, and then use the product condition.
2. To show that there is  $\alpha$ -hit cutoff if and only if there is hit cutoff, we show that for say  $\alpha \leq \beta$  there is  $\text{hit}_\alpha$  cutoff if and only if there is  $\text{hit}_\beta$  cutoff. For this we simply compare  $\text{hit}_\alpha$  to  $\text{hit}_\beta$  using both the monotone decreasing property of size, and the comparison with an extra factor of  $t_{\text{rel}}$ .
3. Finally we have that if there is hit- $\alpha$  cutoff, then  $\text{hit}_\alpha(1/4)/\text{hit}_{\frac{1}{2}}(1/4) \rightarrow 1$ . This is done say by first assuming  $\alpha < 1/2$ , and sandwiching  $\text{hit}_{1/2}(1/4)$  on both sides. Monotonicity gives  $\text{hit}_{1/2}(1/4) \leq \text{hit}_\alpha(1/4)$ , and then with cutoff and comparing different sizes up to  $t_{\text{rel}}$ , we see that

$$(1 - o(1))\text{hit}_\alpha(1/4) \leq \text{hit}_{1/2}(1/4).$$

## 14.4 Trees

We have discussed a few times already how the product condition, i.e:  $t_{\text{rel}} \ll t_{\text{mix}}$  is indeed a **necessary** condition for mixing time cutoff to hold. We haven't yet shown whether the converse holds. In general it turns out not to, but in the specific case of random walks on trees, we can actually prove that product condition is equivalent to cutoff. The key to this will be the result of the previous section, where we showed that given product condition holds, hit cutoff is equivalent to mixing time cutoff.

**Definition 14.16** (Markov chain on trees) Recall that a Markov chain transition matrix  $P$  on a tree  $T=(V, E)$  (i.e: a graph with no cycles), is reversible with respect to its invariant measure. A vertex  $v \in V$  is called a central vertex if each connected component of  $T \setminus \{v\}$ , has  $\pi$ -measure at most  $1/2$ . Fix a central vertex  $o$  of  $T$  (this 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  $\ell(x) = (x_0 = x, x_1, \dots, x_k = o)$  the (unique) path from  $x$  to  $o$ . We call  $x_1$  a parent of vertex  $x$  and label it  $p_x = x_1$ . We also define a subtree from  $u$ ,  $\mathcal{T}_u$  to be  $\mathcal{T}_u = \{v \in T : u \in \ell(v)\}$ . Finally we define the  $\epsilon$  hitting time of  $o$  as

$$\tau_o(\epsilon) = \min\{t : \mathbf{P}_x(T_o > t) \leq \epsilon, \forall x \in V\}$$

In plain words:  $\tau_o(\epsilon)$  is the least time you have to wait so that the probability of hitting the root starting from anywhere is no less than  $1 - \epsilon$ .

**Remark 14.17** The fact that for every tree there is at least one and at most two central vertices is proven in ES4.

**Lemma 14.18** (Comparison of  $\tau_o$  and  $\text{hit}_{1/2}$ ) Let  $s_\delta = \lceil 4t_{\text{rel}} \log(4\delta/9) \rceil$ . Then for all  $0 < \delta < \epsilon < 1$ ,

$$\tau_o(\epsilon) \leq \text{hit}_{1/2}(\epsilon) \leq \tau_o(\epsilon - \delta) + s_\delta$$

*Proof.* Fix any  $x \in V$ . Let  $A_x$  be the union of  $\{o\}$  and all the connected components of  $T \setminus \{o\}$  that do not contain  $x$ . Then since the connected component that contains  $x$  has  $\pi$ -measure less than or equal to  $1/2$ , it must be that  $\pi(A_x) \geq 1/2$ . Moreover, since  $T$  is a tree, hitting  $A_x$  in time less than or equal to  $t$  is equivalent to hitting  $o$  with time less than or equal to  $t$ , and as such



$\tau_o(\epsilon) \leq \text{hit}_{1/2}(\epsilon)$ . Indeed,

$$\begin{aligned}\tau_o(\epsilon) &= \min\{t : \mathbf{P}_x(T_o > t) \leq \epsilon, \forall x\} \\ &= \min\{t : \mathbf{P}_x(T_{A_x} > t) \leq \epsilon, \forall x\} \\ &\leq \min\left\{t : \max_{\pi(A) \leq 1/2} \mathbf{P}_x(T_A > t) \leq \epsilon, \forall x\right\} \\ &= \text{hit}_{1/2}(\epsilon).\end{aligned}$$

This establishes the first inequality. For the second, fix  $x \in V$ , and  $A \subseteq V$  with  $\pi(A) \geq 1/2$ . Then we have that

$$\{T_A > \tau_o(\epsilon - \delta) + s_\delta\} \subseteq \{T_o > \tau_o(\epsilon - \delta)\} \cup \{\text{starting from } o, T_A > s_\delta\}$$

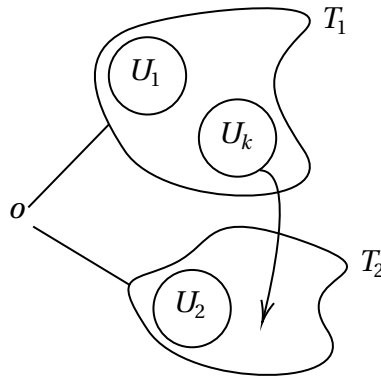
using the Markov property and a union bound it follows that

$$\begin{aligned}\mathbf{P}_x(T_A > \tau_o(\epsilon - \delta) + s_\delta) &\leq \mathbf{P}_x(T_o > \tau_o(\epsilon - \delta)) + \mathbf{P}_o(T_A > s_\delta) \\ &\leq \epsilon - \delta + \mathbf{P}_o(T_A > s_\delta)\end{aligned}$$

Where in the second inequality we have used the definition of  $\tau_o$ . Therefore we must now show the bound  $\mathbf{P}_o(T_A > s_\delta) < \delta$ . If  $o \in A$ , then the probability is zero, so it is trivially true. Now consider  $o \notin A$ . We now have the following claim.

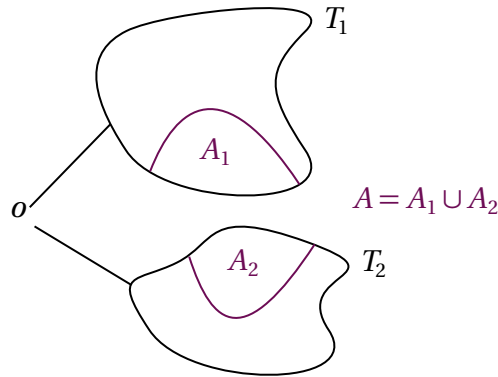
**Claim:**  $T \setminus \{o\}$  can be split as  $T \setminus \{o\} = T_1 \cup T_2$  where both  $T_1$  and  $T_2$  are each unions of the connected components of  $T \setminus \{o\}$ , and each  $\pi(T_1)$  and  $\pi(T_2)$  is  $\leq 2/3$ .

**Proof of Claim:** Label the connected components of  $T \setminus \{o\}$  as  $U_1, U_2, \dots, U_k$  in order of size. i.e:  $\pi(U_1) \geq \pi(U_2) \geq \dots \pi(U_k)$ . If there are only two such components, say  $U_1$  and  $U_2$ , by definition of  $o$ , it must be that  $\pi(U_1), \pi(U_2) \leq 1/2$  and as such we are done. Otherwise, suppose that  $k \geq 3$ . Note that for any  $j \geq 3$ ,  $\pi(U_j) \leq 1/3$ , otherwise, the sums of the sizes would go over 1. Consider any splitting  $T \setminus \{o\} = T_1 \cup T_2$  where  $U_1 \in T_1$  and  $U_2 \in T_2$ . Then if one of  $T_1$  or  $T_2$ , say without loss of generality has  $\pi(T_1) \geq 2/3$ , then it must be that  $\pi(T_2) \leq 1/3$ . Now grab any  $U_j \in T_1$  with  $j \geq 3$  and move it to  $T_2$ . We have decreased the size of  $T_1$ , and the size of  $T_2$  is of course at most  $2/3$  (because originally the size was at most  $1/3$  and now we have added a component  $U_j$  of size at most  $1/3$ ). If  $\pi(T_1)$  is still  $\geq 2/3$ , then repeat, get a component  $U_j \in T_1$  with  $j \geq 3$  and move it to  $T_2$ . It will still be the case that we have decreased the size of  $T_1$  whilst not increasing the size of  $T_2$  above  $2/3$ . The picture is something like this:



We can keep repeating this until eventually  $\pi(T_1)$  must drop below  $2/3$ . If at a certain iteration of this algorithm, we have that suddenly  $\pi(T_2)$  has increased from size  $\leq 2/3$  to size  $\geq 2/3$ , then it must be that after this step,  $\pi(T_1) \leq 1/3$ , and so we can undo this last step, and now we will have that both  $T_1$  and  $T_2$  have sizes at most  $2/3$ . Claim proven.

Now back to the main argument, we can split our original set  $A$  which was arbitrary and assumed to be of size  $\geq 1/2$  into the components that live in  $T_1$  and  $T_2$ , say  $A_i = T_i \cap A$ . Of course it must be that since  $A_1 \cup A_2 = A$ , it can't be that the size of both  $A_1$  and  $A_2$  are simultaneously less than  $1/4$ .



So say without loss of generality that  $\pi(A_1) \geq 1/4$ . Note that if starting at zero you haven't hit  $A$  by time  $s_\delta$ , then you for sure haven't hit  $A_1$ , so we have the following

$$\begin{aligned} \mathbf{P}_o(T_A > s_\delta) &\leq \mathbf{P}_o(T_{A_1} > s_\delta) \\ &\leq \mathbf{P}_{\pi_{T_2 \cup \{o\}}}(T_{A_1} > s_\delta) \\ &\leq \pi(T_2 \cup \{o\})^{-1} \mathbf{P}_\pi(T_{A_1} > s_\delta) \end{aligned}$$

Where the first inequality was explain above, the second inequality comes from the fact that if starting at  $o$  you haven't hit  $A_1$  by time  $s_\delta$ , then you also haven't hit  $A_1$  when starting anywhere in  $T_2 \cup \{o\}$ , and the third inequality comes from the law of total probability, i.e:  $\mathbf{P}_\pi(\cdots) =$

$\mathbf{P}_{\pi_{T_2 \cup \{0\}}}(\dots)\pi(T_2 \cup \{0\}) + \underbrace{\mathbf{P}_{\pi_{T_1}}(\dots)\pi(T_1)}_{\geq 0}$ . Rejoice for we are almost done, now upper bound

$$\mathbf{P}_{\pi}(T_{A_1} > s_{\delta}) \leq \pi(A_1^c) \exp\left(-\frac{s_{\delta}\pi(A_1)}{t_{\text{rel}}}\right)$$

Since  $\pi(T_1) \leq 2/3$ , then  $\pi(T_2 \cup \{0\}) \geq 1/3$ , and using the value of  $s_{\delta}$  we defined at the very start, we have shown finally that

$$\mathbf{P}_x(T_A > \tau_o(\epsilon - \delta) + s_{\delta}) \leq \epsilon$$

which means that  $\text{hit}_{1/2}(\epsilon) \leq \tau_o(\epsilon - \delta) + s_{\delta}$  as we intended to prove. ♡

We can now state the main result of this section.

**Theorem 14.19 (Cutoff for Markov chains on trees)** For a Markov chain on a tree  $T = (V, E)$  with  $|V| \geq 3$ , and for all  $\epsilon \leq 1/4$ , we have that

$$t_{\text{mix}}(\epsilon) - t_{\text{mix}}(1 - \epsilon) \lesssim \sqrt{\epsilon^{-1} t_{\text{mix}} t_{\text{rel}}}$$

We first need the following

**Lemma 14.20** For a Markov chain on a tree  $T = (V, E)$  with  $|V| \geq 3$ , we have that

$$t_{\text{mix}}(\epsilon) - t_{\text{mix}}(1 - \epsilon) \lesssim \tau_o(\epsilon) - \tau_o(1 - \epsilon)$$

*Proof.* [come back to this](#) ♡

This means that to prove the Theorem, it is enough to show that  $\tau_o(\epsilon) - \tau_o(1 - \epsilon) \lesssim \sqrt{\frac{t_{\text{mix}} t_{\text{rel}}}{\epsilon}}$ . To show this, we will inspect how  $\tau_o$  is concentrated around its mean. To show concentration around its mean we need to be able to bound the variance of hitting times:

**Proposition 14.21** Let  $P$  be reversible and irreducible with invariant distribution  $\pi$ . Let  $A$  be a non-empty and proper subset of  $\Omega$ , and define the distribution  $\psi_{A^c}$  on  $A^c$  by

$$\psi_{A^c}(y) = \mathbf{P}_{\pi_A}(X_1 = y \mid X_1 \in A^c)$$

, i.e: the distribution where you are after one time step in  $A^c$  given that you exited  $A$ . Then for all  $t \geq 1$ :

$$\frac{\mathbf{P}_{\pi_{A^c}}(T_A = t)}{\Phi(A^c)} = \mathbf{P}_{\psi_{A^c}}(T_A \geq t).$$

Where  $\Phi(A)$  is the conductance of a set, i.e:  $\Phi(A) = \frac{Q(A, A^c)}{\pi(A)}$ . This implies that

$$\mathbf{E}_{\psi_{A^c}}[T_A] = \frac{1}{\Phi(A^c)} \quad \text{and} \quad \mathbf{E}_{\psi_{A^c}}[T_A^2] = \mathbf{E}_{\psi_{A^c}}[T_A](2\mathbf{E}_{\pi_{A^c}}[T_A] - 1) \leq \frac{2\mathbf{E}_{\psi_{A^c}}[T_A]t_{\text{rel}}}{\pi(A)}$$

