MARKOV CHAINS, MIXING TIMES, AND CUTOFF

Jackie Lok

Supervisor: Prof. Catherine Greenhill

School of Mathematics and Statistics
UNSW Sydney

November 2020

Submitted in partial fulfillment of the requirements of the degree of
Bachelor of Science with Honours

Plagiarism statement

I declare that this thesis is my own work, except where acknowledged, and has not been submitted for academic credit elsewhere.

I acknowledge that the assessor of this thesis may, for the purpose of assessing it:

- Reproduce it and provide a copy to another member of the University; and/or,
- Communicate a copy of it to a plagiarism checking service (which may then retain a copy of it on its database for the purpose of future plagiarism checking).

I certify that I have read and understood the University Rules in respect of Student Academic Misconduct, and am aware of any potential plagiarism penalties which may apply.

By signing this declaration I am agreeing to the statements and conditions above.

Signed: Jackie Lok Date: 20/11/2020

Acknowledgements

Firstly, I would like to express my deepest thanks to my supervisor Catherine Greenhill, who has suggested a most interesting topic, and provided much fruitful guidance and support throughout the year on good mathematics and style. I have also benefited from the many excellent teachers that I have been fortunate to have had. In particular, I would like to thank David Angell for introducing me to the wonders of linear algebra. I would also like to thank Pinhas Grossman and Daniel Chan, who showed that doing rigorous mathematics could still be fun. Thanks also go to Ian Doust and Denis Potapov, for providing a good introduction to university-level mathematics.

Furthermore, I would like to express my gratitude to my parents, as well as my grandma, for providing a comfortable environment, especially during this particularly challenging year. Thanks also go to Kelvin. Finally, I would like to express my special gratitude to Jessica for her constant encouragement and support.

Jackie Lok, November 2020.

Abstract

Understanding the rate of convergence to stationarity, or mixing time, of a Markov chain is of interest for problems from shuffling cards, to providing rigorous bounds for the runtime of Monte Carlo algorithms. In this thesis, we will provide a detailed survey of some selected probabilistic and analytical techniques that can be used to bound the mixing times of discrete Markov chains.

We will illustrate an application of rapidly mixing Markov chains by developing an efficient randomised algorithm for approximately counting graph colourings. Furthermore, we will discuss the cutoff phenomenon, which describes how mixing occurs very abruptly at, and not before, a precise point for certain Markov chains. Finally, we will culminate with a detailed study of the Glauber dynamics for the mean-field Ising model, which exhibits distinct mathematical behaviour at different temperature parametrisations, including cutoff, rapid mixing, and torpid mixing.

Contents

Chapter 1

Introduction

Markov chains, introduced by the eponymous Markov [44], are a class of stochastic processes for which the evolution only depends on the current state. Many classical results in this area have been established (e.g. see Feller [24]), including the fundamental result that a Markov chain will converge towards a stationary distribution under certain conditions.

Markov chains are a natural basis for modelling many real-world processes. For example, shuffling a pack of cards can be viewed as a random walk on the symmetric group S_n, which converges to the uniform distribution (under certain constraints). The Markov chain Monte Carlo (MCMC) method, where random sampling is used to solve computationally intractable problems, also relies upon running a Markov chain until convergence in order to draw (approximate) samples from a particular stationary distribution.

Motivated by these applications, we can ask some simple questions:

(1) How many shuffles does it take to mix up a pack of cards?

(2) How long should a MCMC sampler be run so that the samples drawn are acceptably "close" to the stationary distribution?

Answering these questions requires the analysis of the rate of convergence towards stationarity, or mixing time, of a Markov chain, which is the focus of this thesis. The exact distribution of a chain at time t seldom has a tractable expression, but the distance to stationarity can often be bounded. Very roughly, we want to understand whether a given chain "mixes rapidly", or otherwise "mixes torpidly" (i.e. slowly).

The study of the mixing times of finite Markov chains began in the early 1980s (see [1, 2] for some of the early papers by Aldous and Diaconis). This has grown into a burgeoning research area, and we can briefly mention some selected results. In the study of random walks on finite groups, Bayer and Diaconis [6] prove that seven riffle shuffles is sufficient to mix up a pack of cards. In theoretical computer science, provably rapidly mixing Markov chains for sampling combinatorial objects are the basis of efficient randomised algorithms [20, 31, 34]. The efficiency and mathematical behaviour of chains used to sample from models in statistical physics have also been studied [32, 37, 42].

The excellent expository article of Diaconis [14], aptly named "The Markov Chain Monte Carlo Revolution", describes how the analysis of mixing times is not only of practical interest, but has also lead to (and leans on) fascinating mathematics from various fields. This includes techniques from probability (coupling and stopping times), analysis (spectral methods and functional inequalities), algebra (group representations and algebraic geometry), and combinatorics.

Furthermore, an interesting phenomenon known as cutoff has also been observed for certain Markov chains, which describes how the transition to stationarity occurs very abruptly at, and not before, a precise point. Since the first example where a sharp cutoff was demonstrated [16], it has been proved to occur in many more different classes of chains, which suggests that it may be quite a general phenomenon.

Structure of thesis.

The aim of this thesis is to survey some of the mathematical techniques and problems related to the analysis of mixing times of Markov chains. We will focus on discrete time Markov chains with finite state spaces. The topics discussed are mainly in the field of discrete probability, with some connections to linear algebra, functional analysis, and combinatorics. The structure of this thesis is illustrated in Figure 1.0.1.

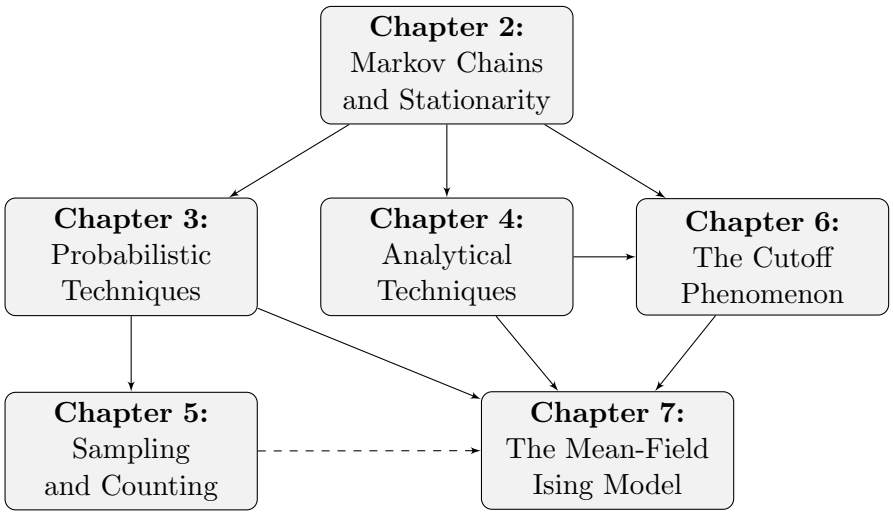

Figure 1.0.1: The arrows indicate the logical flow of the topics covered in the thesis.

The remainder of the introduction will establish the notation that will be used, as well as the assumed knowledge. In Chapter 2, we will state and prove some classical results on the stationary distributions of Markov chains, and precisely define mixing times. Next, we will provide a detailed survey of some probabilistic methods (Chapter 3) and analytical methods (Chapter 4) that can be used to analyse mixing times. The development of the theory will closely follow the textbook of Levin, Peres, and Wilmer [38], which provides an excellent introduction to this area of research.

Using the mathematical tools developed so far, the remainder of this thesis will examine some problems involving the analysis of mixing times. In Chapter 5, we will discuss how rapidly mixing Markov chains can be used as the basis of efficient randomised algorithms for "difficult" counting problems. We will prove that a rapidly mixing chain exists for sampling proper q-colourings of a graph of maximum degree Δ, when $q \geq 2\Delta + 1$.

In Chapter 6, we will provide and prove the equivalence of various precise characterisations of cutoff seen in the literature, which will require delicate control of mixing times. In Chapter 7, this thesis will culminate with a detailed analysis of the Glauber

dynamics for the mean-field Ising model from statistical physics. Following the paper of Levin, Luczak, and Peres [37], we will prove that this chain exhibits distinct mathematical behaviour at different temperature parametrisations, including cutoff, rapid mixing, and torpid mixing. Finally, we will finish with some concluding remarks in Chapter 8.

Contributions.

Although this is a survey of known results, I have endeavoured to provide original proofs and examples where possible.

- In Chapter 6, I provided original proofs for the equivalences of the various characterisations of cutoff, as I have not seen such results in the literature.
- Rather than consider the standard Metropolis chain in Chapter 5, I obtained similar results for an alternative Markov chain, known as the heat-bath Glauber dynamics.

Otherwise, if a good proof already exists, I have provided additional clarifying and expository details in several places, including the proof of the ergodic theorem (Theorem 2.17) in Chapter 2, and the proof of the path coupling theorem (Theorem 3.12) in Chapter 3.

I have also selected nice examples for the theory from research papers, such as the inverse riffle shuffle for strong stationary times from [2] (in Chapter 3), and the riffle shuffle for cutoff [6] (in Chapter 6). For the major result on the Glauber dynamics for the mean-field Ising model in Chapter 7, I have attempted to give a concise explanation of the argument of [37]. I have used ideas from [18] to modify some parts of the argument (e.g. Lemmas 7.6 and 7.13) that seems to improve the proof.

Assumed knowledge.

We will assume a basic background in linear algebra, including familiarity with concepts such as matrices, eigenvalues, etc. (for example, refer to [27]). We will also assume familiarity with concepts from probability theory, such as discrete probability spaces, properties of the probability measure, random variables (independence, distribution, expected value, variance, etc.), conditional probability, and discrete stochastic processes (for example, see any introductory textbook such as [24, 26]).