*Proof.* To prove the first equality, we start by noticing that

$$\{T_A = t\} = \{X_0 \notin A, \dots, X_{t-1} \notin A, X_t \in A\}$$

and

$$\{T_A^+ = t + 1\} = \{X_1 \notin A, \dots, X_t \notin A, X_{t+1} \in A\}$$

Stationarity of  $\pi$  means that

$$\mathbf{P}_{\pi}(T_A = t) = \mathbf{P}_{\pi}(T_A^+ = t + 1)$$

and so for all  $t \geq 1$ , we can compute

$$\begin{aligned} \pi(A^c)\mathbf{P}_{\pi_{A^c}}(T_A = t) &= \mathbf{P}_{\pi}(T_A = t) = \mathbf{P}_{\pi}(T_A^+ = t + 1) \\ &= \mathbf{P}_{\pi}(X_1 \notin A, \dots, X_t \notin A, X_{t+1} \in A) \\ &= \mathbf{P}_{\pi}(X_1 \notin A, \dots, X_t \notin A) - \mathbf{P}_{\pi}(X_1 \notin A, \dots, X_{t+1} \notin A) \\ &= \mathbf{P}_{\pi}(X_1 \notin A, \dots, X_t \notin A) - \mathbf{P}_{\pi}(X_0 \notin A, \dots, X_t \notin A) \\ &= \mathbf{P}_{\pi}(X_0 \in A, X_1 \notin A, \dots, X_t \notin A) \\ &\stackrel{(1)}{=} \pi(A)\Phi(A)\mathbf{P}_{\psi_{A^c}}(X_1 \notin A, \dots, X_t \notin A) \\ &\stackrel{(2)}{=} \pi(A)\Phi(A)\mathbf{P}_{\psi_{A^c}}(T_A \geq t) \end{aligned}$$

Where (1) comes from the fact that we have conditioned on having exited  $A$  in one step of the chain, and as such had to multiply by the probability of having exited the chain in one step which is precisely  $\pi(A)\Phi(A)$ . Equality (2) comes from the fact that  $\mathbf{P}_{\psi_{A^c}}(X_1 \notin A, \dots, X_t \notin A)$  is alternatively written as

$$\mathbf{P}_{X_1 \sim \psi_{A^c}}(X_1 \notin A, \dots, X_t \notin A)$$

and as such we can shift indices back and now (2) can easily be seen. This completes the first equality because  $Q(A, A^c) = Q(A^c, A)$  due to reversibility and as such  $\Phi(A)\pi(A) = \Phi(A^c)\pi(A^c)$ . With the first equality in mind, we can now compute the expectations, this is not too bad:

$$\mathbf{E}_{\psi_{A^c}}[T_A] = \sum_{t=0}^{\infty} \mathbf{P}_{\psi_{A^c}}(T_A \geq t) = \frac{1}{\Phi(A^c)} \sum_{t=0}^{\infty} \mathbf{P}_{\pi_{A^c}}(T_A = t) = \frac{1}{\Phi(A^c)}$$

For the second one, we note the following trick:

**Trick:** If a random variable takes values in  $\mathbf{N}$ , just as we had  $\mathbf{E}[X] = \sum_{t \geq 1} \mathbf{P}(X \geq t)$ , we also have that  $\mathbf{E}[X^2] = \sum_{t \geq 1} (2t - 1) \mathbf{P}(X \geq t)$ .

**Proof of trick:** this follows immediately after noting that  $t^2 = \sum_{s=1}^t (2s - 1)$ .

Back to the computation, we now have that

$$\begin{aligned} \mathbf{E}_{\psi_{A^c}}[T_A^2] &= \sum_{s \geq 1} (2s - 1) \mathbf{P}_{\psi_{A^c}}(T_A \geq s) \\ &= \sum_{s \geq 1} (2s - 1) \frac{\mathbf{P}_{A^c}(T_A = s)}{\Phi(A^c)} \\ &= \frac{2\mathbf{E}_{\pi_{A^c}}[T_A] - 1}{\Phi(A^c)} \\ &\leq \frac{2\mathbf{E}_{\psi_{A^c}}[T_A] t_{\text{rel}}}{\pi(A)} \end{aligned}$$

Where the only mysterious step is the last one, and in here we have used the fact that  $\Phi(A^c)$  has just been shown to be  $\frac{1}{\mathbf{E}_{\psi_{A^c}}}$  and we know from Lemma 14.6, we can bound (noting that if you start from  $A$ ,  $\mathbf{E}_{\pi_A}[T_A] = 0$ ):

$$\mathbf{E}_{\pi_{A^c}}[T_A] = \frac{1}{\pi(A^c)} \mathbf{E}_{\pi}[T_A] \leq \frac{1}{\pi(A^c)} \pi(A^c) \frac{t_{\text{rel}}}{\pi(A)}.$$

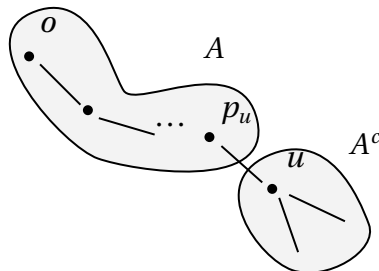
♡

This gives the following Lemma:

**Lemma 14.22 (Expected hitting time of the parent)** For  $u \neq o$ , and writing  $p_u$  for the parent vertex of  $u$ , and  $\mathcal{T}_u$  is the subtree of the vertices that connect  $o$  to  $u$ , we have that

$$\mathbf{E}_u[T_u] = \frac{\pi(\mathcal{T}_{p_u})}{\pi(u)P(u, p_u)} \quad \text{and} \quad \mathbf{E}_u[T_u^2] = 2\mathbf{E}_u[T_{p_u}]\mathbf{E}_{\pi_{\mathcal{T}_u}}[T_{p_u}] - \mathbf{E}_u[T_{p_u}]$$

*Proof.* We wish to use the previous proposition, and as such we need to cleverly choose a set  $A$ . We will use  $A$  to be the chain of vertices that connect  $o$  with the parent of  $u$ :



Now we note that if you start at  $A$ , and the distribution of where you are in one step conditioned on having reached  $A^c$  in one step is precisely a point mass at  $u$ . Moreover, if you start at  $u$ , the hitting time to reach set  $A$  is precisely the hitting time of  $p_u$ . Therefore

$$\mathbf{E}_{\psi_{A^c}}[T_A] = \mathbf{E}_u[T_{p_u}] = \frac{\pi(\mathcal{T}_u)}{\pi(u)P(u, p_u)}$$

as required. The second result also follows immediately now. ♡

With this in mind, we can now provide a concentration on the hitting time of an ancestor, we can give something slightly more general:

**Proposition 14.23 (Concentration of hitting time)** Let  $y \prec x$ , i.e:  $y$  is in the path from  $o$  to  $x$ . Let  $c > 0$  and define

$$\sigma_{xy} = \sqrt{4\mathbf{E}_x[T_y]t_{\text{rel}}}$$

Then

$$\text{Var}_x[T_y] \leq \sigma_{xy}^2$$

In particular, by Markov's inequality we have the concentration inequalities:

$$\mathbf{P}_x(T_y \geq \mathbf{E}_x[T_y] + c\sigma_{xy}) \leq \frac{1}{1+c^2} \quad \mathbf{P}_x(T_y \leq \mathbf{E}_x[T_y] - c\sigma_{xy}) \leq \frac{1}{1+c^2}$$

*Proof.* Consider the path from  $x$  to  $y$ :

$$v_0 = x \succ v_1 \succ \cdots \succ v_k = y$$

The key idea is that the variance of the hitting time of  $y$  starting from  $x$  is the sum of the variances of hitting time of  $v_{i+1}$  starting from  $v_i$  because all of these random variables are independent. Therefore

$$\begin{aligned} \text{Var}_x[T_y] &\stackrel{(1)}{=} \sum_{i=0}^{k-1} \text{Var}_{v_i}[T_{v_{i+1}}] \\ &\stackrel{(2)}{\leq} \sum_{i=0}^{k-1} \mathbf{E}_{v_i}[T_{v_{i+1}}^2] \\ &\stackrel{(3)}{\leq} \sum_{i=0}^{k-1} 4t_{\text{rel}} \mathbf{E}_{v_i}[T_{v_{i+1}}] \\ &\stackrel{(4)}{=} \sigma_{xy}^2 \end{aligned}$$

Where (1) comes from the discussion above, (2) comes from the fact that  $\text{Var}[X] = \mathbf{E}[X^2] - \mathbf{E}[X]^2$ .

Step (3) comes from the fact that from the previous Lemma and Proposition 14.21

$$\mathbf{E}_u[T_{p_u}^2] = 2\mathbf{E}_u[T_{p_u}]\mathbf{E}_{\pi_{T_u}}[T_{p_u}] - \mathbf{E}_u[T_{p_u}] \leq \frac{2\mathbf{E}_u[T_{p_u}]\mathbf{t}_{\text{rel}}}{\pi(\mathcal{T}_u^c)}$$

and when  $u = v_i$ ,  $p_u = v_{i+1}$  so this quantity can be brutally upper bounded by  $4\mathbf{E}_{v_i}[T_{v_{i+1}}]\mathbf{t}_{\text{rel}}$ . Step (4) comes from the fact that the expected hitting time of  $y$  starting at  $x$  is the sum of the expected hitting times along the path from  $x$  to  $y$ . To show the last two inequalities, which just to remark are slightly better than Chebyshev's inequality (hence why we can't directly use this inequality), we make the following computation:

$$\begin{aligned} \mathbf{P}_x(T_y - \mathbf{E}_x[T_y] \geq c\sigma_{xy}) &= \mathbf{P}_x\left(T_y - \mathbf{E}_x[T_y] + \frac{\sigma_{xy}}{c} \geq c\sigma_{xy} + \frac{\sigma_{xy}}{c}\right) \\ &= \mathbf{P}_x\left(\left[T_y - \mathbf{E}_x[T_y] + \frac{\sigma_{xy}}{c}\right]^2 \geq \left[c\sigma_{xy} + \frac{\sigma_{xy}}{c}\right]^2\right) \\ &\leq \frac{1}{\left[c\sigma_{xy} + \frac{\sigma_{xy}}{c}\right]^2} \mathbf{E}_x\left[\left[T_y - \mathbf{E}_x[T_y] + \frac{\sigma_{xy}}{c}\right]^2\right] \\ &= \frac{1}{\left[c\sigma_{xy} + \frac{\sigma_{xy}}{c}\right]^2} \left(\mathbf{E}_x\left[(T_y - \mathbf{E}_x[T_y])^2\right] + 2\frac{\sigma_{xy}}{c} \underbrace{\mathbf{E}_x[T_y - \mathbf{E}_x[T_y]]}_0 + \frac{\sigma_{xy}^2}{c^2}\right) \\ &= \frac{1}{\left[c\sigma_{xy} + \frac{\sigma_{xy}}{c}\right]^2} \left(\text{Var}_x[T_y] + \frac{\sigma_{xy}^2}{c^2}\right) \\ &\stackrel{(!)}{\leq} \frac{1}{\left[c\sigma_{xy} + \frac{\sigma_{xy}}{c}\right]^2} \left(\sigma_{xy}^2 + \frac{\sigma_{xy}^2}{c^2}\right) = \frac{1}{1+c^2} \end{aligned}$$

The only non-trivial step is (!), where we used the inequality on variance which we just proved above. The other inequality follows by identical calculations.

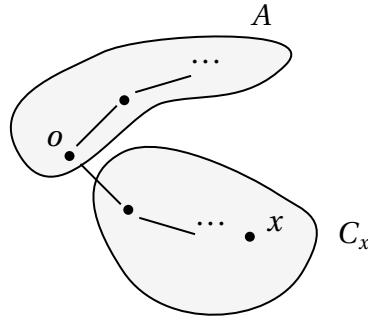
♡

Armed with these concentration inequalities we are ready to prove the main result of this section, namely that the product condition on trees is in fact sufficient for cutoff.

*Proof of Theorem 14.19.* We divide the proof into steps:

1. For any  $x \in V$ ,  $\mathbf{E}_x[T_o] \leq 4\mathbf{t}_{\text{mix}}$ :

If  $x = o$ , there's nothing to prove. Suppose that  $x \neq o$ , then let  $C_x$  be the connected component of  $T \setminus \{o\}$  that contains  $x$ , and let  $A = T \setminus C_x$ :



We will cleverly bound  $\mathbf{E}_x[T_o]$  as follows. Let  $\tau_A$  be the random variable defined as follows

$$\tau_A = \min\{k : X_{k t_{\text{mix}}} \in A\}$$

Since the only way to reach  $A$  by starting from  $x$  is by going through  $o$ , we can safely say that

$$T_o \leq \tau_A t_{\text{mix}}. \quad (\star)$$

We also have the observation that since  $C_x$  is a connected component of  $T \setminus \{o\}$ , by definition of a root node, it follows that  $\pi(C_x) \leq 1/2$ , so in particular  $\pi(A) \geq 1/2$ . Moreover, by definition of the mixing time, we have that for any starting point  $y \in V$ ,  $|P^{t_{\text{mix}}}(y, A) - \pi(A)| \leq 1/4$  so

$$P^{t_{\text{mix}}}(y, A) \geq \pi(A) - \frac{1}{4} \geq \frac{1}{4}.$$

In other words,  $\mathbf{P}_y(X_{t_{\text{mix}}} \in A) < \frac{1}{4}$ . This means that we can bound  $\tau_A$  by making repeated walks of length  $t_{\text{mix}}$  starting at the endpoints of the previous walks to bound  $\tau_A$  stochastically by a geometric random variable with parameter  $1/4$ . (Indeed, a failure of such a trial occurs with probability less than  $3/4$ ), therefore  $\mathbf{E}_x[\tau_A] \leq 4$ . And combining with  $(\star)$  we get that  $\mathbf{E}_x[T_o] \leq 4 t_{\text{rel}}$ .

## 2. Bound $\tau_o(\epsilon)$ and $\tau_o(1-\epsilon)$ :

Start by fixing  $\epsilon \in (0, 1/4]$ , and consider the worst expected time to hit the root,  $\tau := \max_{x \in V} \mathbf{E}_x[T_o]$ . Define also the constant  $\kappa_\epsilon := \sqrt{4\epsilon^{-1}\tau t_{\text{rel}}}$ , which as a preliminary observation, note that  $\kappa_\epsilon \asymp \sqrt{t_{\text{rel}}}$ . Let also  $c = \sqrt{\epsilon^{-1} - 1}$ . Then we have the following: for any



$y \in V$

$$\begin{aligned} \mathbf{P}_y(T_o \geq \tau + c\sqrt{4\tau t_{\text{rel}}}) &\stackrel{(1)}{\leq} \mathbf{P}_y\left(T_o \geq \mathbf{E}_y[T_o] + c\underbrace{\sqrt{4\mathbf{E}_y[T_o] t_{\text{rel}}}}_{\sigma_{xy}}\right) \\ &\stackrel{(2)}{\leq} \frac{1}{1+c^2} = \epsilon \end{aligned}$$

Where (1) comes from the fact that if starting from  $y$  you haven't hit the root by time  $\tau + c\sqrt{4\tau t_{\text{rel}}}$ , given that  $\tau \geq \mathbf{E}_y[T_o]$ , you definitely haven't hit the root by time  $\mathbf{E}_y[T_o] + c\sqrt{4\mathbf{E}_y[T_o] t_{\text{rel}}}$ . Step (2) comes from the concentration inequalities developed in the Proposition 14.23. Now we may quickly recall that  $\tau_o(\epsilon)$  is defined as

$$\tau_o(\epsilon) = \min\{t : \mathbf{P}_y(T_o > t) \leq \epsilon \text{ for all } y \in V\}$$

And so from this it follows that  $\tau_o(\epsilon) \leq \tau + c\sqrt{4\tau t_{\text{rel}}} \leq \tau + \kappa_\epsilon$ , where in this last inequality we noted that  $c := \sqrt{\epsilon^{-1}-1} < \sqrt{\epsilon^{-1}}$ . We have now upper bounded  $\tau_o(\epsilon)$ , to complete the proof we need to use Lemma 14.20, and so we also need to lower bound  $\tau_o(1-\epsilon)$ . To do this, we will use the second concentration inequality we developed in Proposition 14.23.

Let  $x$  be the vertex which attains the maximum in the definition of  $\tau := \max_{x \in V} \mathbf{E}_x[T_o]$ . Then by an immediate application of the second concentration inequality, we obtain that

$$\mathbf{P}_x(T_o \leq \tau - c\sqrt{4\tau t_{\text{rel}}}) \leq \frac{1}{1+c^2} := \epsilon$$

which in particular means that

$$\max_{y \in V} \mathbf{P}_y(T_o > \tau - c\sqrt{4\tau t_{\text{rel}}}) \geq \mathbf{P}_x(T_o > \tau - c\sqrt{4\tau t_{\text{rel}}}) > 1 - \epsilon$$

Quickly noting again that of course,  $\tau_o(1-\epsilon)$  is defined as

$$\tau_o(1-\epsilon) = \min\left\{t : \max_{y \in V} \mathbf{P}_y(T_o > t) \leq 1 - \epsilon\right\}$$

It must be that  $\tau(1-\epsilon) > \tau - c\sqrt{4\tau t_{\text{rel}}} > \tau - \kappa_\epsilon$ .

3. Finish off:

Lemma 14.20

$$\begin{aligned}
 t_{\text{mix}}(\epsilon) - t_{\text{mix}}(1 - \epsilon) &\stackrel{(1)}{\lesssim} \tau_o(\epsilon) - \tau(1 - \epsilon) \\
 &\stackrel{(2)}{\lesssim} \tau + \kappa_\epsilon - \tau + \kappa_\epsilon \\
 &\stackrel{(3)}{=} 4\sqrt{\epsilon^{-1}\tau t_{\text{rel}}} \\
 &\stackrel{(4)}{\lesssim} \sqrt{\epsilon^{-1}t_{\text{mix}} t_{\text{rel}}}
 \end{aligned}$$

Where (1) comes from Lemma 14.20, (2) comes from the bounds we have just established on  $\kappa_\epsilon$ , (3) comes from the definition of  $\kappa_\epsilon$ , and (4) comes from the fact that since for any  $x$ , we showed in the first part of this long proof that  $\mathbf{E}_x[T_o] \leq 4t_{\text{mix}}$ , it follows in particular that  $\tau \leq 4t_{\text{mix}}$  which shows of course that  $\tau \lesssim t_{\text{mix}}$



# Chapter 15

## Electrical Networks

In this chapter we work with reversible Markov Chains only. We will develop a way of thinking about Markov chains in terms of electrical networks, in particular, by thinking of Markov chains as weighted random walks on graphs.

**Definition 15.1** (Conductance and resistance) Let  $G = (V, E)$  be an undirected graph and let  $\{c(e)\}_{e \in E}$  be a collection of non-negative real numbers, which we call **conductances**. For an edge  $e = (x, y)$  we write  $c(x, y) = c(e) = c(y, x)$  and we call the reciprocal  $r(e) = 1/c(e)$  the **resistance** of an edge  $e$ .

A Markov chain can be defined on  $G$  with the conductances  $\{c(e)\}_{e \in E}$  by setting

$$P(x, y) = \frac{c(x, y)}{\sum_{x \sim z} c(x, z)} \equiv \frac{c(x, y)}{c(x)}$$

**Remark 15.2** (Invariant distribution and reversibility) For a Markov chain described as above if we set  $C_G = \sum_{x \in V} c(x)$ , then we can check that  $\pi(x) = \frac{c(x)}{C_G}$  is the invariant distribution. Indeed:

$$\begin{aligned} (\pi P)(x) &= \sum_{y \in \Omega} \pi(x) P(x, y) \\ &= \sum_{y \in \Omega} \frac{c(x)}{C_G} \times \frac{c(x, y)}{c(x)} \\ &= \sum_{y \in \Omega} \frac{c(x, y)}{C_G} = \frac{c(x)}{C_G} = \pi(x) \end{aligned}$$

And moreover, we can also check that  $\pi$  is in detailed balance with  $P$ , hence establishing that the

chain is reversible. Indeed:

$$\begin{aligned}\pi(x)P(x, y) &= \frac{c(x)}{c_G} \frac{c(x, y)}{c(x)} \\ &= \frac{c(y)}{c_G} \frac{c(y, x)}{c(y)} \\ &= \pi(y)P(y, x)\end{aligned}$$

Moreover, any reversible chain can be represented as a weighted random walk on a graph by setting the conductances to be what we previously understood as conductances:  $c(x, y) = \pi(x)P(x, y)$ . Then the assumption of reversibility ensures that  $c(x, y) = c(y, x)$ . Moreover, with this definition,  $c(x) = \sum_y c(x, y) = \sum_y \pi(x)P(x, y) = \pi(x)$ , so the transition probabilities of the weighted random walk coincide with  $P(x, y)$ .

**Definition 15.3** (Harmonic function) A function  $f : V \rightarrow \mathbf{R}$  is called **harmonic** for  $P$  at a vertex  $x \in V$  is

$$f(x) = \sum_{y \in V} P(x, y)f(y),$$

or in more compact notation, if  $f$  satisfies  $f(x) = (Pf)(x)$ . A function is called harmonic if it is harmonic at all vertices  $x \in V$ , i.e: if we have that as functions:  $f = Pf$ .

**Remark 15.4** This definition of Harmonic coincides with the intuition from Analysis, it just says that the value of the function at a point  $x$  is equal to the average of the function after one step of the chain.

Harmonic functions enjoy a special extension property:

**Proposition 15.5** (Unique extension of harmonic functions) Let  $X$  be an irreducible Markov chain on  $\Omega$  with transition matrix  $P$ . Let  $B \subseteq \Omega$  be a set and let  $f : B \rightarrow \mathbf{R}$  be a function defined on  $B$ . Then the function defined by

$$h(x) = \mathbf{E}_x[f(X_{T_B})]$$

is the unique extension  $h : \Omega \rightarrow \mathbf{R}$  of  $f$  which is harmonic in  $\Omega \setminus B$ .

*Proof.* We first check that  $h$  indeed extends  $f$ . If  $x \in B$ , then the hitting time  $T_B = 0$  and so since the expectation is for the chains started at  $x$ , we have that  $\mathbf{E}_x[f(X_{T_B})] = f(x)$ . Now we check that  $h$  is indeed harmonic on  $\Omega \setminus B$ . For this we will condition on the first step of the chain started at

$x$ , (1), and then use the Markov Property, (2):

$$\begin{aligned} h(x) &\stackrel{(1)}{=} \sum_{y \in \Omega} P(x, y) \mathbf{E}_x [f(X_{T_B}) | X_1 = y] \\ &\stackrel{(2)}{=} \sum_{y \in \Omega} P(x, y) \mathbf{E}_y [f(X_{T_B})] \\ &= (Ph)(x) \end{aligned}$$

as required. Now we check uniqueness of the extension. Suppose two functions  $h$  and  $h'$  satisfy these properties. Then let  $g = h - h'$ . It is easy to see that harmonic functions form a vector space, indeed, if  $f$  and  $g$  are harmonic with respect to  $P$ , then

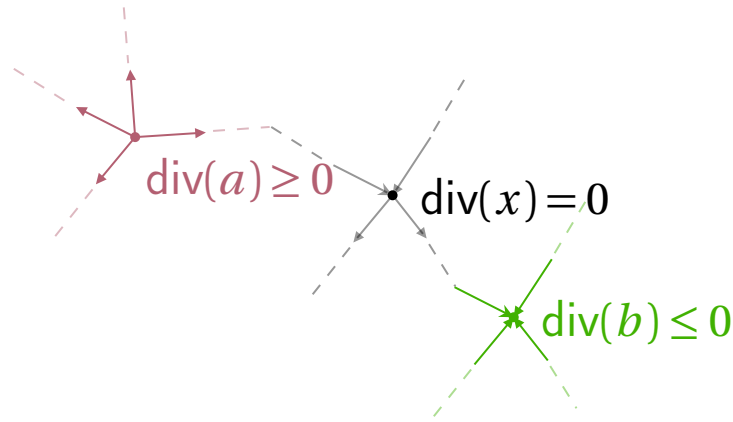
$$[P(\alpha f + \beta g)](x) = \alpha(Pf)(x) + \beta(Pg)(x) = (\alpha f + \beta g)(x)$$

So  $g$  is also harmonic on  $S \setminus B$ . Since we have assumed both functions extend  $f$ , it must be that  $g \equiv 0$  on  $B$ . Now we have the following argument. We will show that  $\max_y g(y) \leq 0$  and that  $\min_y g(y) \geq 0$ . Therefore  $g \equiv 0$  on all  $\Omega$ .

Assume  $\max_y g(y) > 0$  and let  $a = \operatorname{argmax}_y g(y)$ . Then suppose that there exists some neighbour  $x$  of  $a$  (i.e:  $P(a, x) > 0$ ) with  $g(x) < g(a)$ . Then

$$g(a) = \sum_{y \in \Omega} P(a, y) g(y) = \underbrace{g(x) P(a, x)}_{< g(a)} + \underbrace{\sum_{y \neq x} P(a, y) g(y)}_{\leq g(a) \sum_{y \neq x} P(a, y)} < g(a) \sum_y P(a, y) = g(a)$$

which is a contradiction. Therefore, for any neighbour  $x$  of  $a$ , we must have that  $g(x) = g(a) > 0$ . We could now repeat this argument with the fact that  $x = \operatorname{argmax}_y g(y)$ , which shows that for any connected path that contains  $a$ , all vertices  $v$  in said path must have  $g(v) > 0$ . But since the chain is irreducible, it must be that there is some path that leads to  $B$ , but we know that on  $B$ , the function  $g$  is equivalent to zero, which contradicts the above. We could now repeat the argument switching maximums for minimums and show the other required inequality. This shows that  $g \equiv 0$  and as such  $h = h'$ . ♡

Figure 15.1: Flow from  $a$  to  $b$ , an illustration

**Definition 15.6** (Flow and divergence) A **flow**  $\theta$  on  $G = (V, E)$  is a function defined on the set of directed edges  $\{(x, y)\}_{x \sim y}$  for which

$$\theta(x, y) = -\theta(y, x)$$

The **divergence** of the flow  $\theta$  at a vertex  $x$  is given by

$$\text{div}(\theta)(x) = \sum_{y \sim x} \theta(x, y)$$

We say  $a$  is a **source** and  $b$  is a **sink** for what we call a **flow from  $a$  to  $b$**   $\theta$ , if it satisfies Kirchhoff's node law:

$$\text{div}(\theta)(x) = 0 \quad \text{for any } x \notin \{a, b\}$$

and  $\text{div}(\theta)(a) \geq 0$ . For a flow from  $a$  to  $b$ , we define its **strength**, to be  $\|\theta\| := \text{div}(\theta)(a)$ , and we call it a unit flow if  $\|\theta\| = 1$ .

**Remark 15.7** (Kirchhoff's Law and Flows from  $a$  to  $b$ ) Observe that

$$\sum_{x \in V} \text{div}(\theta)(x) = \sum_{x \in V} \sum_{y \in x} \theta(x, y) = \sum_{(x, y) \in E} \theta(x, y) + \theta(y, x) = 0$$

Therefore, by assumption of Kirchhoff's Law being satisfied, when we have a flow from  $a$  to  $b$ , then  $\text{div}(\theta)(a) = -\text{div}(\theta)(b)$ .

**Definition 15.8** (Voltage, Current flow) Given two nodes  $a, b$ , a **voltage** between  $a$  and  $b$  is a Harmonic function on  $V \setminus \{a, b\}$  (And due to the harmonic extension theorem, it always exists for any given pair of boundary conditions  $W(a), W(b)$ , and is in fact unique).

Given a voltage  $W$ , we define a current flow  $I$ , a function of directed edges:

$$I(x, y) = \frac{W(x) - W(y)}{r(x, y)} = c(x, y)[W(x) - W(y)]$$

It is not a surprise that a current flow is a flow, since  $I(x, y) = -I(y, x)$

It is easy to check that given a current  $I$ , and a cycle of directed edges  $(x_1, x_2), \dots, (x_k, \underbrace{x_{k+1}}_{x_1})$ , we have that:

$$\sum_{i=1}^k r(x_i, x_{i+1})I(x_i, x_{i+1}) = 0$$

We refer to this condition as the cycle law.

**Proposition 15.9** (A flow that satisfies cycle law is a current flow) Let  $\theta$  be a flow from  $a$  to  $z$  that satisfies the cycle law and  $I$  be a current flow from  $a$  to  $z$ . If  $\|\theta\| = \|I\|$ , then  $\theta = I$ .