The following inequalities relating the first and second moments of a random variable to its tail probabilities will be particularly important.

Theorem 1.1 (Markov's inequality ([26, p.311]))**.** *Let X be a random variable with finite mean $\mathbb{E}[X]$. Then for any $a > 0$,*

$$\mathbb{P}(|X| \geq a) \leq \frac{\mathbb{E}[|X|]}{a}. \tag{1.1}$$

Theorem 1.2 (Chebyshev's inequality ([26, p.319]))**.** *Let X be a random variable with finite mean $\mathbb{E}[X]$ and variance $\mathrm{Var}(X)$. Then for any $a > 0$,*

$$\mathbb{P}(|X - \mu| \geq a) \leq \frac{\mathrm{Var}(X)}{a}. \tag{1.2}$$

Notation.
We write $\mathbb{N} = \{0, 1, 2, 3, \dots\}$ and $\mathbb{Z}_+ = \{1, 2, 3, \dots\}$. Given $x \in \mathbb{R}$, we denote the floor (resp. ceiling) of x by $\lfloor x \rfloor$ (resp. $\lceil x \rceil$), which is x rounded down (resp. up) to the nearest integer. Floors and ceilings will typically be used with real arguments for functions with integer-valued domains, however may sometimes be omitted for notational simplicity.

The basic object of study will be a discrete time Markov chain with finite state space Ω, which we denote by $\mathcal{M} = (X_t)_{t \in \mathbb{N}} = (X_0, X_1, X_2, \dots)$. If A is a subset of Ω, we write $A \subseteq \Omega$, and denote its complement by $A^c \coloneqq \Omega \setminus A$. We write $\mathbb{P}$ to denote the probability measure of the underlying probability space under consideration. Recall that a Markov chain satisfies the Markov property: for all $x, y \in \Omega$,

$$\mathbb{P}\left(X_{t+1} = y \mid X_0 = x_0,\, X_1 = x_1,\, \dots,\, X_t = x\right) = \mathbb{P}\left(X_{t+1} = y \mid X_t = x\right). \tag{1.3}$$

We will consider time homogeneous chains, such that (1.3) does not depend on t. The dependence of the chain on its starting state $X_0 = x$ will be indicated by $\mathbb{P}_x$.

This allows the chain to be defined by an $|\Omega| \times |\Omega|$ transition matrix P, such that $P(x, y) = \mathbb{P}\left(X_{t+1} = y \mid X_t = x\right)$ for all $t \in \mathbb{N}$. By convention, P is right stochastic (i.e. each row sums to one). The t-step probabilities of the chain are given by matrix multiplication, such that $P^t(x, y) = \mathbb{P}_x\left(X_t = y\right)$. We also write $\mathbb{P}_x\left(X_t \in A\right) = \sum_{y \in A} P^t(x, A)$ to denote the probability of starting in x and ending in a subset A of Ω after t steps.

We denote probability distributions on Ω by π, μ, ν, etc. These correspond to non-negative row vectors that sum to one. Given a distribution μ and $A \subseteq \Omega$, we can denote the probability of an event A by $\mu(A) = \sum_{x \in A} \mu(A)$. We say that a row vector π is stationary for P if $\pi P = \pi$. If, in addition, π sums to one, then we say that π is a stationary distribution for P.

Column vectors correspond to real-valued functions on Ω, which we denote by f, g, etc. We will write $\mathbb{E}_\mu\left[f\right] = \sum_{x \in \Omega} f(x)\mu(x)$ to denote the expected value of f with respect to distribution μ. Similarly, we write $\mathrm{Var}_\mu(f) = \sum_{x \in \Omega}(f(x) - \mathbb{E}_\mu\left[f\right])^2 \mu(x)$ to denote the variance of f with respect to distribution μ.

We write $\mathbb{1}_A : \Omega \to \{0, 1\}$ for the indicator function on Ω, given $A \subseteq \Omega$. This satisfies $\mathbb{1}_A(y) = 1$ if $y \in A$, and $\mathbb{1}_A(y) = 0$ if $y \notin A$. We write $\delta_x \coloneqq \mathbb{1}_{\{x\}}$ for the Dirac delta function, which represents a distribution with point mass at x.

Some asymptotic notation will also be used. If $(a_n)_{n \in \mathbb{Z}_+}$ and $(b_n)_{n \in \mathbb{Z}_+}$ are sequences of non-negative numbers, we write $b_n = o(a_n)$ to mean that b_n/a_n tends to zero as $n \to \infty$. We write $b_n = O(a_n)$ to mean that b_n/a_n is bounded from above by a constant for sufficiently large n. Similarly, we write $b_n = \Omega(a_n)$ to mean that b_n/a_n is bounded from below by a constant for sufficiently large n. Finally, we write $a_n \asymp b_n$ to mean that $a_n = O(b_n)$ and $a_n = \Omega(b_n)$, i.e. a_n and b_n are of the same order asymptotically.

Chapter 2

The Stationary Distribution of a Markov Chain

This chapter will describe some fundamental results on the theory of Markov chains. In Section 2.1, we prove the existence and uniqueness of a stationary distribution for an irreducible chain. Section 2.2 will show that if such a chain is also aperiodic, then it will converge towards its stationary distribution. Furthermore, Section 2.3 will relate the stationary distribution to an ergodic theorem for Markov chains. Finally, Section 2.4 will introduce the total variation distance between distributions, and mixing times.

The main reference for this chapter is the textbook by Levin, Peres and Wilmer [38]. For an extension of the theory to chains in continuous time and with countable state spaces, see [38, Chapters 20, 21]. For results on chains with general state spaces, see [50].

2.1 Existence and uniqueness of stationary distributions

We begin by introducing some terminology used in the classification of Markov chains (for a classical reference, refer to Feller [24]). Let $x, y \in \Omega$ be any two states. We say that y is **accessible** from x, written $x \to y$, if $P^r(x,y) > 0$ for some $r \in \mathbb{Z}_+$. If $x \to y$ and $y \to x$, we say that x and y **communicate**, written $x \leftrightarrow y$. It is straightforward to verify that $\leftrightarrow$ is an equivalence relation on Ω, and we call the induced equivalence classes **communicating classes**. We say that x is **absorbing** if $P(x,x) = 1$.

Definition 2.1. (i) A Markov chain $\mathcal{M}$ is **irreducible** if $x \leftrightarrow y$ for all $x, y \in \Omega$. In other words, there is a single communicating class, which means that for any $x, y \in \Omega$, there exists an $r \in \mathbb{Z}_+$ such that $P^r(x,y) > 0$.

(ii) For $x \in \Omega$, let $\mathcal{T}(x) := \{t \in \mathbb{Z}_+ : P^t(x,x) > 0\}$. The **period** of x is the greatest common divisor of $\mathcal{T}(x)$. It can be shown that this is a class property that is shared by all the members of a communicating class, and so the period of an irreducible chain is well-defined. An irreducible chain is **aperiodic** if its period is one.

Definition 2.2. The **hitting time** of $x \in \Omega$ is defined by $\tau_x := \min\{t \in \mathbb{N} : X_t = x\}$, where we take $\tau_x = \infty$ if the state is never reached. Similarly, for when only positive hitting times are desired, we define $\tau_x^+ := \min\{t \in \mathbb{Z}_+ : X_t = x\}$. If the chain starts at x, we call τ_x^+ the **first return time** of x.

A state $x \in \Omega$ is **recurrent** if $\mathbb{P}_x(\tau_x^+ < \infty) = 1$ (i.e. its first return time is almost surely bounded), otherwise it is **transient**. A recurrent state can further be classified as positive recurrent if $\mathbb{E}_x[\tau_x^+] < \infty$, or null recurrent otherwise. It can be shown that recurrence and transience are also class properties.

Recurrence and transience are more important for chains with countable (or larger) state spaces, which is beyond the scope of this thesis. The next lemma shows that an irreducible chain with a finite state space is automatically (positive) recurrent.

Lemma 2.3. *Let P be the transition matrix of an irreducible Markov chain with finite state space Ω. Then for any $x, y \in \Omega$, $\mathbb{E}_x\left[\tau_y^+\right] < \infty$, which implies that $\mathbb{P}_x\left(\tau_y^+ < \infty\right) = 1$.*

Proof. By irreducibility, there exists $r \in \mathbb{Z}_+$ and real $0 < \epsilon < 1$, such that for any pair of states $z, w \in \Omega$, there exists a positive integer $j \leq r$ such that $P^j(z, w) \geq \epsilon$. In particular, for any time t and value of X_t, the probability of hitting state y in the next r steps is at least ϵ. Hence, by the Markov property, $\mathbb{P}_x\left(\tau_y^+ > kr\right) \leq (1-\epsilon)\mathbb{P}_x\left(\tau_y^+ > (k-1)r\right)$ for any $k \in \mathbb{Z}_+$. Iterating this expression implies that

$$\mathbb{P}_x\left(\tau_y^+ > kr\right) \leq (1-\epsilon)^k. \tag{2.1}$$

Since τ_y^+ is a non-negative random variable, the tail sum formula for the expectation (which follows from interchanging the order of summation, e.g. see [26, p.84]) can be used to write

$$\mathbb{E}_x\left[\tau_y^+\right] = \sum_{t=0}^{\infty} \mathbb{P}_x\left(\tau_y^+ > t\right). \tag{2.2}$$

Note that the tail probabilities $\mathbb{P}_x\left(\tau_y^+ > t\right)$ are non-increasing in t. Therefore, they can be grouped into blocks of r, and then bounded from above by the first time in each block. Hence, by using (2.1),

$$\mathbb{E}_x\left[\tau_y^+\right] \leq \sum_{k=0}^{\infty} r \cdot \mathbb{P}_x\left(\tau_y^+ > kr\right) \leq r \sum_{k=0}^{\infty} (1-\epsilon)^k = \frac{r}{\epsilon},$$

which is finite. Next, note that $\mathbb{P}_x\left(\tau_y^+ = \infty\right) = \lim_{m\to\infty} \mathbb{P}_x\left(\tau_y^+ \geq m\right)$ by continuity from above. By Markov's inequality (Theorem 1.1), $\mathbb{P}_x\left(\tau_y^+ \geq m\right) \leq \frac{\mathbb{E}_x\left[\tau_y^+\right]}{m}$. This tends to zero as $m \to \infty$ since $\mathbb{E}_x\left[\tau_x^+\right]$ is finite, and therefore $\mathbb{P}_x\left(\tau_y^+ < \infty\right) = 0$. □

We will now prove, using a probabilistic approach, the existence and uniqueness of a stationary distribution for an irreducible Markov chain. Intuitively, the assumption of irreducibility means that the chain cannot be decomposed into separate chains, each of which supports a different stationary distribution.

Proposition 2.4. *Let P be the transition matrix of an irreducible Markov chain with finite state space Ω. Then the strictly positive distribution $\pi(x) = 1/\mathbb{E}_x\left[\tau_x^+\right]$ satisfies $\pi = \pi P$. Moreover, π is the unique stationary distribution of P.*

Proof. Existence. Fix any arbitrary state $x \in \Omega$. For $y \in \Omega$, define

$$\widetilde{\pi}_x(y) := \sum_{t=0}^{\infty} \mathbb{P}_x\left(X_t = y, \tau_x^+ > t\right) = \mathbb{E}_x\left[\#\text{ visits to } y \text{ before returning to } x\right]. \tag{2.3}$$

Note that $\mathbb{P}_x\left(X_t = y, \tau_x^+ > t\right) \le \mathbb{P}_x\left(\tau_x^+ > t\right)$, and hence $\widetilde{\pi}_x(y) \le \mathbb{E}_x\left[\tau_x^+\right]$. By Lemma 2.3, $\mathbb{E}_x\left[\tau_x^+\right]$ is finite, and so $\widetilde{\pi}_x$ is well-defined. We claim that $\widetilde{\pi}_x$ is stationary for P. For any $y \in \Omega$,

$$\widetilde{\pi}_x\, P(y) = \sum_{z\in\Omega} \widetilde{\pi}_x(z) P(z,y) = \sum_{z\in\Omega} \sum_{t=0}^{\infty} \mathbb{P}_x\left(X_t = z, \tau_x^+ > t\right) P(z,y).$$

Note that the event $\{\tau_x^+ > t\}$ is equivalent to $\{\tau_x^+ \ge t+1\}$. Moreover, by the Markov property, $\mathbb{P}_x\left(X_t = z,\, \tau_x^+ > t\right) P(z,y) = \mathbb{P}_x\left(X_t = z,\, X_{t+1} = y,\, \tau_x^+ \ge t+1\right)$. Since the summands are non-negative, we can interchange the order of summation, which shows that

$$\begin{aligned}
\widetilde{\pi}_x\, P(y) &= \sum_{t=0}^{\infty} \sum_{z\in\Omega} \mathbb{P}_x\left(X_t = z, X_{t+1} = y, \tau_x^+ \ge t+1\right) \\
&= \sum_{t=1}^{\infty} \mathbb{P}_x\left(X_t = y, \tau_x^+ \ge t\right) \\
&= \widetilde{\pi}_x(y) - \mathbb{P}_x\left(X_0 = y, \tau_x^+ > 0\right) + \sum_{t=1}^{\infty} \mathbb{P}_x\left(X_t = y, \tau_x^+ = t\right). \qquad (2.4)
\end{aligned}$$

For the last equality, (2.3) was used. Observe that the second term in the final expression is equal to $\mathbb{P}_x\left(X_0 = y\right) = \mathbb{1}_{\{y=x\}}$, and the last term is equal to $\mathbb{P}_x\left(X_{\tau_x^+} = y\right) = \mathbb{1}_{\{y=x\}}$. If $y = x$ then both terms are one and cancel, and if $y \ne x$ then both terms are zero.

Therefore, $\widetilde{\pi}_x\, P = \widetilde{\pi}_x$, which shows that $\widetilde{\pi}_x$ is stationary for P, and it suffices to normalise $\widetilde{\pi}_x$. By interchanging the order of summation, and then using the tail sum formula for the expectation again, as in (2.2), the normalising factor is

$$\sum_{y\in\Omega} \widetilde{\pi}_x(y) = \sum_{t=0}^{\infty} \sum_{y\in\Omega} \mathbb{P}_x\left(X_t = y, \tau_x^+ > t\right) = \sum_{t=0}^{\infty} \mathbb{P}_x\left(\tau_x^+ > t\right) = \mathbb{E}_x\left[\tau_x^+\right].$$

Hence,

$$\pi(y) := \frac{\widetilde{\pi}_x(y)}{\mathbb{E}_x\left[\tau_x^+\right]} \quad \text{for all } y \in \Omega \qquad (2.5)$$

is a probability distribution satisfying $\pi P = \pi$. Since the choice of x was arbitrary, choosing $x = y$ shows that $\pi(y) = 1/\mathbb{E}_y\left[\tau_y^+\right]$.

Uniqueness. Suppose that $\widetilde{\pi}$ is another stationary distribution of P, such that $\widetilde{\pi}P = \widetilde{\pi}$. Since Ω is finite, there exists a state $y \in \Omega$ that minimises the ratio $\pi(x)/\widetilde{\pi}(x)$. In other words,

$$c := \frac{\pi(y)}{\widetilde{\pi}(y)} \le \frac{\pi(x)}{\widetilde{\pi}(x)} \qquad \text{for all } x \in \Omega.$$

Suppose that $x \in \Omega$ satisfies $P(x,y) > 0$ and $\pi(x)/\widetilde{\pi}(x) > c$. Using $\pi P = \pi$ and $\widetilde{\pi}P = \widetilde{\pi}$ with the inequalities $\widetilde{\pi}(x) < \pi(x)/c$ and $\widetilde{\pi}(z) \le \pi(z)/c$ for all $z \in \Omega$ implies that

$$\widetilde{\pi}(y) = \widetilde{\pi}(x)P(x,y) + \sum_{z\ne x} \widetilde{\pi}(z)P(z,y) < \frac{1}{c}\left(\pi(x)P(x,y) + \sum_{z\ne x} \pi(z)P(z,y)\right) = \frac{\pi(y)}{c}.$$

This shows that $\pi(y)/\widetilde{\pi}(y) > c$, contradicting the choice of y. Therefore, if $P(x,y) > 0$, then $\pi(x)/\widetilde{\pi}(x) = c$. For any $x \in \Omega$, irreducibility of P implies that there exists a finite sequence $x = x_0, x_1, \ldots, x_k = y$ with $P(x_{i-1}, x_i) > 0$. By the above argument, we have $c = \pi(y)/\widetilde{\pi}(y) = \pi(x_{k-1})/\widetilde{\pi}(x_{k-1}) = \cdots = \pi(x)/\widetilde{\pi}(x)$, and so $\pi(x) = c\widetilde{\pi}(x)$ for all $x \in \Omega$. Since π and $\widetilde{\pi}$ sum to one, it follows that $c = 1$ and $\pi = \widetilde{\pi}$. □

Definition 2.5. Let P be the transition matrix of a Markov chain. If π is a probability distribution on Ω which satisfies the **detailed balance equations**

$$\pi(x)P(x,y) = \pi(y)P(y,x) \qquad \text{for all } x, y \in \Omega, \tag{2.6}$$

then we say that the chain (or equivalently, P) is **reversible** with respect to π.

Reversibility is a very important property. The following shows that it automatically provides a stationary distribution (which is unique, if the chain is irreducible). In Section 4.1, reversible chains will also be shown to have nice spectral properties.

Proposition 2.6. *Let π be a distribution on a finite state space Ω. If P is a transition matrix which is reversible with respect to π, then π is stationary for P.*

Proof. Since π satisfies the detailed balance conditions (2.6), and $\sum_{y\in\Omega} P(x,y) = 1$,

$$\pi P(x) = \sum_{y\in\Omega} \pi(y)P(y,x) = \sum_{y\in\Omega} \pi(x)P(x,y) = \pi(x), \qquad \text{for all } x \in \Omega.$$

Hence, π is stationary since it satisfies $\pi P = \pi$. □

Remark 2.7. (i) A symmetric transition matrix satisfies $P(x,y) = P(y,x)$ for all $x, y \in \Omega$. If P is symmetric, then it is reversible with respect to the uniform distribution π, given by $\pi(x) = |\Omega|^{-1}$ for all $x \in \Omega$.

(ii) A random walk on a group G with increment distribution μ has transition probabilities $P(g, hg) = \mu(h)$ for all $g, h \in G$. i.e. μ specifies an element with which we multiply the current state on the left. We say that μ is symmetric if $\mu(g) = \mu(g^{-1})$ for all $g \in G$.

Let π be the uniform distribution on G. A quick calculation shows that $\pi = \pi P$ holds, since $\pi(g) = \sum_{h\in G} \frac{1}{|G|} \cdot P(h,g) = \frac{1}{|G|} \sum_{k\in G} \mu(k) = \frac{1}{|G|}$ for all $g \in G$, where we re-indexed by $k = gh^{-1}$ in the last sum. Thus, π is stationary for P. Moreover, P is reversible with respect to π if and only if μ is symmetric.

2.2 Convergence to stationary distribution

If $\mathcal{M}$ is an irreducible Markov chain, then it has a unique stationary distribution π by Proposition 2.4. We will prove that under the further assumption that the chain is aperiodic, then it will converge towards π.

First, we will need the following lemma which shows that an aperiodic, irreducible matrix can be raised to a suitable power such that it becomes strictly positive. The proof, which will be omitted, relies on some number-theoretic facts about subsets of $\mathbb{N}$.