**Main idea:** We consider the function  $f = \theta - I$ , and by Harmonicity of voltage, show that  $\text{div} I(x) = 0$  for all  $x \in V \setminus \{a, b\}$  and hence  $\text{div} f(x) = 0$  for all  $x$ . Moreover, since both  $\theta$  and  $I$  satisfy the cycle law, then so does  $f$ . Now we can show that  $f$  is identically zero: pick any edge  $(x_1, x_2)$ , if  $f$  assigns a positive value to it, then by Kirchhoff Law, it must be that there is some edge  $(x_2, x_3)$  with strictly positive flow. Continue and eventually you return to  $(x_1, x_2)$ , violating Cycle Law.

*Proof.* Let  $f = \theta - I$ . Let us check that  $f$  satisfies Kirchhoff's node law at all nodes. If  $x \notin \{a, z\}$ , then

$$\text{div} f(x) = \sum_{y \sim x} \theta(x, y) - I(x, y) = \text{div} \theta(x) - \text{div} I(x)$$

From the definition of  $\theta$  being a flow from  $a$  to  $b$  it follows that  $\text{div} \theta(x) = 0$ , so all left to check

is that  $\text{div} I(x) = 0$ , this comes from the fact that voltage is harmonic. Indeed:

$$\begin{aligned} \text{div} I(x) &= \sum_{y \sim x} \frac{W(x) - W(y)}{r(x, y)} \\ &= \sum_{y \sim x} c(x, y)W(x) - c(x, y)W(y) \\ &= W(x)c(x) - \sum_{y \sim x} c(x, y)W(y) = 0 \end{aligned}$$

Where the last equality comes from the definition of transition probabilities in an electrical network and  $W$  being harmonic. If  $x \in \{a, z\}$ , then we first note that by the remark above,  $\text{div} f(a) = -\text{div} f(z)$ , and since  $\|\theta\| = \|I\|$ , then the divergences agree at  $a$ , so  $\text{div} f(a) = 0$ , and hence  $\text{div} f(x) = 0$  for all  $x \in V$ . Since  $\theta$  is assumed to satisfy the cycle law and  $I$  being a current flow automatically satisfies the cycle law it follows that  $f$  satisfies the cycle law too.

Now suppose that  $\theta \neq I$ , we may suppose WLOG that there is some edge  $(x_1, x_2)$  such that  $f(x_1, x_2) > 0$ . Then by Kirchoff's Law, we have that since the flow  $f$  coming into  $x_2$  from  $x_1$  is positive, there must be some edge  $x_3$ , for which the flow going out of  $x_2$  into  $x_3$  is positive, i.e:  $f(x_2, x_3) > 0$ . Repeating this step as many times as needed and using the fact that the electrical grid is finite by assumption, gives that eventually we will reach a vertex we have already visited, and thus have created a cycle  $x_1, x_2, \dots, x_k, x_{k+1} = x_1$  with  $f(x_i, x_{i+1}) > 0$ . This violates the cycle law, which states that

$$\sum_{i=1}^k r(x_i, x_{i+1})f(x_i, x_{i+1}) = 0$$



**Remark 15.10** The take home message of this proposition is that there is a unique unit current flow. To be pedantic, since we have shown that a current flow  $I$  satisfies Kirchoff's Law on vertices distinct from  $a$  and  $b$ , it also follows that  $I$  is a flow from  $a$  to  $b$ .



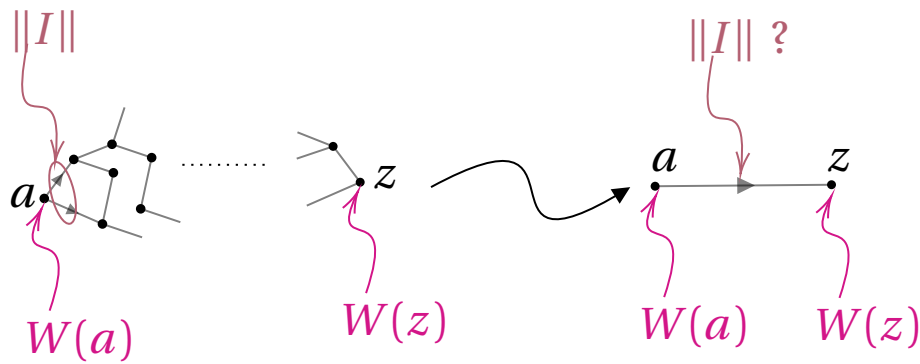


Figure 15.2: Effective resistance, the meaning behind the definition

## 15.1 Effective Resistance

Consider the following thought experiment: you have an electrical network  $(V, E)$  with resistances  $\{r(e)\}_{e \in E}$ . You plug in a voltage  $W$  from  $a$  to  $z$ . This voltage creates a current flow  $I$  from  $a$  to  $z$ , with strength  $\|I\| = \text{div} I(a)$ . Suppose we were to replace the entire network by one single edge from  $a$  to  $z$ , what resistance would we need to assign this edge so that the strength of the current on this new network coincided with the strength of the current on the previous network? We call this resistance  $R_{\text{eff}}(a, z)$  and by a trivial calculation we get the next definition:

**Definition 15.11** (Effective resistance) For a graph  $G = (V, E)$  with resistances  $\{r(e)\}_{e \in E}$ , and  $a, z \in V$ , the effective resistance between said edges is

$$R_{\text{eff}}(a, z) = \frac{W(a) - W(z)}{\|I\|}$$

Where  $W$  is any voltage from  $a$  to  $z$ , and  $I$  is the current corresponding to it. The effective conductance is defined as  $C_{\text{eff}} = 1/R_{\text{eff}}$ .

**Proposition 15.12** (Effective resistance is well defined) Effective resistance is a property of a graph and not of the choice of voltage.

**Main idea:** The key aspect of this proof relies on the fact that any voltage  $W$  from  $a$  to  $z$  is an affine transformation of the simplest voltage, the one with  $W_0(a) = 1$  and  $W_0(z) = 0$ .

*Proof.* Let us make the following observation:

- The space of Harmonic functions for  $P$  is closed under affine transformations. Indeed: if  $f$

is Harmonic for  $P$ , and  $A, B$  are real constants, then

$$(P(Af + B))(x) = \sum_y AP(x, y)f(y) + BP(x, y) = A(Pf)(x) + B = Af(x) + B$$

Then, if  $W$  is any voltage from  $a$  to  $z$  with boundary conditions  $W(a) = A$  and  $W(z) = Z$ , we can construct the following affine transformation of  $W_0$ , the voltage with  $W_0(a) = 1$  and  $W_0(z) = 0$ :

$$(A - Z)W_0(z) + Z$$

It is clear that the boundary conditions of this new function agree with the boundary conditions of our voltage  $W$ . Moreover, since this function is an affine transformation of the voltage  $W_0$ , it follows by uniqueness of Harmonic functions that

$$W(z) = (A - Z)W_0(z) + Z$$

(Note that this is powerful as it holds for all voltages!) Define now  $\|I_0\|$  to be the strength of the current flow on the network associated with the voltage  $W_0$ , and let  $\|I\|$  be the strength of the current associated to the given voltage  $W$ . We can now calculate that

$$\|I\| = \sum_{x \sim a} \frac{W(a) - W(x)}{r(a, x)} = \sum_{x \sim a} \frac{(A - Z)(W_0(a) - W_0(x))}{r(a, x)} = (A - Z)\|I_0\|$$

Therefore, for any voltage from  $a$  to  $z$ :

$$\frac{W(a) - W(z)}{\|I\|} = \frac{1}{\|I_0\|}$$



We now start to provide probabilistic connections between electrical networks and random walks. In particular we will use the effective resistance from  $a$  to  $z$  to calculate the probability that a Markov Chain started at  $a$  hits a vertex  $z$  before returning to  $a$ :

**Proposition 15.13 (Effective resistance and escape probability)** Let  $X$  be a reversible chain on a graph  $G = (V, E)$  and let  $\{c_e\}_{e \in E}, \{r_e\}_{e \in E}$  be the conductances and resistances respectively. Then for any  $a, z \in V$ , we have that

$$\mathbf{P}_a(T_z < T_a^+) = \frac{1}{c(a)R_{\text{eff}}(a, z)} =: \frac{C_{\text{eff}}(a, z)}{c(a)}$$

**Main idea:** The function  $f(x) = \mathbf{P}_x[T_z < T_a]$  is harmonic on  $V \setminus \{a, z\}$ , and since it has the same boundary conditions as an affine transformation of a voltage, it can be expressed as such. Then we can just compute  $\mathbf{P}_a[T_z < T_a^+]$  with a simple computation.

*Proof.* Let  $f(x) = \mathbf{P}_x[T_z < T_a]$ . We start by making the following two observations:

- $f(x)$  is Harmonic on  $V \setminus \{a, z\}$ : this is a routine trick:

$$\begin{aligned} f(x) &= \mathbf{P}_x[T_z < T_a] \\ &= \sum_y \mathbf{P}_x[X_1 = y] \mathbf{P}_x[T_z < T_a \mid X_1 = y] \\ &\stackrel{(!)}{=} \sum_y P(x, y) \mathbf{P}_y[T_z < T_a] = (Pf)(x) \end{aligned}$$

Where step (!) comes from the Markov Property to change the measure  $\mathbf{P}_x[\cdot \mid X_1 = y]$  into the measure  $\mathbf{P}_y[\cdot]$  and  $T_z$  remains unchanged, because  $x \notin \{a, z\}$ . (Indeed, if we were trying to perform this trick with the probability  $\mathbf{P}_x[T_z^+ < T_a]$ , it would have not worked because  $T_z^+$  is the first non-zero time, whereas  $T_z$  is the first time including time equals zero).

- $f(x)$  has the same boundary conditions as  $1 - W_0(x)$ , where  $W_0(x)$  is the "unit voltage" between  $a$  and  $z$ . (i.e:  $W_0(a) = 1 = 1 - W_0(z)$ ), therefore, given that  $1 - W_0(x)$  is an affine transformation of a harmonic function, it is also harmonic, and therefore by the uniqueness result  $f(x) = 1 - W_0(x)$ .

From this we are ready to prove the result:

$$\begin{aligned} \mathbf{P}_a[T_z < T_a^+] &= \sum_y P(a, y) \mathbf{P}_y[T_z < T_a] \\ &= \sum_y \frac{c(a, y)}{c(a)} (1 - W_0(y)) \\ &= \frac{1}{c(a)} \sum_y \frac{W_0(a) - W_0(y)}{r(a, y)} = \frac{\|I_0\|}{c(a)} \end{aligned}$$

and recall from the computation we did in the proof of Effective Resistance is Well Defined:  $R_{\text{eff}}(a, z) = 1/\|I_0\|$ . This finishes the claim. ♥

**Remark 15.14** The interpretation of this result is that the probability of visiting  $z$  before returning to  $a$  is inversely proportional to the effective resistance between  $a$  and  $z$ . A big effective resistance signifies a higher probability of returning to  $a$  before ever visiting  $z$ . The normalisation constant is to account for the degree of  $a$ .

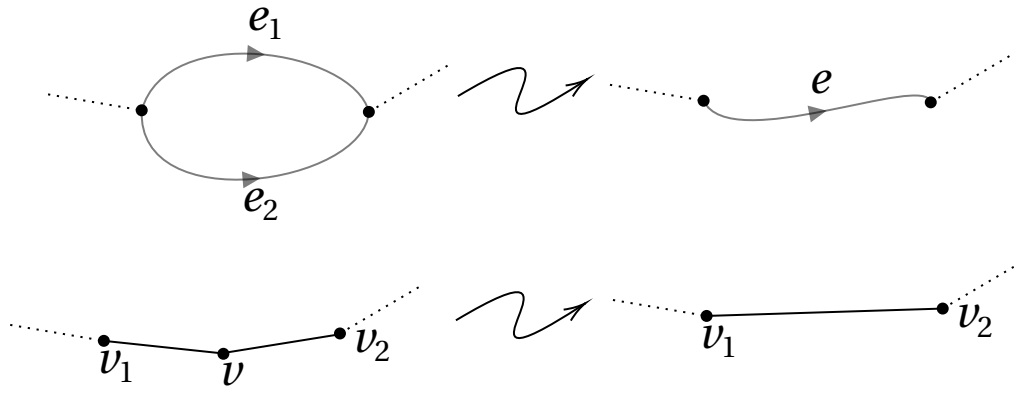


Figure 15.3: Simplifying electrical networks: conductances in parallel and resistances in series

We now define Green's function, which counts the expected number of times that a random walk started at  $a$  will hit  $x$  before a stopping time  $T$  rings.

**Definition 15.15 (Green's function)** For any stopping time  $T$  we define a **Green's function** for the Random Walk stopped at  $T$  by

$$G_T(a, x) = \mathbf{E}_a \left[ \sum_{i=1}^{\infty} \mathbf{1}(\{X_i = x\} \cap \{i < T\}) \right]$$

i.e. simply the function that counts how many times you've been in state  $x$ , having started from  $a$ , until the clock rings.

**Corollary 15.16 (Green's function and effective resistance)** For a reversible chain  $X$ , we have that

$$G_{T_z}(a, a) = c(a) R_{\text{eff}}(a, z).$$

*Proof.* By the Strong Markov Property, started from  $a$ , the chain will visit  $z$  before returning to  $a$  with probability  $\frac{1}{c(a) R_{\text{eff}}(a, z)}$ . If it returns to  $a$ , since  $T_a$  is a stopping time, we can restart the chain there, and the probability of visiting  $z$  again before returning to  $a$  is again  $\frac{1}{c(a) R_{\text{eff}}(a, z)}$ . From this argument it is evident that  $G_{T_z}(a, a) \sim \text{Geo}(1/c(a) R_{\text{eff}}(a, z))$ . Hence the result follows. ♡

The reason this result is quite useful is that we have the following simplification rules that make calculating effective resistance quite easy:

**Proposition 15.17 (Simplifying the electrical network: conductances in parallel)** Let  $e_1$  and  $e_2$  be edges with conductances  $c_1$  and  $c_2$  respectively. Suppose these edges share endpoints  $v_1$  and  $v_2$  (i.e a multigraph). Then both edges can be replaced with a single edge  $e$  of conductance  $c_1 + c_2$

without affecting the rest of the network, that is to say: all voltages in  $V$  and currents in  $E \setminus \{e_1, e_2\}$  are unchanged, and  $I(e) = I(e_1) + I(e_2)$ .

*Proof.* By looking at the diagram, it is clear that the transition probabilities of this new graph are in fact the same as the one we had before. Since the construction of a voltage (and hence its unicity) depends only on the transition probabilities and the boundary values, it follows that the voltage  $W$  on all vertices is the same. If we take any edge  $l$  other than  $\{e_1, e_2\}$ , it is clear that  $r(l)$  hasn't changed, therefore  $I(l)$  remains also unchanged. Finally, if  $e = (x, y)$  is the new edge  $I(e) = c(e)[W(y) - W(x)] = (c(e_1) + c(e_2))[W(y) - W(x)] = I(e_1) + I(e_2)$ . ♡

**Proposition 15.18** (Simplifying the electrical network: resistances in series) Let  $v \in V$  be a vertex of degree 2. Let  $v_1$  and  $v_2$  be its neighbouring vertices, then the edges  $(v_1, v)$  and  $(v_2, v)$  can be replaced by a single edge of resistance  $r(v_1, v) + r(v, v_2)$ . All the voltages in  $V \setminus \{v\}$  are unchanged, and all currents in  $E \setminus \{(v_1, v), (v, v_2)\}$  are unchanged. Moreover,  $I(v_1, v_2) = I(v_1, v) = I(v, v_2)$

**Proposition 15.19** (Simplifying the electrical network: short circuiting) Suppose two vertices  $x_1$  and  $x_2$  have the same voltage, then one can identify them as a single vertex  $z$  (morally this corresponds to adding a wire of zero resistance or infinite conductance, hence the name of the proposition) while keeping all edges and without changing the voltage or current.

Let us introduce the following Lemma.

**Lemma 15.20** (Self loops don't affect voltage) Suppose a network has a voltage  $W$  from  $a$  to  $b$ , and suppose  $y \in V$  has voltage  $W(y)$ , then if we add a self-loop of conductance 1 to  $y$ , its voltage remains the same

*Proof.* We simply use the characterisation of voltage as  $W(y) = \mathbf{E}_y[W(X_{T\{a,b\}})]$ . If we add a self-loop with conductance 1, a priori we have a new voltage  $W'$  and if we expand this:

$$W'(y) = \sum_{x \sim y} \frac{c(x, y)}{c(y) + 1} W'(x) = \frac{1}{c(y) + 1} W'(y) + \sum_{x \neq y, y \sim x} \frac{1}{c(y) + 1} W'(x)$$

Rearranging gives that

$$W'(y) = \frac{1}{c(y)} \sum_{y \sim x: x \neq y} c(x, y) W'(x)$$

Therefore our new voltage  $W'$  satisfies the exact same equations as  $W$  so they are the same. ♡

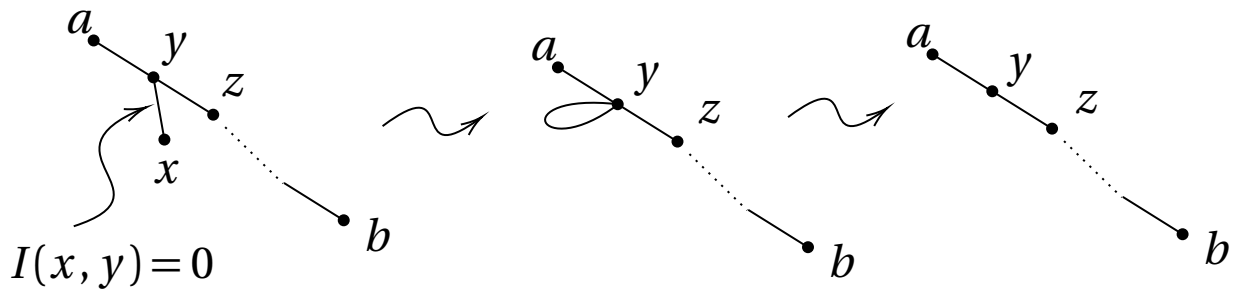


Figure 15.4: Effective resistance on trees

**Example 15.21 (Tree)** Let  $T$  be a finite connected tree which has resistance 1 on each edge. Then we have that for two vertices  $a$  and  $b$ ,  $R_{\text{eff}}(a, b)$  equals the graph distance between  $a$  and  $b$ .

*Proof.* We are going to simplify the network. Consider any vertex  $x$  that is not on the geodesic from  $a$  to  $b$ , let  $y$  be the closest vertex to  $x$  that lies on the geodesic. Since  $W(x) = \mathbf{E}_x[W(X_{T\{a,b\}})]$  it is clear that this will be also the same as  $\mathbf{E}_y[W(X_{T\{a,b\}})]$  because starting the walk at  $x$ , in order to get to  $\{a, b\}$ , we must go through  $y$ . Therefore, we can glue  $x$  to  $y$ , which gives a self-loop at  $y$ . However, by the previous Lemma, we can just ignore this loop, as it doesn't affect the voltage. Repeating this, we can kill all vertices that are not on the direct path from  $a$  to  $b$ , now we can just iterate the Resistances in Series Lemma. ♡

**Remark 15.22** Just to be super clear, I claim that the effective resistance between  $a$  and  $z$  in the first diagram is equal to the effective resistance in the last one. Indeed: effective resistance between  $a$  and  $z$  is precisely  $\frac{W(a) - W(z)}{\|I\|}$ , however, since after the looping business we have not altered the voltages nor the resistances,  $\|I\|, W(a), W(z)$  stay the same. Then one can start simplifying, to express  $R_{\text{eff}}(a, z) = R_{\text{eff}}(a, y) + R_{\text{eff}}(y, z)$ . This (allegedly) follows from the Resistors in Series Add Lemma, but let's show it. By choosing the unit voltage  $W_0$  from  $a$  to  $z$  its clear that

$$R_{\text{eff}}(a, z) = \frac{1}{\|I_0\|}$$

and  $\|I_0\| = I(a, y) = \frac{1 - W_0(y)}{r(a, y)}$ . Note that since  $I$  is a flow,  $I(a, y) = I(y, z) = I$ . Now we claim that if we remove  $y$  and put an edge with the appropriate resistance  $r(a, y) + r(y, z)$ , we also have that  $\|I_0\| := I(a, z)$  equals  $I$ . Indeed:  $(r(a, y) + r(y, z))I = (1 - W) + (W - 0) = 1$ , which shows that  $I = \frac{1 - 0}{r(a, z)}$  as required. Now you can keep iterating this and you get that  $R_{\text{eff}}(a, b)$  is the distance from  $a$  to  $b$ .

**Definition 15.23 (Energy)** Let  $\theta$  be a flow on a finite connected graph  $G$ . We define its energy by

$$\mathcal{E}(\theta) = \sum_{e \in E} (\theta(e))^2 r(e),$$

The following Theorem gives an equivalent characterisation of effective resistance as the minimal energy of unit flows:

**Theorem 15.24 (Thomson's principle)** For all  $a, z \in V$ , we have that

$$R_{\text{eff}}(a, z) = \inf\{\mathcal{E}(\theta) : \|\theta\| = 1 \text{ is a flow from } a \text{ to } b\}$$

Moreover this infimum is uniquely attained by the unit current flow from  $a$  to  $z$ .

*Proof.* Let  $I$  be the unit current flow from  $a$  to  $z$  with associated voltage  $W$ . We have the following:

$$\begin{aligned} \mathcal{E}(I) &= \frac{1}{2} \sum_{x, y: x \sim y} I(x, y)^2 r(x, y) \\ &\stackrel{(1)}{=} \frac{1}{2} \sum_{x, y: x \sim y} I(x, y)(W(x) - W(y)) \\ &\stackrel{(2)}{=} \sum_{x, y: x \sim y} I(x, y)W(x) \\ &\stackrel{(3)}{=} \sum_{x \in V} W(x) \text{div} I(x) \\ &\stackrel{(4)}{=} W(a) - W(z) \\ &\stackrel{(5)}{=} R_{\text{eff}}(a, z) \end{aligned}$$

Where (1) comes from Ohm's Law, (2) comes from noting that  $I(x, y) = -I(y, x)$  and relabelling indices on the second sum, (3) comes from the fact that summing over all edges  $(x, y)$  amounts to summing over all vertices  $x$  and then summing over all connections from  $x$ , as well as the definition of divergence, (4) comes from the fact that  $I$  is a flow from  $a$  to  $z$ , so the divergence at any  $x \notin \{a, z\}$  is zero, as well as the fact that  $-\text{div} I(z) = \text{div} I(a) = 1$ , and (5) comes from the definition of effective resistance using the fact that  $I$  is assumed to be unit strength.

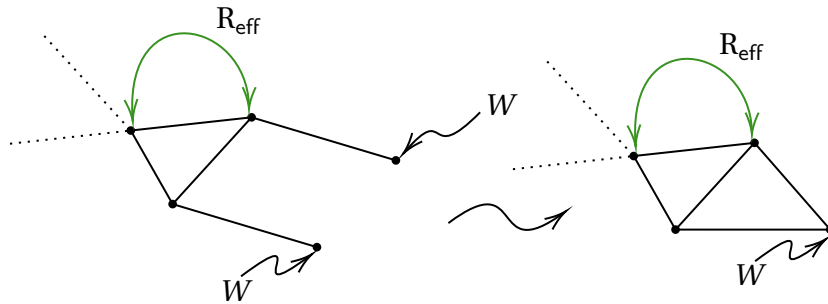
We now show that for any other unit flow from  $a$  to  $z$  (the key difference is that it may not be a current flow)  $J$  we have that  $\mathcal{E}(J) \geq \mathcal{E}(I)$ . Define the flow  $K = J - I$ , and notice that it has

strength zero by assumption of both  $I$  and  $J$  being flows from  $a$  to  $z$ . Then

$$\begin{aligned}
 \mathcal{E}(J) &:= \sum_{e \in E} J(e)^2 r(e) \\
 &:= \sum_{e \in E} [I(e) + K(e)]^2 r(e) \\
 &= \sum_{e \in E} I(e)^2 r(e) + \sum_{e \in E} K(e)^2 r(e) + 2 \sum_{e \in E} I(e)K(e)r(e) \\
 &\stackrel{(1)}{=} \mathcal{E}(I) + \mathcal{E}(K) + \sum_{x,y: x \sim y} (W(x) - W(y))K(x, y) \\
 &\stackrel{(2)}{=} \mathcal{E}(I) + \mathcal{E}(K) + 2 \sum_{x,y: x \sim y} W(x)K(x, y) \\
 &\stackrel{(3)}{=} \mathcal{E}(I) + \mathcal{E}(K) + 2 \sum_{x \in V} W(x) \operatorname{div} K(x) \\
 &\stackrel{(4)}{=} \mathcal{E}(I) + \mathcal{E}(K)
 \end{aligned}$$

Where step (1) comes from the fact that when summing over all pairs  $(x, y)$  where  $x \sim y$  counts all edges twice, as well as using the definition of  $I(x, y)$ . Step (2) comes from distributing and using anti-symmetry of  $K(x, y)$  and relabelling the sum. Step (3) comes from the definition of divergence, and finally step (4) comes from the fact that  $K$  has zero strength, so its divergence is zero at all points. Now finally, using the fact that  $\mathcal{E}(K)$  is a non-negative quantity. ♡

**Corollary 15.25** (Gluing does not increase effective resistance) The procedure of gluing two vertices with the same voltage does not increase effective resistance between any vertices that have not been glued together.



*Proof.* Every flow in the previous network is also a flow in the glued one. ♡



**Theorem 15.26 (Rayleigh's monotonicity principle)** Suppose that  $G$  is a graph and  $\{r_e\}_{e \in E}$ ,  $\{r'_e\}_{e \in E}$  are two sets of resistances for  $G$  with the property that  $r(e) \leq r'(e)$  for all edges. Then if we denote  $R_{\text{eff}}(a, z; r)$  to be the effective resistance between  $a$  and  $z$  corresponding to the set of resistances  $\{r(e)\}_{e \in E}$ , we have the following monotonicity principle:

$$R_{\text{eff}}(a, z; r) \leq R_{\text{eff}}(a, z; r').$$

*Proof.* This just follows from Thomson's principle. Let  $I$  and  $I'$  be the unit flows corresponding to the two sets of resistances, then by Thomson's principle:

$$R_{\text{eff}}(a, z; r) = \sum_{e \in E} I(e)^2 r(e) \leq \sum_{e \in E} I(e)^2 r'(e),$$

since the infimum is attained by  $I$ . Now we can just use the assumption on the resistances and deduce the claim. ♡

As an application of this result, we can ask the following question which is not immediate from a probabilistic perspective. Suppose  $G$  is a graph with resistances  $\{r(e)\}_{e \in E}$ , and suppose that we add one more edge with whatever resistance we wish. How does the escape probability  $\mathbf{P}_a(T_z < T_a^+)$  compare after this addition? It could seem that by placing this new edge in a clever way we can reduce this probability, say by making it more likely to return to  $a$ . This however turns out to not be the case: indeed, we can think of this new edge  $e'$  as already being there in the previous network, but having an associated resistance  $r(e')$  of infinity. By "adding" this new edge, we have reduced the resistance from infinity to some finite number, which by the monotonicity principle gives that the effective resistance of this new graph is at most the effective resistance of the previous graph. Since the return probabilities are inversely proportional to the effective resistances, we see then that in this new graph, the probability of escaping to  $z$  before returning to  $a$  is at least as large as the original escape probability.

## 15.2 Lower Bounds on Effective Resistance

We have seen that by constructing unit flows, one may upper bound effective resistance through the use of Thomson's principle. We now provide a method to lower bound effective resistances. The end goal is that as we will see, we will be able to say things about hitting times and cover times using effective resistance.

**Definition 15.27 (Cutset)** Let  $G = (V, E)$ . A set  $\Pi \subseteq E$  of edges is called a cutset separating  $a$  and  $z$  if every path between  $a$  and  $z$  contains an edge in  $\Pi$ .