Lemma 2.8 ([38, Proposition 1.7])**.** *Suppose that P is an aperiodic, irreducible transition matrix. Then there exists $r \in \mathbb{Z}_+$ such that $P^r > 0$, or in other words, $P^r(x, y) > 0$ for all $x, y \in \Omega$.*

The following facts about the powers of stochastic matrices (i.e. each row sums to one) will also be useful.

Lemma 2.9. (i) *If A and B are $n \times n$ stochastic matrices, then their product AB is also stochastic. In particular, this implies that A^k is stochastic for any $k \in \mathbb{N}$.*

(ii) *If π is a row vector that is stationary for a stochastic matrix S, then π is also stationary for S^k for any $k \in \mathbb{N}$.*

Proof. (i) The fact that the rows of A and B sum to one can be written as $A\mathbf{1} = \mathbf{1}$ and $B\mathbf{1} = \mathbf{1}$, where $\mathbf{1}$ is the column vector consisting of ones. Therefore, $AB\mathbf{1} = A\mathbf{1} = \mathbf{1}$ by the associativity of matrix multiplication, so the rows of AB also sum to one. It is clear that if all the entries of A and B are non-negative, then they will also be for AB.

(ii) This is clear from $\pi S^k = (\pi S)S^{k-1} = \pi S^{k-1} = \cdots = \pi$. □

Theorem 2.10 (Convergence theorem)**.** *Let P be the transition matrix of an irreducible and aperiodic Markov chain with finite state space Ω and stationary distribution π. Then there exist constants $0 < \alpha < 1$ and $C > 0$ such that*

$$\sum_{y \in \Omega} |P^t(x, y) - \pi(y)| \leq C\alpha^t \quad \text{for all } x \in \Omega. \tag{2.7}$$

In particular, for any starting state $z \in \Omega$, $\pi(x) = \lim_{t\to\infty} P^t(z, x)$.

Proof. Let Π be the $|\Omega| \times |\Omega|$ stochastic matrix such that each row is equal to π. The strategy will be to decompose P into an independent mixture of Π and another stochastic matrix Q, such that, in the long run, most of the draws come from Π.

By Lemma 2.8, there exists $r \in \mathbb{Z}_+$ such that $P^r > 0$. From Proposition 2.4, the stationary distribution satisfies $\pi(y) > 0$ for all $y \in \Omega$. Therefore, $P^r(x, y) = \delta_{xy}\,\pi(y)$ for some constants $\delta_{xy} > 0$. By taking the minimum of δ_{xy} over all states, there exists a constant $0 < \delta < 1$ satisfying $\delta \leq \delta_{xy}$ for all $x, y \in \Omega$ such that

$$P^r(x, y) \geq \delta\pi(y) \quad \text{for all } x, y \in \Omega. \tag{2.8}$$

Let $\theta := 1 - \delta$. Then (2.8) allows us to define a stochastic matrix Q such that

$$P^r = (1 - \theta)\Pi + \theta Q. \tag{2.9}$$

We will show by induction that

$$P^{rk} = (1 - \theta^k)\Pi + \theta^k Q^k. \tag{2.10}$$

This holds for $k = 1$ by (2.9). Assuming that this holds for $k = m$, we can write

$$\begin{aligned} P^{r(m+1)} = P^{rm}P^r &= (1 - \theta^m)\Pi P^r + \theta^m Q^m P^r \\ &= (1 - \theta^m)\Pi P^r + (1 - \theta)\theta^m Q^m \Pi + \theta^{m+1} Q^{m+1}, \end{aligned}$$

where (2.9) was used in the second term of the sum. Note that $A\Pi = \Pi$ for any stochastic matrix A. By Lemma 2.9, Q^m is a stochastic matrix, and therefore $Q^m\Pi = \Pi$. Furthermore, if S is a stochastic matrix satisfying $\pi S = \pi$, then it also satisfies $\Pi S = \Pi$. The same lemma shows that π is stationary for P^r, and therefore $\Pi P^r = \Pi$. Thus,

$$\begin{aligned} P^{r(m+1)} &= (1 - \theta^m)\Pi + (1 - \theta)\theta^m \Pi + \theta^{m+1} Q^{m+1} \\ &= (1 - \theta^{m+1})\Pi + \theta^{m+1} Q^{m+1}. \end{aligned}$$

Hence (2.10) holds for $k = m + 1$, and consequently for all k. For any $j \in \mathbb{N}$, note that $\Pi P^j = \Pi$. Therefore, multiplying (2.10) on the right by P^j and rearranging shows that

$$P^{rk+j} - \Pi = \theta^k (Q^k P^j - \Pi). \tag{2.11}$$

For any fixed row $x \in \Omega$, this gives the difference in distribution between the chain started at x after $rk + j$ steps, and the stationary distribution. Therefore,

$$\sum_{y \in \Omega} |P^{rk+j}(x, y) - \pi(y)| = \theta^k \sum_{y \in \Omega} |Q^k P^j(x, y) - \pi(y)| \leq 2\theta^k. \tag{2.12}$$

The inequality follows from $\sum_{y \in \Omega} |Q^k P^j(x, y) - \pi(y)| \leq \sum_{y \in \Omega} (Q^k P^j(x, y) + \pi(y)) = 2$, using the triangle inequality, and the fact that $Q^k P^j$ is stochastic from Lemma 2.9. For any $t \in \mathbb{N}$, we can write $t = rk + j$, where $0 \leq j < r$. By defining $\alpha := \theta^{1/r} \in (0, 1)$ and $C := 2\theta^{(r-1)/r} > 0$, we have the inequality $2\theta^k = 2\theta^{-j/r}\theta^{t/r} \leq C\alpha^t$. Since this works for all $x \in \Omega$, putting these constants into (2.12) immediately implies (2.7). □

The following simple example shows that aperiodicity is necessary for convergence.

Example 2.11. Consider a Markov chain with $\Omega = \{x, y\}$ and transition matrix

$$P = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}.$$

It is clear that P is irreducible, has period two, and has unique stationary distribution $\pi = \begin{pmatrix} \frac{1}{2} & \frac{1}{2} \end{pmatrix}$. However, P cannot converge to π as t tends to infinity, since $P^t(x, y) = 1$ if t is even, and $P^t(x, y) = 0$ if t is odd.

The following proposition explains that if a chain is periodic, then its states can be partitioned into equivalence classes on which it moves between in a cyclic manner.

Proposition 2.12 ([38, Exercise 1.6]). *Let P be the transition matrix of an irreducible Markov chain with finite state space Ω. If P has period $b \in \mathbb{Z}_+$, then Ω can be partitioned into b classes $\mathcal{C}_0, \mathcal{C}_1, \ldots, \mathcal{C}_{b-1}$ such that $P(x,y) > 0$ only if $x \in \mathcal{C}_k$ and $y \in \mathcal{C}_{k+1}$ (where the indices are taken modulo b).*

Proof. Fix $x_0 \in \Omega$. Recall that the period of the chain is b if and only if $\gcd \mathcal{T}(x_0) = b$, where $\mathcal{T}(x_0) = \{t \in \mathbb{Z}_+ : P^t(x_0, x_0) > 0\}$. For $k = 0, 1, \ldots, b-1$, define

$$\mathcal{C}_k \coloneqq \{x \in \Omega : P^{mb+k}(x_0, x) > 0 \text{ for some } m \in \mathbb{N}\}, \tag{2.13}$$

which is the set of states that are accessible from x_0 in $k \pmod b$ steps. It can be checked that $\{\mathcal{C}_k\}_{k=0}^{b-1}$ partitions Ω. In other words, the sets are pairwise disjoint and cover Ω. Furthermore, it can also be checked that if $x \in \mathcal{C}_k$ and $P(x,y) > 0$, then $y \in \mathcal{C}_{k+1}$ necessarily. However, the details for these claims will be omitted. □

Remark 2.13. In practice, periodicity is not a real problem, since any chain can be made aperiodic by adding in self-loops, such that $\widetilde{P}(x,x) \geq 1/2$ for all $x \in \Omega$. Such a chain is called a **lazy** chain. Lazy chains are further discussed in Lemma 4.4 and its accompanying Remark 4.5 in Section 4.1.

While the proofs in this chapter have taken a probabilistic route so far, there are in fact multiple proofs of these classical results in Markov chain theory. In particular, an algebraic approach uses the Perron-Frobenius theorem, which has the benefit of highlighting the important connection to the eigenvalues of P. We will now present some of these main results without proof, however the full details can be found in [53].

Theorem 2.14 (Perron-Frobenius Theorem, [53, Theorem 1.1, Corollary 1]). *We say that a matrix A is **primitive** if $A^k > 0$ for some $k \in \mathbb{Z}_+$. Suppose that A is a non-negative, primitive matrix. Then there exists a real eigenvalue $r > 0$, which we call the **Perron-Frobenius eigenvalue** of A, such that*

(i) *There exist strictly positive left and right eigenvectors corresponding to r. Moreover, the eigenvectors corresponding to r are unique up to constant multiples.*

(ii) *For any other eigenvalue λ of A, $|\lambda| < r$.*

(iii) *r is bounded from below by the minimum row (resp. column) sum and from above by the maximum row (resp. column) sum.*

Let P be an aperiodic, irreducible transition matrix, then Lemma 2.8 shows that P is primitive. Since the row sums of P are all one, Theorem 2.14 immediately implies that the Perron-Frobenius eigenvalue of P is $r = 1$. Moreover, the left eigenvectors of P corresponding to r live in a one-dimensional subspace, and hence there exists a unique vector π with entries summing to one that satisfies $\pi P = \pi$, which implies the existence

and uniqueness of a stationary distribution. Finally, the convergence theorem follows from [53, Theorem 1.2], which implies that $P^t \to \mathbf{1}\pi$ elementwise as $t \to \infty$.