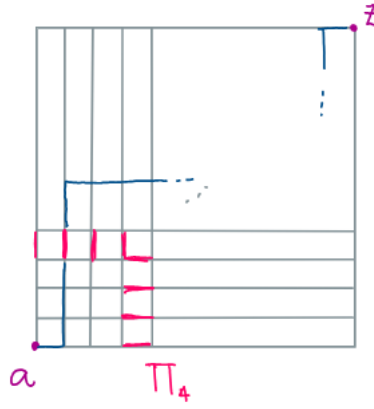


Figure 15.5: A cutset for a grid

Intuitively, if we can fit a lot of disjoint cutsets between  $a$  and  $z$ , it makes sense that the effective resistance between these two points is big. This intuition is made precise by the Nash-Williams inequality

**Theorem 15.28 (Nash-Williams)** Let  $\{\Pi_k : 1 \leq k \leq m\}$  be a disjoint collection of cutsets separating  $a$  and  $z$ . Then

$$R_{\text{eff}}(a, z) \geq \sum_{k=1}^m \left( \sum_{e \in \Pi_k} c(e) \right)^{-1}$$

*Proof.* All we need to show is that for any unit flow  $\theta$ , we have that

$$\mathcal{E}(\theta) \geq \sum_{k=1}^m \left( \sum_{e \in \Pi_k} c(e) \right)^{-1}$$

Then by minimising over  $\theta$  we will be done. First we note that since the collection  $\{\Pi_k\}$  are assumed to be disjoint, we have that

$$\sum_{e \in E} \theta(e)^2 r(e) \geq \sum_{k=1}^m \sum_{e \in \Pi_k} \theta(e)^2 r(e)$$

Now we note that

$$\left( \sum_{e \in \Pi_k} c(e) \right) \left( \sum_{e \in \Pi_k} \theta(e)^2 r(e) \right) \stackrel{(1)}{\geq} \left( \sum_{e \in \Pi_k} \sqrt{c(e)r(e)} |\theta(e)| \right)^2 = \left( \sum_{e \in \Pi_k} |\theta(e)| \right)^2 \stackrel{(2)}{\geq} 1$$

Step (1) is nothing but the Cauchy-Schwarz inequality. For step (2) we have to be a bit more

careful. Consider the set  $A = \{x \in V : x \leftrightarrow a \text{ in } G \setminus \Pi_k\} \cup \{a\}$ . Then we have that

$$\begin{aligned}
 1 &\stackrel{(1)}{=} \operatorname{div} \theta(a) \\
 &\stackrel{(2)}{=} \sum_{x \in A} \operatorname{div} \theta(x) \\
 &\stackrel{(3)}{=} \sum_{x \in A} \sum_{y \sim x} \theta(x, y) \\
 &\stackrel{(4)}{=} \sum_{x \in A} \sum_{y \sim x: y \in A} \theta(x, y) + \sum_{x \in A} \sum_{y \sim x: y \in A^c} \theta(x, y) \\
 &\stackrel{(5)}{=} \sum_{x \in A} \sum_{y \sim x: y \in A^c} \theta(x, y) \\
 &\stackrel{(6)}{=} \sum_{e \in \Pi_k} \theta(e) \leq \sum_{e \in \Pi_k} |\theta(e)|
 \end{aligned}$$

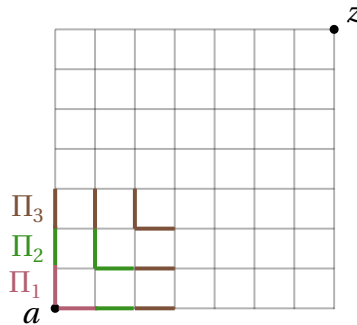
Where (1) follows from  $\theta$  having unit strength, (2) follows from the fact that  $\theta$  being a flow *from*  $a$  to  $z$  has that  $\operatorname{div} \theta(x) = 0$  for all  $x \notin \{a, z\}$ , and  $z$  certainly is not in  $A$ , because to connect to it one needs to go through the cutset. Step (3) is the definition of divergence, step (4) is trivial. Step (5) comes from the fact that if we are summing over the values of  $y$  that are also in  $A$ , we are going to be summing over both the pairs  $(x, y)$  and  $(y, x)$  so by the antisymmetry of the flow, the entire sum will be killed. Step (6) comes from the fact that the edges that go from  $A$  to  $A^c$  are precisely those from  $\Pi_k$  and the last step is a trivial bound.  $\heartsuit$

We are now ready to present an example, where we compute bounds for the effective resistance of a grid.

**Proposition 15.29 (Effective resistance on the grid)** Let  $a = (1, 1)$  and  $z = (n, n)$  be the lower and upper corners of the box  $B_n = \{1, 2, \dots, n\}^2$ . Suppose each edge has unit conductance, then

$$\frac{\log(n-1)}{2} \leq R_{\text{eff}}(a, z) \leq 2 \log n$$

*Proof.* We begin with the lower bound as its simpler. For this we need to come up with a collection of cutsets  $\{\Pi_k\}$ . A natural collection of cutsets to consider is the following:

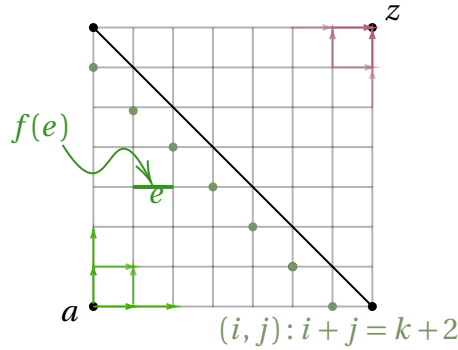


This is obviously a collection of cutsets and they are disjoint. By looking at the diagram its easy to convince yourself that we will need  $n-1$  such cutsets, and the  $i^{\text{th}}$  cutset has  $2i$  edges, therefore, by applying the Nash-William inequality, we have that

$$\begin{aligned} R_{\text{eff}}(a, z) &\stackrel{(1)}{\geq} \sum_{i=1}^{n-1} \left( \sum_{e \in \Pi_i} 1 \right)^{-1} \\ &\stackrel{(2)}{=} \frac{1}{2} \sum_{i=1}^{n-1} \frac{1}{i} \\ &\geq \frac{1}{2} \int_1^{n-1} \frac{1}{x} dx = \frac{\log(n-1)}{2} \end{aligned}$$

Where (1) comes from the fact that we are assuming unit conductance, (2) comes from the fact that we have just discussed that  $|\Pi_k| = 2k$ , and the last steps are just a simple integral comparison of a sum.

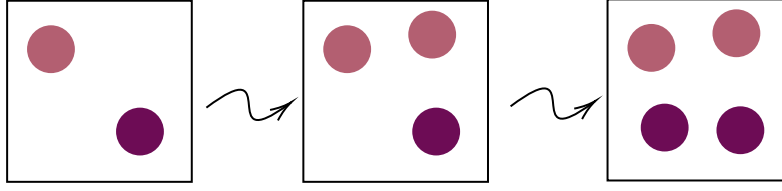
Now we set to obtain the upper bound, which will involve creating a unit flow  $f$  from  $a$  to  $z$ , and therefore by Thomson's principle, it will follow that  $R_{\text{eff}}(a, z) \leq \mathcal{E}(f)$ . Let's keep in mind *the diagram that says it all*



We are going to direct all edges  $e$  in the grid flowing out of  $a$  and into  $z$  and we will cleverly run a process  $X_t$  on the grid which will run from  $a$  to  $z$  in a very "symmetrical way", and then assign for an edge  $e$ ,  $f(e)$  to be the probability that the process goes through that edge. The process that we will run on the grid is Polya's urn:

**Polya's urn:** consider an urn that initially contains one white ball and one black ball, we choose one of the two balls at random, and then return it to the box, alongside a new ball of the same colour of the ball we just picked. The sequence of the number of white and black balls on the urn form a Markov process on the grid  $\{1, \dots, n\}^2$  with transition probabilities

$$P[(i, j), (i+1, j)] = \frac{i}{i+j} \quad P[(i, j), (i, j+1)] = \frac{j}{i+j}$$



We will run this process  $X_t = (B_t, W_t)$  on the grid, and stop once we have reached the main diagonal  $x + y = n + 1$  (highlighted in black in *the diagram*). We shall assign for an edge  $e$  in the lower diagonal region

$$f(e) = \mathbf{P}[X_t = e \text{ for some } t \geq 0]$$

To complete the construction of  $f$ , we reflect the values of  $f$  onto the upper diagonal, so that for example, the total flow out of  $a$  equals the total flow into  $z$  (as also shown in the diagram). We now claim the following:

**Claim:** for a given diagonal, i.e: the points  $(i, j)$  such that  $i + j = k + 2$  for some  $k \geq 1$  (notice that for illustration, the case  $k = 0$  corresponds to the two points coming out of  $a$ ), the probability of reaching any such points is the same. For illustration again, this means that the process is equally likely to pass through any of the points highlighted in dark green in the diagram. More precisely, we claim the probability to visit any point  $(i, j)$  such that  $i + j = k + 2$ , we have that

$$\mathbf{P}[X_t \text{ passes through } (i, j)] = \frac{1}{i + j - 1} = \frac{1}{k + 1}$$

**Proof of Claim:** We go by induction. For the base case, which corresponds to  $k = 1$ , i.e: the first diagonal after  $a$ , it is immediately clear that the probability that each  $(1, 2)$  or  $(2, 1)$  is visited is of  $1/2$  each. Assume now that the hypothesis holds for the diagonal  $i + j = (k - 1) + 2$ , let us show the claim also holds for the diagonal  $i + j = k + 2$ . Then the probability that the process goes through  $(i, j)$  is the probability that it passes through  $(i - 1, j)$  and then goes to  $(i, j)$  plus the probability that it passes through  $(i, j - 1)$  and then passes through  $(i, j)$ . Thus

$$\begin{aligned} \mathbf{P}[\text{process passes through } (i, j)] &= \frac{1}{(i - 1) + j - 1} P[(i - 1, j), (i, j)] + \frac{1}{i + (j - 1) - 1} P[(i, j - 1), (i, j)] \\ &= \frac{1}{i + j - 1} \end{aligned}$$

by direct computation.

Thus the claim is proven. Since the probability of visiting one vertex is exactly the sum of going through any of the edges that lead to said vertex, it follows that for any vertex on the diagonal  $i + j = k + 2$ , the sum of flows going into each vertex  $(i, j)$  is constant and equal to  $\frac{1}{k + 1}$ . Naturally,

the probability of going into one vertex is also the probability of leaving it, so the flow into any vertex equals the flow out of said vertex which means that this construction of  $f$  really is a unit flow from  $a$  to  $z$  (and just to be super clear, the flow out of  $a$  is 1 because the probability of exiting  $a$  is one given that the process starts there). We are now ready, using these tools to bound the energy of  $f$ . Labelling  $L$  for the lower diagonal region (and abusing notation for  $L$  being the edge set and the vertex set):

$$\begin{aligned}
 \mathcal{E}(f) &:= \sum_{e \in E} f(e)^2 r(e) \\
 &\stackrel{(1)}{=} \sum_{e \in E} f(e)^2 \\
 &\stackrel{(2)}{\leq} 2 \sum_{e \in L} f(e)^2 \\
 &\stackrel{(3)}{=} \sum_{x \in L} \sum_{e: e \text{ comes into } x} f(e)^2 \\
 &\stackrel{(4)}{\leq} 2 \sum_{x \in L} \left( \sum_{e: e \text{ comes into } x} f(e) \right)^2 \\
 &\stackrel{(5)}{=} 2 \sum_{k=1}^{n-1} (k+1) \frac{1}{(k+1)^2} \\
 &\stackrel{(6)}{\leq} 2 \log n
 \end{aligned}$$

Where (1) comes from the assumption of unit conductance, and hence unit resistance, (2) comes from the fact that to count all edges, we can count twice the ones on the lower region, thus double counting the main diagonal, hence the inequality, and since we are squaring  $f$ , we don't really care if the edges are directed and undirected to justify that this also takes care of the upper diagonal. Step (3) comes from the fact that to count all edges in the lower diagonal, it suffices to count all vertices and then the edges that go into each vertex, step (4) comes from the fact that  $a^2 + b^2 \leq (a + b)^2$  for positive  $a$  and  $b$  and we are summing positive flows. Step (5) comes from the fact that to count all edges, we can simply count over all edges on the diagonals  $i + j + k + 2$ , and note that for this  $k$  will run from 1, corresponding to the first diagonal, all the way to  $n - 1$ , corresponding to the main diagonal. Moreover, we have established that the sum of flows of edges going into any such vertex is constant and equal to  $\frac{1}{k+1}$ , and moreover, on any such diagonal there are  $k + 1$  vertices. Step (6) comes from once again the straightforward integral comparison of a sum.



We now give one last connection between electrical networks to properties of the random walk before moving to the next section. First let us state and prove the following classic result:

**Lemma 15.30 (Green's function and stationary distribution)** Let  $X$  be an irreducible Markov chain on a finite state space and let  $T$  be a positive stopping time of finite mean, such that  $\mathbf{P}_a(X_T = a) = 1$  for some state  $a$ . Then for all  $x$  we have that

$$G_T(a, x) = \pi(x) \mathbf{E}_a[T].$$

**Main idea:** The claim follows from the fact that if a function  $h$  on the state space satisfies  $hP = h$ , then  $h$  is a constant multiple of  $\pi$ .

*Proof.* We have the following observations:

- We can express  $G_T(a, x) := \mathbf{E}_a \left[ \sum_{t=0}^{T-1} \mathbf{1}\{X_t = x\} \right]$  as

$$\begin{aligned} \sum_{t \geq 0} \mathbf{P}_a(X_t = x, t < T) &= \underbrace{\mathbf{P}_a(X_0 = x, 0 < T)}_{\mathbf{1}\{x=a\}} + \sum_{t \geq 1} \mathbf{P}_a(X_t = x, t < T) \\ &= \mathbf{1}\{x = a\} + \sum_{t \geq 1} \mathbf{P}_a(X_t = x, t \leq T) - \mathbf{P}_a(X_t = x, t = T) \end{aligned}$$

but  $\sum_{t \geq 1} \mathbf{P}_a(X_t = x, T = t) = \mathbf{P}_a(X_T = x) = \mathbf{1}\{x = a\}$ . Therefore, Green's function can also be expressed as  $\sum_{t \geq 1} \mathbf{P}_a(X_t = x, t \leq T)$ .