Another consequence of the Perron-Frobenius theorem is that all the eigenvalues of P can be written in decreasing order $1 = \lambda_1 > \lambda_2 \geq \cdots \geq \lambda_{|\Omega|} > -1$. We will particularly be interested in the eigenvalue with the largest absolute value, which is either λ_2 or $\lambda_{|\Omega|}$, since it can be used to bound mixing times as Section 4.1 will describe later.

2.3 Ergodic theorem for Markov chains

We will finish this chapter by proving how the stationary distribution of an irreducible Markov chain relates to an ergodic theorem. Intuitively, this describes how an "average over time" equals an "average over the state space" in the long run.

First, we will introduce another probabilistic concept. We say that a random variable τ taking values in $\mathbb{N} \cup \{+\infty\}$ is a **stopping time** for the stochastic process $(X_t)_{t \in \mathbb{N}}$ if the event $\{\tau = t\}$ is determined by $X_0, X_1, \ldots, X_t$ for every $t \in \mathbb{N}$.

The following proposition proves the **strong Markov property**, which means that a chain will also "forget the past" whenever it is stopped by a stopping time. This is not too difficult to show for discrete-time chains with countable state spaces, since this avoids any complications from working in continuous time, or with uncountable state spaces.

Proposition 2.15. *Let τ be a stopping time for a discrete-time Markov chain $\mathcal{M}$ with finite state space Ω. Let A denote the event $\{X_0 = x_0, X_1 = x_1, \ldots, X_{\tau-1} = x_{\tau-1}\}$. Then for any $\ell \in \mathbb{Z}_+$, conditional on $\tau < \infty$,*

$$\mathbb{P}\left(X_{\tau+1} = y_1, \ldots, X_{\tau+\ell} = y_\ell \mid X_\tau = x, A\right) = \mathbb{P}_x\left(X_1 = y_1, \ldots, X_\ell = y_\ell\right). \tag{2.14}$$

Proof. It will be sufficient to prove (2.14) for $\ell = 1$ (i.e. one time step). By the definition of conditional probability and the Law of Total Probability (e.g. see [26, p.22]),

$$\begin{aligned}
\mathbb{P}_{x_0}\left(X_{\tau+1} = y \mid X_\tau = x, A\right) &= \frac{\mathbb{P}\left(X_{\tau+1} = y, X_\tau = x, A\right)}{\mathbb{P}\left(X_\tau = x, A\right)} \\
&= \frac{\sum_{k=0}^{\infty} \mathbb{P}\left(X_{k+1} = y, X_k = x, A\right) \mathbb{P}\left(\tau = k\right)}{\sum_{k=0}^{\infty} \mathbb{P}\left(X_k = x, A\right) \mathbb{P}\left(\tau = k\right)} \\
&= \frac{\sum_{k=0}^{\infty} \mathbb{P}\left(X_{k+1} = y \mid X_k = x, A\right) \mathbb{P}\left(X_k = x, A\right) \mathbb{P}\left(\tau = k\right)}{\sum_{k=0}^{\infty} \mathbb{P}\left(X_k = x, A\right) \mathbb{P}\left(\tau = k\right)}.
\end{aligned}$$

By the Markov property, $\mathbb{P}\left(X_{k+1} = y \mid X_k = x, A\right) = \mathbb{P}_x\left(X_1 = y\right)$. Therefore,

$$\mathbb{P}_{x_0}\left(X_{\tau+1} = y \mid X_\tau = x, A\right) = \frac{\mathbb{P}_x\left(X_1 = y\right) \sum_{k=0}^{\infty} \mathbb{P}\left(X_k = x, A\right) \mathbb{P}\left(\tau = k\right)}{\sum_{k=0}^{\infty} \mathbb{P}\left(X_k = x, A\right) \mathbb{P}\left(\tau = k\right)} = \mathbb{P}_x\left(X_1 = y\right),$$

as desired. □

The following technical lemma provides conditions under which convergence of a series along a subsequence implies convergence of the entire series.

Lemma 2.16. *Let $(a_n)_{n\in\mathbb{N}}$ be a bounded sequence. Suppose that $(a_{n_k})_{k\in\mathbb{N}}$ is a subsequence with indices satisfying $\lim_{k\to\infty} \frac{n_k}{n_{k+1}} = 1$. Then*

$$\lim_{k\to\infty} \frac{a_{n_1} + a_{n_2} \cdots + a_{n_k}}{n_k} = a \qquad \textit{implies that} \qquad \lim_{n\to\infty} \frac{a_1 + a_2 + \cdots + a_n}{n} = a.$$

Proof. Since $(a_n)_{n\in\mathbb{N}}$ is bounded, there exists a real $B \geq 0$ such that $|a_n| \leq B$ for all $n \in \mathbb{N}$. Fix $n \in \mathbb{N}$. Then we can choose k such that $n_k \leq n < n_{k+1}$. Let $A_n = \frac{1}{n}\sum_{j=1}^{n} a_j$. Then

$$A_n = \frac{n_k}{n} A_{n_k} + \frac{\sum_{j=n_k+1}^{n} a_j}{n}. \tag{2.15}$$

Since $\frac{n_k}{n_{k+1}} < \frac{n_k}{n} \leq 1$, and $\frac{n_k}{n_{k+1}} \to 1$ by assumption, this implies $\frac{n_k}{n} \to 1$ as $n \to \infty$ by sandwiching. Moreover, $A_{n_k} \to a$ as $n \to \infty$, also by assumption. The absolute value of the second term of (2.15) is bounded from above by $\frac{B(n_{k+1}-n_k)}{n_k}$, which tends to zero as $n \to \infty$. Hence, the second term also tends to zero. Therefore, $A_n \to a$ as $n \to \infty$. □

We can now prove the ergodic theorem. The key idea will be to use the hitting time τ_x^+ defined in Definition 2.2 (which is clearly a stopping time), and the strong Markov property to separate the chain into independent and identically distributed (i.i.d.) blocks.

Theorem 2.17 (Ergodic theorem)**.** *Let $\mathcal{M} = (X_t)_{t\in\mathbb{N}}$ be an irreducible Markov chain with finite state space Ω, and f be a real-valued function on Ω. Then for any starting distribution μ on Ω, $\frac{1}{t}\sum_{s=0}^{t-1} f(X_s) \to \mathbb{E}_\pi[f]$ as $t \to \infty$ almost surely. Equivalently,*

$$\mathbb{P}_\mu\left(\lim_{t\to\infty} \frac{1}{t}\sum_{s=0}^{t-1} f(X_s) = \mathbb{E}_\pi[f]\right) = 1. \tag{2.16}$$

In particular, taking $f = \delta_x$ shows that π describes the average time spent in each state in the long run:

$$\mathbb{P}_\mu\left(\lim_{t\to\infty} \frac{1}{t}\sum_{s=0}^{t-1} \mathbb{1}_{\{X_s=x\}} = \pi(x)\right) = 1. \tag{2.17}$$

Proof. Fix a state $x \in \Omega$. Consider the sequence of stopping times $(\tau_{x,k}^+)_{k\in\mathbb{N}}$, where

$$\tau_{x,0}^+ := 0, \qquad \tau_{x,k}^+ := \min\{t > \tau_{x,k-1}^+ : X_t = x\}, \quad k \in \mathbb{N}. \tag{2.18}$$

For convenience, write $\tau_x^+ := \tau_{x,1}^+$, which is the first return time in Definition 2.2. Note that τ_x^+ is almost surely bounded by Lemma 2.3. Define

$$Y_k := \sum_{s=\tau_{x,(k-1)}^+}^{\tau_{x,k}^+ - 1} f(X_s), \quad \text{and} \quad S_t := \sum_{s=0}^{t-1} f(X_s). \tag{2.19}$$

Note that for all $k \in \mathbb{N}$, $X_{\tau_{x,(k-1)}^+} = x$. Since each Y_k is completely determined by $X_{\tau_{x,(k-1)}^+}, X_{\tau_{x,(k-1)}^+ + 1}, \ldots, X_{\tau_{x,k}^+ - 1}$, the strong Markov property (Proposition 2.15) implies

that $(Y_k)_{k\in\mathbb{N}}$ forms an i.i.d. sequence. Each Y_k has common mean

$$\mathbb{E}_x[Y_1] = \mathbb{E}_x\left[\sum_{s=0}^{\tau_x^+-1} f(X_s)\right] = \mathbb{E}_x\left[\sum_{s=0}^{\tau_x^+-1}\sum_{y\in\Omega} f(y)\mathbb{1}_{\{X_s=y\}}\right] = \sum_{y\in\Omega} f(y)\mathbb{E}_x\left[\sum_{s=0}^{\tau_x^+-1}\mathbb{1}_{\{X_s=y\}}\right].$$

For the last equality, the sum was interchanged, and the linearity of the expectation was used. Observe that the expectation in the last term is equal to $\widetilde{\pi}_x(y)$, the expected number of visits to y before returning to x, as defined in (2.3). This was shown to be equal to $\pi(y)\,\mathbb{E}_x[\tau_x^+]$ in (2.5). Therefore,

$$\mathbb{E}_x[Y_1] = \sum_{y\in\Omega} f(y)\pi(y)\mathbb{E}_x\left[\tau_x^+\right] = \mathbb{E}_\pi[f]\;\mathbb{E}_x\left[\tau_x^+\right]. \tag{2.20}$$

Since $\mathbb{E}_x[\tau_x^+] < \infty$ from Lemma 2.3, (2.20) shows that $\mathbb{E}_x[Y_1]$ is also finite. Therefore, since $S_{\tau_{x,n}^+} = \sum_{k=1}^n Y_k$ is a sum of the n i.i.d. Y_k, the Strong Law of Large Numbers ([26, p.329]) implies that

$$\mathbb{P}_x\left(\lim_{n\to\infty}\frac{S_{\tau_{x,n}^+}}{n} = \mathbb{E}_x[Y_1]\right) = 1. \tag{2.21}$$

Next, we can also write $\tau_{x,n}^+ = \sum_{k=1}^n(\tau_{x,k}^+ - \tau_{x,(k-1)}^+)$. By the strong Markov property again, each increment is i.i.d., since the chain effectively resets after each return to x. Therefore, using the Strong Law of Large Numbers again implies that

$$\mathbb{P}_x\left(\lim_{n\to\infty}\frac{\tau_{x,n}^+}{n} = \mathbb{E}_x\left[\tau_x^+\right]\right) = 1. \tag{2.22}$$

Combining the almost sure statements in (2.21) and (2.22), and using (2.20), shows that

$$\mathbb{P}_x\left(\lim_{n\to\infty}\frac{S_{\tau_{x,n}^+}}{\tau_{x,n}^+} = \mathbb{E}_\pi[f]\right) = 1. \tag{2.23}$$

Note that $f(X_t)$ is bounded, since the state space Ω is finite. Furthermore, for all $n\in\mathbb{N}$, $\tau_{x,(n+1)}^+ \geq n+1$, and $\tau_{x,(n+1)}^+ - \tau_{x,n}^+$ has the same distribution as τ_x^+. Therefore, each increment $\tau_{x,(n+1)}^+ - \tau_{x,n}^+$ is also almost surely bounded, and so there exists a real $M>0$ such that

$$0 \leq \frac{\tau_{x,(n+1)}^+ - \tau_{x,n}^+}{\tau_{x,(n+1)}^+} \leq \frac{M}{n+1} \qquad \text{almost surely.}$$

Hence, $\lim_{n\to\infty}\frac{\tau_{x,n}^+}{\tau_{x,(n+1)}^+} = 1$ almost surely. By Lemma 2.16, (2.23) implies that

$$\mathbb{P}_x\left(\lim_{t\to\infty}\frac{1}{t}\sum_{s=0}^{t-1} f(X_s) = \mathbb{E}_\pi[f]\right) = 1.$$

Since this holds for any $x \in \Omega$, we can verify that (2.16) holds for any starting distribution μ on Ω by averaging across all the states. That is,

$$\mathbb{P}_\mu\left(\lim_{t\to\infty}\frac{1}{t}\sum_{s=0}^{t-1} f(X_s) = \mathbb{E}_\pi[f]\right) = \sum_{x\in\Omega}\mathbb{P}_x\left(\lim_{t\to\infty}\frac{1}{t}\sum_{s=0}^{t-1} f(X_s) = \mathbb{E}_\pi[f]\right)\mu(x) = 1,$$

as desired. □

2.4 Total variation distance and mixing times

In this section, we will make precise the notion of distance between two distributions. This will allow the rate of convergence towards stationarity to be analysed.

Definition 2.18. Let μ and ν be two probability distributions on a finite state space Ω. The **total variation distance** between μ and ν is defined by

$$\|\mu - \nu\|_{\mathrm{TV}} \coloneqq \max_{A\subseteq\Omega} |\mu(A) - \nu(A)|, \tag{2.24}$$

where $\mu(A) = \sum_{x\in A}\mu(x)$ and similarly for ν. Since we are interested in bounding the distance of P^t to stationarity uniformly over all starting states, define

$$d(t) \coloneqq \max_{x\in\Omega}\left\|P^t(x,\cdot) - \pi\right\|_{\mathrm{TV}}, \tag{2.25}$$

$$\overline{d}(t) \coloneqq \max_{x,y\in\Omega}\left\|P^t(x,\cdot) - P^t(y,\cdot)\right\|_{\mathrm{TV}}. \tag{2.26}$$

Proposition 2.19. *The total variation distance has the equivalent representations:*

(i) *Half the usual L^1 norm:*

$$\|\mu - \nu\|_{\mathrm{TV}} = \frac{1}{2}\sum_{x\in\Omega}|\mu(x) - \nu(x)|. \tag{2.27}$$

(ii) *The usual operator norm for the dual of the space of bounded, measurable functions:*

$$\|\mu - \nu\|_{\mathrm{TV}} = \sup_{\|f\|_\infty\le 1}|\,\mathbb{E}_\mu[f] - \mathbb{E}_\nu[f]\,|. \tag{2.28}$$

Here $f : \Omega \to \mathbb{R}$, $\|f\|_\infty = \max_{x\in\Omega}|f(x)|$, and $\mathbb{E}_\mu[f] = \sum_{x\in\Omega} f(x)\mu(x)$ (as for ν).

Proof. (i) Let $B \coloneqq \{x \in \Omega : \mu(x) - \nu(x) \ge 0\}$, and $A \subseteq \Omega$ be any event. Then

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

Since $\mu(x) - \nu(x) < 0$ for any $x \in B^c$, and $\mu(x) - \nu(x) \ge 0$ for any $x \in B$, it follows that

$$\mu(A) - \nu(A) \le \sum_{x\in B}[\mu(x) - \nu(x)] = \mu(B) - \nu(B). \tag{2.29}$$

Similarly, the same argument applied to B^c shows that

$$\nu(A) - \mu(A) \leq \nu(B^c) - \mu(B^c). \tag{2.30}$$

Observe that $\big[\mu(B) - \nu(B)\big] - \big[\nu(B^c) - \mu(B^c)\big] = \mu(\Omega) - \nu(\Omega) = 0$, and hence the upper bounds in (2.29) and (2.30) are actually equal. Since this holds for any $A \subseteq \Omega$, it follows that either set B or B^c provides the total variation distance. Hence,

$$\|\mu - \nu\|_{\mathrm{TV}} = \sum_{\substack{x\in\Omega \\ \mu(x)\geq\nu(x)}} [\mu(x) - \nu(x)] = \sum_{\substack{x\in\Omega \\ \mu(x)<\nu(x)}} [\nu(x) - \mu(x)]. \tag{2.31}$$

Using (2.31), it follows that

$$\|\mu - \nu\|_{\mathrm{TV}} = \frac{1}{2}\left(\sum_{\substack{x\in\Omega \\ \mu(x)\geq\nu(x)}} [\mu(x) - \nu(x)] + \sum_{\substack{x\in\Omega \\ \mu(x)<\nu(x)}} [\nu(x) - \mu(x)] \right) = \frac{1}{2} \sum_{x\in\Omega} |\mu(x) - \nu(x)|. \tag{2.32}$$

(ii) Suppose that f satisfies $\|f\|_\infty \leq 1$. Then, using the triangle inequality,

$$\begin{aligned} \frac{1}{2}\left| \sum_{x\in\Omega} f(x)\mu(x) - \sum_{x\in\Omega} f(x)\nu(x) \right| &\leq \frac{1}{2} \sum_{x\in\Omega} |f(x)| \cdot |\mu(x) - \nu(x)| \\ &\leq \frac{1}{2} \sum_{x\in\Omega} |\mu(x) - \nu(x)| = \|\mu - \nu\|_{\mathrm{TV}}. \end{aligned}$$

Hence, $\sup_{\|f\|_\infty \leq 1} |\mathbb{E}_\mu[f] - \mathbb{E}_\nu[f]| \leq \|\mu - \nu\|_{\mathrm{TV}}$. To show the opposite inequality, consider the function f defined by $f(x) = 1$ if $\mu(x) \geq \nu(x)$ (i.e. x is in the set B from part (i)), and $f(x) = -1$ if $\mu(x) < \nu(x)$. Then f satisfies $\|f\|_\infty \leq 1$, and

$$\frac{1}{2}\left| \sum_{x\in\Omega} f(x)\mu(x) - \sum_{x\in\Omega} f(x)\nu(x) \right| = \frac{1}{2}\left(\sum_{\substack{x\in\Omega \\ \mu(x)\geq\nu(x)}} [\mu(x) - \nu(x)] + \sum_{\substack{x\in\Omega \\ \mu(x)<\nu(x)}} [\nu(x) - \mu(x)] \right),$$

which is $\|\mu - \nu\|_{\mathrm{TV}}$ from (2.32). Therefore, $\sup_{\|f\|_\infty \leq 1} |\mathbb{E}_\mu[f] - \mathbb{E}_\nu[f]| \geq \|\mu - \nu\|_{\mathrm{TV}}$. □

Remark 2.20. (i) Proposition 2.19 shows that the total variation distance is a metric between distributions on Ω.

(ii) By (2.24), the total variation distance has the probabilistic interpretation as the maximum difference in probability assigned to any event.

(iii) Since two distributions are close only if they are uniformly close on all subsets of Ω, the total variation distance can be quite unforgiving of small, local deviations as the following example from [2] illustrates. Consider the state space S_n. Let π be the uniform distribution on S_n, representing a well-shuffled deck. Suppose that you

know the first card is c. Then your knowledge is represented by μ, the distribution that is uniform on all permutations σ with $\sigma(c) = 1$. Then, by (2.27),

$$\|\mu - \pi\|_{\mathrm{TV}} = (n-1)! \cdot \left[\frac{1}{(n-1)!} - \frac{1}{n!}\right] + (n! - (n-1)!) \cdot \frac{1}{n!} = 1 - \frac{1}{n}.$$

The following describes some key properties of the total variation distance.