- If we sum over the previous step when  $t \geq 1$  and use the Markov property and the fact that the event  $\{t \leq T\}$  is measurable with respect to  $X_1, \dots, X_{t-1}$ :

$$\mathbf{P}_a(X_t = x, t \leq T) = \sum_y \mathbf{P}_a(X_t = x, X_{t-1} = y, t \leq T) = \sum_y \mathbf{P}_a(X_{t-1} = y, t \leq T) P(y, x).$$

Combining the two observations,

$$G_T(a, x) = \sum_{t \geq 1} \mathbf{P}_a(X_t = x, t \leq T) = \sum_{t \geq 1} \sum_y \mathbf{P}_a(X_{t-1} = y, t \leq T) P(y, x),$$

and now swapping the two summations and relabelling the indices, we are indeed left with  $G_T(a, x) = \sum_y G_T(a, y) P(y, x)$ . This means that  $G_T(a, \cdot)$  is a constant multiple  $C\pi(\cdot)$ . To see the value of this constant, we simply sum over all the state space:

$$C = \sum_x G_T(a, x) = \sum_{t \geq 0} \sum_x \mathbf{P}_a(X_t = x, t \leq T) = \mathbf{E}_a[T],$$

as required. ♡

We now have the following connection between effective resistance and commute time:

**Proposition 15.31 (Commute time identity)** Let  $T_{a,b}$  be the commute time from  $b$  to  $a$ , that is to say, the first time the chain returns to  $a$  after having visited  $b$ . Formally:  $T_{a,b} = \inf\{t > 0 : X_t = a \text{ and } X_u = b \text{ for some } u \in \{1, \dots, t-1\}\}$ . Then

$$\mathbf{E}_a[T_{a,b}] = c(G)R_{\text{eff}}(a, b).$$

*Proof.* We first note that the stopping time  $T_{a,b}$  satisfies  $\mathbf{P}_a(X_{T_{a,b}} = a) = 1$ . Then, we also note that

$$G_{T_{a,b}}(a, a) = G_{T_b}(a, a).$$

This is because if we think of  $T_{a,b}$ , once  $b$  has been visited,  $a$  will by definition not be visited until the stopping time rings. Since the time when the stopping time rings is not counted in the sum defining Green's function, we therefore see that the expected number of visits to  $a$  before  $T_{a,b}$  rings is the exact same as the expected number of visits to  $a$  before  $T_b$  rings. However, by the connection between Green's function and effective resistance, we have that  $G_{T_b}(a, a) = c(a)R_{\text{eff}}(a, b)$ . Finally, we can now combine with the previous Lemma, and since  $\pi(a) = \frac{c(a)}{c(G)}$ , we are done. ♡

## 15.3 Cover times

We reach the final section of these notes, where we will use the theory of electric networks to obtain bounds on the cover times.

**Definition 15.32 (Cover times and hitting times)** We define the **maximal hitting time**  $t_{\text{hit}}$ , as

$$t_{\text{hit}} = \max_{x,y} \mathbf{E}_x[T_y],$$

and the **cover time**,  $t_{\text{cov}}$  as

$$t_{\text{cov}} = \max_x \mathbf{E}[T_{\text{cov}}],$$

where  $T_{\text{cov}}$  is the first time that the Markov chain has visited every state in the state space.

We now introduce a preliminary relationship between  $t_{\text{hit}}$  and  $t_{\text{cov}}$ :



**Theorem 15.33** (Relationship between  $t_{\text{hit}}$  and  $t_{\text{cov}}$  V1) Let  $n$  be the size of the state space, then

$$t_{\text{hit}} \leq t_{\text{cov}} \leq t_{\text{hit}} \left( 1 + \frac{1}{2} + \frac{1}{3} + \cdots + \frac{1}{n-1} \right)$$

*Proof.* The lower bound is obvious, since for all  $y$ ,  $T_y \leq T_{\text{cov}}$ . For the upper bound we will proceed as follows: suppose without loss of generality that our state space is  $\{1, \dots, n\}$ , and that the chain starts at  $n$ . Let  $\sigma$  be a permutation of  $\{1, \dots, n-1\}$  that is chosen uniformly at random from the group  $\mathfrak{S}_{n-1}$ . We will look to cover states in order  $\sigma(1)$ , then  $\sigma(2)$ , etc. In particular, let  $\tau_k$  be the first time that the states  $\sigma(1), \dots, \sigma(k)$  have all been visited, and let  $L_k = X_{\tau_k}$  be the position of the chain at the time that the states  $\sigma(1), \dots, \sigma(k)$  were all first visited. We make the following observations:

- If  $L_k \neq \sigma(k)$ , meaning: the position of the chain at the time at which we finished visiting every state in  $\{\sigma(1), \dots, \sigma(k)\}$  was not  $\sigma(k)$  (so it was one of  $\sigma(1), \dots, \sigma(k-1)$ ), then  $L_k$  is in fact equal to  $L_{k-1}$ , because the time at which we first visited every state in  $\{\sigma(1), \dots, \sigma(k)\}$  is also the first time we visited every state in  $\{\sigma(1), \dots, \sigma(k-1)\}$ .
- If  $L_k = \sigma(k) = r$ , and  $L_{k-1} = s$ , then the difference  $\tau_k - \tau_{k-1}$  is precisely the hitting time of state  $r$  when starting from  $s$ . Hence:

$$\mathbf{E}_n[\tau_k - \tau_{k-1} | \{\sigma(k) = L_k = r\} \cap \{L_{k-1} = s\}] = \mathbf{E}_s[T_r] \leq t_{\text{hit}}.$$

Averaging over  $r$  and  $s$  gives that

$$\mathbf{E}[\tau_k - \tau_{k-1} | \{\sigma(k) = L_k\}] \leq t_{\text{hit}}.$$

- Finally, we note that since  $\sigma$  is chosen uniformly at random, the probability that  $\sigma(k)$  is the last state to be visited in  $\{\sigma(1), \dots, \sigma(k)\}$  before they are all visited is precisely  $1/k$ .

Now we combine all this:

$$\begin{aligned} t_{\text{cov}} &\leq \mathbf{E}_n[\tau_{n-1}] = \sum_{i=1}^{n-1} \mathbf{E}_n[\tau_i - \tau_{i-1}] \\ &= \sum_{i=1}^{n-1} \mathbf{E}_n[\tau_i - \tau_{i-1} | \{\sigma(i) = L_i\}] \mathbf{P}_n(\sigma(i) = L_i) + \mathbf{E}_n[\tau_i - \tau_{i-1} | \{\sigma(i) \neq L_i\}] \mathbf{P}_n(\sigma(i) \neq L_i) \\ &\leq \sum_{i=1}^{n-1} \frac{t_{\text{hit}}}{i}. \end{aligned}$$



In a very similar spirit to the above, we have the following (possibly) stronger lower bound on  $t_{\text{cov}}$ :

**Theorem 15.34** (Relationship between  $t_{\text{hit}}$  and  $t_{\text{cov}}$  V2) For a set  $A \subseteq \Omega$  of states, define  $t_{\min}^A := \min_{x, y \in A: x \neq y} \mathbf{E}_x[T_y]$ . Then

$$t_{\text{cov}} \geq \max_{A \subseteq \Omega} t_{\min}^A \left( 1 + \frac{1}{2} + \cdots + \frac{1}{|A|-1} \right).$$

*Proof.* The proof strategy is essentially the same as before. We enumerate the states in  $A$  as  $\{1, 2, \dots, |A|\}$ , and using the same notation as before, choosing a uniformly at random permutation  $\sigma$  of these elements, if we let  $\tau_k$  be the first time  $\sigma(1), \dots, \sigma(k)$  have been visited, we will have that:

- Trivially  $t_{\text{cov}} \geq t_{\min}^A$ .
- If  $L_k = \sigma(k) = r$ , and  $L_{k-1} = s$ , then the difference  $\tau_k - \tau_{k-1}$  is precisely the hitting time of state  $r$  when starting from  $s$ . Hence:

$$\mathbf{E}_n[\tau_k - \tau_{k-1} | \{\sigma(k) = L_k = r\} \cap \{L_{k-1} = s\}] = \mathbf{E}_s[T_r].$$

However, this time instead of upper bounding  $\mathbf{E}_s[T_r]$ , we can lower bound it by  $t_{\min}^A$ .

The rest of the proof is identical. ♡

We now see the relationship between the hitting times and electrical networks:

**Theorem 15.35** (Hitting times and effective resistance) We have that

$$\frac{c(G)}{2} \max_{a,b} R_{\text{eff}}(a, b) \leq t_{\text{hit}} \leq c(G) \max_{a,b} R_{\text{eff}}(a, b)$$

*Proof.* The proof is quite simple and follows immediately from the commute time identity. Recall that the commute time identity established that  $\mathbf{E}_x[T_{x,y}] = c(G) R_{\text{eff}}(a, b)$ . However since  $\mathbf{E}_x[T_{x,y}]$  is just the expected time taken to travel from  $x$  to  $y$ , and then from  $y$  to  $x$ , by the Strong Markov Property, the Markov chain starts afresh when it reaches  $y$ , and so  $\mathbf{E}_x[T_{x,y}] = \mathbf{E}_x[T_y] + \mathbf{E}_y[T_x]$ . From this we have the following two observations:

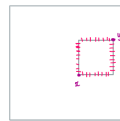
1.  $\mathbf{E}_x[T_{x,y}] \leq 2(\mathbf{E}_x[T_y] \vee \mathbf{E}_y[T_x])$ , and so by taking maximums  $c(G) \max_{a,b} R_{\text{eff}}(a, b) \leq 2 t_{\text{hit}}$ .
2. Clearly  $\mathbf{E}_x[T_y] \leq \mathbf{E}_x[T_{x,y}] = c(G) R_{\text{eff}}(x, y)$ , and so by taking maximums,  $t_{\text{hit}} \leq \max_{a,b} R_{\text{eff}}(a, b)$ .

♡

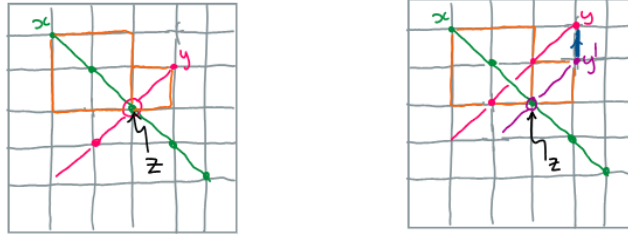
**Example 15.36 (Hitting and cover times on the grid)** We finish these notes by showing how to combine these results to study the hitting and cover times on the grid.

*Solution.* We recall from previous discussions that the effective resistance between the bottom left and top right corners on the grid  $\{1, \dots, n\}^2$  was order  $\log(n)$ . Let us begin by obtaining the order of the maximal hitting time:

- Lower bound: from the previous Theorem we know that  $t_{\text{hit}} \gtrsim c(G) \max_{a,b} R_{\text{eff}}(a, b)$ . Since the conductances are assumed to be one for each edge, we get that  $c(G) \asymp n^2$ , and clearly  $\max_{a,b} R_{\text{eff}}(a, b)$  is lower bounded by the effective resistance between the bottom left and top right corners, from which we get that  $t_{\text{hit}} \gtrsim n^2 \log(n)$ .
- For an upper bound, we similarly just need to show that  $\max_{a,b} R_{\text{eff}}(a, b) \lesssim \log(n)$ . For this we make the following considerations:
  - If  $x$  and  $y$  are two points on the grid that “form a square”, then we can think of “isolating” the square with corners  $x$  and  $y$  from the rest of the grid by removing all edges that connect it, that is to say, setting the resistances of the outgoing edges to be  $\infty$ . Then this new square will itself be a grid of side-length  $|x - y|$ , and so in this new grid, the effective resistance between  $x$  and  $y$  will be order  $\log(|x - y|) \leq \log(n)$ . Now we can note that by the monotonicity principle, when thinking about the original grid, we have “added some edges” (i.e: reduced the resistance from  $\infty$  to 1), and we will have that in the original grid, the effective resistance between  $x$  and  $y$  will be upper bounded up to constants by  $\log(n)$ :



- If  $x$  and  $y$  don't form the corners of a square, we will use the triangle inequality of effective resistance and “put squares in between”. We can consider the lines of slope  $\pm 1$  emanating from  $x$  and  $y$ . We have two cases: if one of these pairs of lines intersect at an integer coordinate, then we have found the “intermediate square”, say  $z$ , and then  $R_{\text{eff}}(x, y) \leq R_{\text{eff}}(x, z) + R_{\text{eff}}(z, y)$ , and since now each of these are points that make the corners of squares, we get that  $R_{\text{eff}}(x, y) \lesssim \log(n)$  (left picture):



On the other hand, if the intersection occurs at a half-integer coordinate, then we can consider another point  $y'$  which is moved by one unit so that the intersection occurs at an integer coordinate, use the square trick, and then simply note that  $R_{\text{eff}}(y, y') = 1$  since they are neighbours (right picture above).

For the cover times:

- The upper bound follows from the fact that  $t_{\text{cov}} \lesssim t_{\text{hit}} \left(1 + \frac{1}{2} + \dots + \frac{1}{n-1}\right) \lesssim \log(n)^2 n^2$ .
- For the lower bound we consider the set  $A = \{(k\lfloor\sqrt{n}\rfloor, k\lfloor\sqrt{n}\rfloor), ((k+1)\lfloor\sqrt{n}\rfloor, (k+1)\lfloor\sqrt{n}\rfloor)\}$ . Then  $t_{\text{cov}} \gtrsim \log(n) \min_{x \neq y: y, x \in A} \mathbf{E}_x[T_y]$ , but by the commute time identity,  $\min \mathbf{E}_x[T_y] \gtrsim c(G) \min_{x, y \in A, x \neq y} R_{\text{eff}}(x, y)$ . This minimum effective resistance will be the effective resistance between the bottom left and top right corner of a box of length  $\sqrt{n}$ , and so this effective resistance will also be order  $\log(n)$ . Therefore  $t_{\text{cov}} \gtrsim n^2 \log(n)^2$ .

♡

**Example 15.37 (Hitting and cover times on a tree)** Just kidding here is one last example, let us find the hitting and cover times for a binary tree of depth  $k$ .

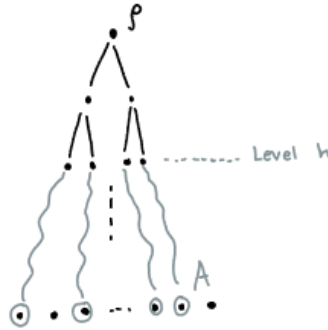
*Hitting times.* It is clear that the maximal hitting time will be achieved by any two pairs of leaves  $a$  and  $b$  whose common ancestor is the root  $\rho$ . By symmetry, the hitting time  $\mathbf{E}_a[T_b]$  is precisely equal to the commute time from the root to any of them, say  $\mathbf{E}_\rho[T_{\rho,a}]$ . By the commute time identity, this is equal to  $c(G)R_{\text{eff}}(\rho, a)$ . As we computed before, the effective resistance in a tree is simply graph distance, so  $R_{\text{eff}}(\rho, a) = k$ . Now to compute  $c(G)$ , we simply note that since each vertex has degree two,  $c(G)$  will be twice the number of edges. A quick computation shows that  $c(G) = 2(n-1)$ , where  $n = 2^{k+1} - 1$  is the number of vertices.

♡

*Cover times.* To bound the cover times, we see that one the one hand:

- Since  $t_{\text{cov}} \leq t_{\text{hit}} \left(1 + \frac{1}{2} + \dots + \frac{1}{n}\right)$ , this will give  $t_{\text{cov}} \leq 2(n-1)k \left(1 + \frac{1}{2} + \dots + \frac{1}{n}\right)$ .
- For a lower bound, we need to use the bound  $t_{\text{cov}} \geq \max_A t_{\text{min}}^A$ . Let us define the families  $A_h$

where  $h \in \{1, \dots, k\}$ , where  $A_h$  is a set of  $2^h$  leaves, such that each vertex at level  $h$  has a unique descendant in  $A$ :



Recall that  $t_{\min}^A = \min_{x \neq y \in A} \mathbf{E}_x[T_y]$ . Clearly this minimum hitting time is achieved by leaves  $a$  and  $b$  in  $A$  whose most recent common ancestor is at level  $h' < h$ . Once again, by symmetry and the commute time identity, this hitting time will be equal to  $2(n-1)(k-h')$ . We can now minimise this by cleverly choosing the leaves in  $A$  so that  $h' = h-1$ , which gives that for any  $h = 1, \dots, k$

$$t_{\text{cov}} \geq 2(n-1)(k-h+1) \left( 1 + \frac{1}{2} + \dots + \frac{1}{2^{h-1}} \right) = (2+o(1)) \log(2)n(k-h)h.$$

We now minimise this with respect to  $h$  by setting  $h = \lfloor k/2 \rfloor$ .



Dear reader, thank you for reaching the end of the notes. They took me an entire year to write, after many long days in the library and some unhealthy habits. Hopefully they aided your studies.  
Yours falsely,  
JOF



# Appendix A

## Product Chains

Sometimes a lot of examples are of something called product chains, where we take the literal product of several chains, and then move on this new bigger chain by selecting one of the sub-chains, and moving it according to its transition matrix.

**Definition A.1 (Product chain)** For  $j = 1, \dots, d$ , let  $P_j$  be an irreducible transition matrix on the state space  $\Omega_j$ . Consider the chain on  $\Omega = \prod_{i=1}^d \Omega_i$ , which moves by selecting one of the  $d$  coordinates, say  $\Omega_i$  at random, and then only the  $i^{th}$  coordinate according to  $P_i$ . The transition matrix is therefore, for two vectors  $\mathbf{x}, \mathbf{y}$ :

$$\tilde{P}(\mathbf{x}, \mathbf{y}) = \frac{1}{d} \sum_{j=1}^d P_j(x_j, y_j) \prod_{i:i \neq j} \mathbf{1}(x_i = y_i)$$

And here's the main result we are concerned with:

**Lemma A.2 (Eigenvalues)** Suppose that for each  $j = 1, 2, \dots, d$  the transition matrix  $P_j$  on state space  $\Omega_j$  has an eigenfunction  $\phi^{(j)}$  with eigenvalue  $\lambda^{(j)}$ . Then the function  $\phi = \prod_j \phi^{(j)}$  is an eigenfunction of the product chain  $\tilde{P}$  with eigenvalue  $\frac{1}{d} \sum_{j=1}^d \lambda^{(j)}$ .

**Lemma A.3 (Orthonormal Basis)** Let  $\pi_1, \dots, \pi_d$  be the invariant distributions of the chains  $P_1, \dots, P_d$ . Then  $\pi = \prod_j \pi_j$  is the invariant distribution of the product chain, and if the set  $B_j$  is an orthonormal basis for  $\ell(\pi_j)$ , i.e: an orthonormal basis for  $\mathbf{R}^{\Omega_j}$  with respect to the inner product with  $\pi_j$ , then

$$\left\{ \prod_j \phi^{(j)} : \phi^{(j)} \in B_j \right\}$$

is a basis for  $\ell(\pi)$ . In practical terms, the set of all products of eigenvectors of the individual chains is an orthonormal basis for the big chain.



# Appendix B

## Induced Chains

**Definition B.1** (Induced Chains ) Let  $X$  be a reversible Markov chain and  $A \subseteq \Omega$  a subset of state-space. Let  $T_{A^+}$  be the first return time to  $A$ . The induced chain on  $A$  is the Markov chain with state space  $A$  and transition matrix

$$P_A(x, y) = \mathbf{P}_x[X_{T_{A^+}} = y]$$

That is to say, the transition probability from  $x$  to  $y$  is the probability that the original chain started at  $x$  first reaches  $y$  when returning to  $A$ .

This corresponds to observing the original chain only during the times that it is at  $A$

**Theorem B.2** (Example Sheet 3, Q1) Let  $\pi$  be the invariant distribution of the original chain. Then the induced chain is reversible with respect to  $\pi_A = \pi(x)\mathbf{1}(x \in A)/\pi(A)$ , and hence  $\pi_A$  is the invariant distribution for the induced chain. Moreover, let  $\gamma_A$  be the spectral gap of the induced chain and  $\gamma$  be the original spectral gap. Then  $\gamma_A \geq \gamma$ .



# Appendix C

## Classic Markov Chains

Here we discuss some Markov Chains which are useful to know results about:

**Example C.1 (Gambler's Ruin)** Consider a gambler betting on the outcome of independent fair coin tosses. If she gets heads, her fortune increases by 1, otherwise it decreases by 1. The gambler will stop once she goes bust or her fortune reaches  $n$ . Starting from a wealth of  $k \in \{0, \dots, n\}$ , define  $\tau$  to be the time at which she stops gambling. Then:

- $\mathbf{P}_k(X_\tau = n) = k/n$ .
- $\mathbf{E}_k(\tau) = k(n - k)$ .

**Example C.2 (Coupon collector)** A company issues  $n$  different types of coupons. A collector needs all  $n$  types to win. Suppose any coupon collected is equally likely, i.e. with probability  $1/n$ . Let  $\tau$  be the number of coupons obtained until he obtains all  $n$  types. Then

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

and for any  $c > 0$

$$\mathbf{P}(\tau > \lceil n \log n + c n \rceil) \leq \exp(-c).$$



# Appendix D

## Background Results

**Theorem D.1** (Recurrence is a class property) Let  $(X_n)$  be an irreducible chain, then one state is recurrent if and only if all states are recurrent.

*Proof.* Let  $i$  and  $j$  be two states. Since  $(X_n)$  is irreducible,  $P^n(i, j) > 0$  and  $P^m(j, i) > 0$  for some  $n, m$  large enough. Then we have that for all  $r \geq 0$ :

$$P^{n+m+r}(i, i) \geq P^n(i, j)P^r(j, j)P^m(j, i)$$

Since

$$\sum_{r=0}^{\infty} P^r(i, i) \geq \sum_{r=0}^{\infty} P^{n+m+r}(i, i) \geq P^n(i, j)P^m(j, i) \sum_{r=0}^{\infty} P^r(j, j)$$

We have that if  $j$  is recurrent, then this last sum is infinite, and as such  $i$  is also recurrent. ♡

**Theorem D.2** An irreducible Markov Chain on a finite state space is positive recurrent.

*Proof.* First we note that

$$\mathbf{E}_x[T_x^+] = \sum_{t=0}^{\infty} \mathbf{P}_x(T_x^+ > t) = \sum_{t=0}^{\infty} \mathbf{P}_x\left(\bigcap_{k=1}^t \{X_k \neq x\}\right)$$

Now observe that since the chain is irreducible, given any pair  $(i, j)$  of states we have some integer  $n_{ij}$  and positive real number  $\epsilon_{ij}$  such that  $P^{n_{ij}}(i, j) = \epsilon_{ij} > 0$ . Thus define the quantities

$$n := \max_{i,j} n_{ij} \quad \epsilon := \min_{i,j} \epsilon_{ij}$$

It is clear that

$$\mathbf{P}_x \left( \bigcup_{k=1}^n \{X_k = x\} \right) \geq \epsilon$$

Or in other words

$$\mathbf{P}_x \left( \bigcap_{k=1}^n \{X_k \neq x\} \right) \leq 1 - \epsilon$$

By the Markov Property we can also note that for any  $m \in \mathbf{N}$

$$\mathbf{P}_x \left( \bigcap_{k=1}^{mn} \{X_k \neq x\} \right) \leq (1 - \epsilon) \mathbf{P}_x \left( \bigcap_{k=1}^{(m-1)n} \{X_k \neq x\} \right) \leq \dots \leq (1 - \epsilon)^m$$

Combining this with our initial observation:

$$\mathbf{E}_x[T_x^+] = \sum_{t=0}^{\infty} \mathbf{P}_x \left( \bigcap_{k=1}^t \{X_k \neq x\} \right) \leq \sum_{t=0}^{\infty} \mathbf{P}_x \left( \bigcap_{k=1}^{nt} \{X_k \neq x\} \right) \leq \sum_{t=0}^{\infty} (1 - \epsilon)^t < \infty$$

♡

**Theorem D.3 (Distribution of visits)** Let  $(X_n)$  be a Markov Chain, and let  $x$  be a transient state. Suppose  $X_0 = x$ . Let

$$V_x = \sum_{t=0}^{\infty} \mathbf{1}_{\{X_t = x\}}$$

denote the number of visits to  $x$ . Then the distribution of  $V_x$  is geometric.

*Proof.* The proof involves a few prior constructions which we do not have the time to transcribe here. See [Nor97, Pages 24-25]

♡

**Theorem D.4 (Strong Markov Property)** Let  $(X_n : n \geq 0)$  be a Markov chain with initial distribution  $\lambda$  and transition matrix  $P$ . Let  $T$  be a stopping time for  $(X_n : n \geq 0)$ , then conditional on  $T < \infty$  and  $X_T = \omega$ ,  $(X_{T+n} : n \geq 0)$  is a Markov chain with initial distribution  $\delta_\omega$  and is independent of  $X_0, X_1, \dots, X_T$ .

**Theorem D.5 (Existence of  $\pi$ )** Let  $(X_n)$  be an irreducible, positive recurrent Markov Chain, then  $(X_n)$  admits an invariant distribution  $\pi$ , and moreover

$$\pi(x) = \frac{1}{\mathbf{E}_x[T_x^+]}$$

*Proof.*



**Lemma D.6 (Number Theoretic Lemma)** Let  $S \subseteq \mathbf{N} \cup \{0\}$  be closed under addition and  $\gcd(S) = 1$ . Then there exists an  $M \in \mathbf{N} \cup \{0\}$  such that whenever  $a \geq M$ , then  $a \in S$ .

*Proof.* See [Fre, Lemma 4.4]



**Theorem D.7 (Alternative interpretation of aperiodicity)** Let  $(X_n : n \geq 0)$  be an irreducible aperiodic Markov Chain on a finite state space. Then for a time  $t$  large enough, given any two states  $x$  and  $y$ , we have that  $P^t(x, y) > 0$ .

*Proof.* For any  $x \in \Omega$  we have that  $T = \{t > 0 : P^t(x, x) > 0\}$  has greatest common divisor of 1 and is closed under addition. Indeed, if  $a, b \in T$  then

$$P^{a+b}(x, x) \geq P^a(x, x)P^b(x, x) > 0$$

Therefore by Lemma D.6 there is some integer  $M_x$  such that whenever  $t \geq M_x$ , then  $P^t(x, x) > 0$ . Define  $M = \max\{M_x : x \in \Omega\}$  which exists due to finiteness of  $\Omega$ . Then it follows that for all  $t \geq M$ , we have that  $P^t(x, x) > 0$  for all  $x \in \Omega$ . Now we have that if  $x$  and  $y$  are two states, due to irreducibility of the Chain, some integers  $m$  and  $n$  such that  $P^m(x, y)$  and  $P^n(y, x)$  are both strictly greater than zero. Due to finiteness, we can pick some  $N$  that is greater than all of these  $m$  and  $n$ . Thus now for any  $t > 2M + N$ , i.e:  $t = 2M + N + r$  we have that

$$P^t(x, y) \geq P^m(x, k)P^{M-m-n+N+r}(k, k)P^n(k, y) > 0$$



**Lemma D.8** Let  $A$  be a stochastic matrix and let  $B$  be a matrix whose rows are all the row vector  $v$ . Then  $AB = B$ .

*Proof.* Simple computation:

$$(AB)(x, y) = \sum_i A(x, i)B(i, y) = \sum_i A(x, i)v(y) = v(y) \sum_i A(x, i) = v(y) = B(x, y)$$

where this last equality comes from the fact that  $A$  is stochastic.



**I**



# Index

- $E$ -Path, 98
- $\mathcal{L}^p$  mixing times, 62
- Coupling of Markov Chains, 33
- Dirichlet form, 89
- Dirty bound, 29
- Distance of states, 143
- Expander graph, 118
- Induced Chain, 217
- Invariant distribution, 9
- LSRW on Box, 100
- Modified Cycle, 134
- Poincaré constant, 92
- Pre-cutoff, 77
- Product condition, 77
- Random walk on  $\mathbf{Z}_n$ , 36
- Separation distance, 44
- Spectral representation of the spectral profile, 132
- Transitive chain, 81
- Wilson's Method, 84



# Bibliography

- [Fre] Freedman. url: <https://math.uchicago.edu/~may/REU2017/REUPapers/Freedman.pdf>.
- [Nor97] J.R Norris. *Markov Chains*. Cambridge University Press, 1997.