Proposition 2.21. *Let μ and ν be two distributions on the finite state space Ω. Let P be the transition matrix of a Markov chain with stationary distribution π. Then*

(i) *The total variation distance is non-decreasing when the chain advances:*

$$\|\mu P - \nu P\|_{\mathrm{TV}} \le \|\mu - \nu\|_{\mathrm{TV}}\,. \tag{2.33}$$

(ii) *The total variation distance is convex:*

$$\left\|\mu P^t - \pi\right\|_{\mathrm{TV}} \le \sum_{x\in\Omega} \mu(x) \left\|P^t(x,\cdot) - \pi\right\|_{\mathrm{TV}}\,. \tag{2.34}$$

Proof. (i) By using (2.27),

$$\|\mu P - \nu P\|_{\mathrm{TV}} = \frac{1}{2}\sum_{x\in\Omega} |\mu P(x) - \nu P(x)| = \frac{1}{2}\sum_{x\in\Omega} \left|\sum_{y\in\Omega} [\mu(y)P(y,x) - \nu(y)P(y,x)]\right|.$$

Using the triangle inequality and interchanging the (finite) sum shows that

$$\|\mu P - \nu P\|_{\mathrm{TV}} \le \frac{1}{2}\sum_{y\in\Omega} |\mu(y) - \nu(y)| \sum_{x\in\Omega} P(y,x).$$

This upper bound is equal to $\frac{1}{2}\sum_{y\in\Omega} |\mu(y) - \nu(y)| = \|\mu - \nu\|_{\mathrm{TV}}$, since P is stochastic.

(ii) Since μ is a distribution, $\sum_{y\in\Omega} \mu(y) = 1$. Hence,

$$\left\|\mu P^t - \pi\right\|_{\mathrm{TV}} = \frac{1}{2}\sum_{x\in\Omega} |\mu P^t(x) - \pi(x)| = \frac{1}{2}\sum_{x\in\Omega} \left|\sum_{y\in\Omega} \mu(y)[P^t(y,x) - \pi(x)]\right|.$$

Using the triangle inequality, and interchanging the sum again, shows that

$$\left\|\mu P^t - \pi\right\|_{\mathrm{TV}} \le \frac{1}{2}\sum_{y\in\Omega} \mu(y) \sum_{x\in\Omega} |P^t(y,x) - \pi(x)| = \sum_{y\in\Omega} \mu(y) \left\|P^t(y,\cdot) - \pi\right\|_{\mathrm{TV}},$$

as desired. □

Corollary 2.22. *Assume that the same conditions as Proposition 2.21 hold.*

(i) *The function* $d(t) = \max_{x\in\Omega} \left\|P^t(x,\cdot) - \pi\right\|_{\mathrm{TV}}$ *from (2.25) is monotone decreasing:*

$$d(t+1) \le d(t). \tag{2.35}$$

In other words, the chain only moves closer to stationarity as time passes.

(ii) *Let* μ *be any arbitrary starting distribution. Then*

$$\left\|\mu P^t - \pi\right\|_{\mathrm{TV}} \le \max_{x\in\Omega} \left\|P^t(x,\cdot) - \pi\right\|_{\mathrm{TV}} = d(t).$$

In other words, the distance to stationarity is maximised when the starting distribution is concentrated at a single point. Hence, it suffices to bound $d(t)$*.*

Proof. (i) Fix $x \in \Omega$ and $t \in \mathbb{N}$. Recall that π is stationary for P^t (Lemma 2.9 (ii)). By taking the starting distribution to be the point mass $\mu = \delta_x$, Proposition 2.21 (i) implies that

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

Since this holds for all $x \in \Omega$,

$$\max_{x\in\Omega} \left\|P^{t+1}(x,\cdot) - \pi\right\|_{\mathrm{TV}} \le \max_{x\in\Omega} \left\|P^t(x,\cdot) - \pi\right\|_{\mathrm{TV}}.$$

Hence, $d(t+1) \le d(t)$, as claimed.

(ii) Since $\sum_{x\in\Omega} \mu(x) = 1$, this is an immediate consequence of Proposition 2.21 (ii) by taking the maximum of $\left\|P^t(x,\cdot) - \pi\right\|_{\mathrm{TV}}$ over all $x \in \Omega$. □

Recall $d(t) = \max_{x\in\Omega} \left\|P^t(x,\cdot) - \pi\right\|_{\mathrm{TV}}$ and $\overline{d}(t) = \max_{x,y\in\Omega} \left\|P^t(x,\cdot) - P^t(y,\cdot)\right\|_{\mathrm{TV}}$ from (2.25) and (2.26). The next lemma establishes a relationship between $d(t)$ and $\overline{d}(t)$. This will be useful since $\overline{d}(t)$ is naturally compatible with the coupling technique, which is discussed in Section 3.1.

Lemma 2.23. *The functions* $d(t)$ *and* $\overline{d}(t)$ *satisfy*

$$d(t) \le \overline{d}(t) \le 2d(t), \qquad t \in \mathbb{N}. \tag{2.36}$$

Proof. Using the triangle inequality shows that

$$\sum_{z\in\Omega} |P^t(x,z) - P^t(y,z)| \le \sum_{z\in\Omega} |P^t(x,z) - \pi(z)| + \sum_{z\in\Omega} |P^t(y,z) - \pi(z)|.$$

Dividing both sides by $\frac{1}{2}$, and taking the maximum over $x, y \in \Omega$, implies that $\overline{d}(t) \le 2d(t)$.

Next, fix $x \in \Omega$. Recall that $\pi = \pi P^t$, or equivalently $\pi(z) = \sum_{y\in\Omega} \pi(y) P^t(y,z)$ for any $z \in \Omega$. Using this property, and the triangle inequality,

$$\left\|P^t(x,\cdot) - \pi\right\|_{\mathrm{TV}} = \frac{1}{2} \sum_{z\in\Omega} |P^t(x,z) - \pi(z)| \le \frac{1}{2} \sum_{z\in\Omega} \sum_{y\in\Omega} \pi(y) |P^t(x,z) - P^t(y,z)|.$$

By interchanging the sum, the right hand side is further upper bounded by

$$\sum_{y\in\Omega}\pi(y)\,\left\|P^t(x,\cdot)-P^t(y,\cdot)\right\|_{\mathrm{TV}}\le\max_{y\in\Omega}\left\|P^t(x,\cdot)-P^t(y,\cdot)\right\|_{\mathrm{TV}}.$$

This inequality follows from upper bounding the total variation distance by the maximum over y, and then using $\sum_{y\in\Omega}\pi(y)=1$. Next, taking the maximum over x shows that $d(t)\le\overline{d}(t)$, as desired. □

Remark 2.24. The total variation distance is the most widely used in the literature, and will be used in the rest of this thesis. We will mention some other reasonable distances. In practice, they are similar in that if one distance is close, then they all tend to be close [13, Section 6]. Some techniques may be naturally compatible with certain distances.

(i) The L^p $(1\le p<\infty)$ distance is the usual L^p norm induced by π:

$$\left\|\frac{P^t(x,\cdot)}{\mu}-\mathbf{1}\right\|_{p,\pi}^p:=\sum_{y\in\Omega}\left|\frac{P^t(x,y)}{\mu(y)}-1\right|^p\pi(y). \tag{2.37}$$

Here $\mathbf{1}$ denotes the function that is one for all $x\in\Omega$. This measures the relative density $P^t(x,\cdot)/\pi$, which converges to one as $t\to\infty$. When $p=1$, this leads to the usual total variation distance:

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

When $p=2$, this is the χ^2 distance (or equivalently, this is $\mathrm{Var}_\pi(P^t(x,\cdot)/\pi)$):

$$\chi(P^t(x,\cdot),\pi)=\sum_{y\in\Omega}\frac{(P^t(x,y)-\pi(y))^2}{\pi(y)}=\left\|\frac{P^t(x,\cdot)}{\mu}-\mathbf{1}\right\|_{2,\pi}^2.$$

These are discussed further in [46], which focuses more on analytic techniques. In particular, the L^2 distance is used more extensively (which provides an upper bound for the total variation distance by the Cauchy-Schwarz inequality).

(ii) The separation distance (introduced in [2]):

$$\mathrm{sep}(P^t(x,\cdot),\pi):=\max_{y\in\Omega}\left(1-\frac{P^t(x,y)}{\pi(y)}\right). \tag{2.38}$$

Note that this is not symmetric, and hence not a metric. Another example of use is in [15]. This is related to the concept of strong stationary times, which is the topic of Section 3.3.

Definition 2.25. The total variation distance will be used to measure the rate of convergence towards stationarity. We define $t_{\mathrm{mix}}(\epsilon):=\min\{t\in\mathbb{N}:d(t)\le\epsilon\}$ to be the **mixing time** of a Markov chain, with tolerance $0<\epsilon\le1$. When the argument is omitted, we take $\epsilon=1/4$ by convention, and write $t_{\mathrm{mix}}:=t_{\mathrm{mix}}(1/4)$.

The following lemma will be proved in Section 3.1 after coupling is introduced.

Lemma 2.26. *The function $\overline{d}(t)$, defined in (2.26), is submultiplicative.*

$$\overline{d}(s+t) \le \overline{d}(s)\,\overline{d}(t), \qquad \text{for all } s,t \in \mathbb{N}. \tag{2.39}$$

Corollary 2.27. *Let ℓ, $t \in \mathbb{N}$. Then*

(i) $d(\ell t) \le \overline{d}(\ell t) \le \overline{d}(t)^{\ell}$;

(ii) $d(\ell t_{\text{mix}}(\epsilon)) \le (2\epsilon)^{\ell}$;

(iii) $d(\ell t_{\text{mix}}) \le 2^{-\ell}$.

Proof. (i) is an immediate consequence of Lemma 2.23 and Lemma 2.26. (ii) follows from (i) with $t = t_{\text{mix}}(\epsilon)$, recalling that $\overline{d}(t_{\text{mix}}(\epsilon)) \le 2d(t_{\text{mix}}(\epsilon)) \le 2\epsilon$. Finally, (iii) follows from (ii) by taking $\epsilon = 1/4$. □

Remark 2.28. Let $\epsilon > 0$. By setting $2^{-\ell} = \epsilon$, which is equivalent to $\ell = \log_2(\epsilon^{-1})$, Corollary 2.27 (iii) implies that $t_{\text{mix}}(\epsilon) \le \lceil \log_2(\epsilon^{-1}) \rceil\, t_{\text{mix}}$. Therefore, while the choice of $\epsilon = 1/4$ in the definition of the mixing time is mostly arbitrary, a good bound on t_{mix} translates to a good bound on $t_{\text{mix}}(\epsilon)$.

To conclude this section, we will show that the mixing time of a random walk on a group (defined in Remark 2.7) is independent of the starting state because of symmetry.

Lemma 2.29. *Consider a random walk on a group G with increment distribution μ, which has uniform stationary distribution π. Then for any $g, h \in G$,*

$$\left\| P^t(g,\cdot) - \pi \right\|_{\text{TV}} = \left\| P^t(h,\cdot) - \pi \right\|_{\text{TV}}.$$

Therefore, $d(t) = \left\| P^t(\text{id},\cdot) - \pi \right\|_{\text{TV}}$, where id *is the identity element of G.*

Proof. Fix $g, h \in G$. For $x \in G$, consider $P^t(g,x)$ and $P^t(h, xg^{-1}h)$. To get from g to x in t steps means that we can write $z_t z_{t-1} \cdots z_1 g = x$, where $z_i \in G$. By multiplying both sides on the right by $g^{-1}y$, this is equivalent to $z_t z_{t-1} \cdots z_1 h = xg^{-1}h$. Both events have probability $\mu(z_1) \cdots \mu(z_t)$, and the sum of the probabilities of all such events by varying the symbols $z_1, \ldots, z_t$ gives $P^t(g,x)$ and $P^t(h, xg^{-1}h)$ respectively. Thus, these two probabilities are the same, and we can write

$$\frac{1}{2}\sum_{x\in G}\left|P^t(g,x) - \frac{1}{|\Omega|}\right| = \frac{1}{2}\sum_{x\in G}\left|P^t(h,xg^{-1}h) - \frac{1}{|\Omega|}\right| = \frac{1}{2}\sum_{\widetilde{x}\in G}\left|P^t(h,\widetilde{x}) - \frac{1}{|\Omega|}\right|,$$

where we relabelled $\widetilde{x} = xg^{-1}h$ in the last sum. By Proposition 2.19 (i), this shows that the two total variation distances are equal, as claimed. □

Chapter 3

Probabilistic Techniques for Analysing Mixing Times

From now on, all Markov chains considered will be irreducible and aperiodic unless otherwise specified. The previous chapter showed that such a chain will converge to a unique stationary distribution. We will now analyse the rate of convergence towards stationarity.

The analysis of mixing times is an area of much recent work and developments. In this chapter, we will describe the probabilistic methods of coupling (Section 3.1), path coupling (Section 3.2), and strong stationary times (Section 3.3) that have been used. The next chapter will focus on analytical methods.

The main reference for these two chapters remains the textbook by Levin, Peres and Wilmer [38]. For an extensive treatise on reversible Markov chains, see [4]. For references to some other techniques beyond the scope of this thesis, see [12] (group representations), and [46, 52] (analytical tools).

3.1 Coupling

This section will describe a probabilistic technique, called coupling, that can be used to bound the total variation distance. This is a powerful tool that is useful in more general situations than considered in this thesis (e.g. see [39] for broader applications within probability theory, and [50] for its application to chains with uncountable state spaces).

An interesting application of coupling is the coupling from the past method of Propp and Wilson [49, 48], although its details remain beyond the scope of this thesis. This algorithm allows for exact samples to be drawn from the stationary distribution, which separates the quality of output from the issue of efficiency.

Definition 3.1. (i) A **coupling** of two probability distributions μ and ν is a pair of random variables (X, Y) defined on a common probability space, which has marginal distributions μ and ν respectively.

(ii) A coupling of two stochastic processes $(Z_t^{(1)})_{t\in\mathbb{N}}$ and $(Z_t^{(2)})_{t\in\mathbb{N}}$ is a stochastic process $(X_t, Y_t)_{t\in\mathbb{N}}$ defined on a common probability space, such that (X_t) and (Y_t) are faithful copies of $(Z_t^{(1)})$ and $(Z_t^{(2)})$ respectively. Moreover, we assume that (X_t) and (Y_t) agree after they meet (or coalesce), by satisfying the **coupling condition**:

$$\text{if } X_s = Y_s, \text{ then } X_t = Y_t \quad \text{for } t \geq s. \tag{3.1}$$

If $(Z_t^{(1)})_{t\in\mathbb{N}}$ and $(Z_t^{(2)})_{t\in\mathbb{N}}$ are copies of a Markov chain with transition matrix P starting at x and y respectively, then we call this a **coupling of Markov chains**. We

write $\mathbb{P}_{xy}$ to denote the dependence of the coupling on the starting states. Note that this definition can naturally be generalised for the coupling of more than two processes.

Proposition 3.2. *Let μ and ν be two probability distributions on a finite state space Ω. Then*

$$\|\mu - \nu\|_{\mathrm{TV}} = \inf\left\{\mathbb{P}\left(X \neq Y\right) : (X, Y) \text{ is a coupling of } \mu \text{ and } \nu\right\}. \tag{3.2}$$

Moreover, there exists an ***optimal coupling*** *(X, Y), such that $\|\mu - \nu\|_{\mathrm{TV}} = \mathbb{P}\left(X \neq Y\right)$.*

Proof. Let (X, Y) be any coupling of μ and ν, and let $A \subseteq \Omega$. Note that $\mathbb{P}\left(X \in A\right)$ can be decomposed into $\mathbb{P}\left(X \in A,\, Y \in A\right) + \mathbb{P}\left(X \in A,\, Y \notin A\right)$. Similarly, $\mathbb{P}\left(Y \in A\right)$ can be decomposed into $\mathbb{P}\left(X \in A,\, Y \in A\right) + \mathbb{P}\left(X \notin A,\, Y \in A\right)$. Therefore,

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

Similarly, $\nu(A) - \mu(A) \leq \mathbb{P}\left(X \neq Y\right)$, and so $|\mu(A) - \nu(A)| \leq \mathbb{P}\left(X \neq Y\right)$. By taking the maximum over all $A \subseteq \Omega$ (recalling the definition of the total variation distance (2.24)),

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

Next, we will construct an optimal coupling (which is possible since Ω is finite). The goal is to force $X = Y$ as often as possible, whilst ensuring that their marginal distributions are correct. Recall that from (2.31),

$$\|\mu - \nu\|_{\mathrm{TV}} = \sum_{\substack{x \in \Omega \\ \mu(x) > \nu(x)}} [\mu(x) - \nu(x)] = \sum_{\substack{x \in \Omega \\ \mu(x) < \nu(x)}} [\nu(x) - \mu(x)]. \tag{3.3}$$

Consider the total probability mass common to μ and ν, which is equal to

$$\begin{aligned}\sum_{x \in \Omega} \min\{\mu(x), \nu(x)\} &= \sum_{\substack{x \in \Omega \\ \mu(x) = \nu(x)}} \mu(x) + \sum_{\substack{x \in \Omega \\ \mu(x) > \nu(x)}} \nu(x) + \sum_{\substack{x \in \Omega \\ \nu(x) > \mu(x)}} \mu(x) \\ &= 1 - \sum_{\substack{x \in \Omega \\ \mu(x) > \nu(x)}} (\mu(x) - \nu(x)) = 1 - \|\mu - \nu\|_{\mathrm{TV}}.\end{aligned}$$

Here $\sum_{x \in \Omega} \mu(x) = 1$ and (3.3) were used for the second and last equalities respectively.

We will now define an optimal coupling (X, Y). With probability $p := 1 - \|\mu - \nu\|_{\mathrm{TV}}$, set $X = Y = z$ together, with probability

$$\gamma_{XY}(z) = \frac{\min\{\mu(z), \nu(z)\}}{p}, \qquad z \in \Omega.$$

With probability $1 - p$, set $X = z$ and $Y = z'$ independently, with probabilities

$$\gamma_X(z) = \frac{\max\{\mu(z) - \nu(z), 0\}}{1 - p}, \qquad \gamma_Y(z') = \frac{\max\{\nu(z') - \mu(z'), 0\}}{1 - p}, \qquad z, z' \in \Omega,$$

respectively. In this case, the probability that X and Y are equal is zero since their common probability mass is exhausted. By construction, $\sum_{x\in\Omega}\gamma_{XY}(x) = 1$, and (3.3) shows that $\sum_{x\in\Omega}\gamma_X(x) = \sum_{y\in\Omega}\gamma_Y(y) = 1$. Hence, this is a valid joint distribution as

$$\sum_{x,y\in\Omega}\mathbb{P}\left(X=x, Y=y\right) = p\sum_{x\in\Omega}\gamma_{XY}(x) + (1-p)\sum_{x\in\Omega}\gamma_X(x)\sum_{y\in\Omega}\gamma_Y(y) = 1.$$

Moreover, $\mathbb{P}\left(X=x\right) = p\,\gamma_{XY}(x) + (1-p)\,\gamma_X(x) = \mu(x)$ shows that the marginal distribution for X is correct (and similarly for Y). Therefore, this is an optimal coupling such that $\mathbb{P}\left(X\neq Y\right) = 1-p = \|\mu-\nu\|_{\mathrm{TV}}$. □

The following important theorem allows us to bound the mixing time by the time it takes for a coupling of two copies of the chain to coalesce.

Theorem 3.3 (Coupling theorem)**.** *Let $(X_t, Y_t)_{t\in\mathbb{N}}$ be a coupling of Markov chains with finite state space Ω, and irreducible, aperiodic transition matrix P. Define the* ***coupling time*** *to be the first time that the two chains coalesce:*

$$\tau_{\mathrm{couple}} := \min\{t\in\mathbb{N} : X_t = Y_t\}. \tag{3.4}$$

Suppose that $X_0 = x$ and $Y_0 = y$. Then

(i) $\left\|P^t(x,\cdot) - P^t(y,\cdot)\right\|_{\mathrm{TV}} \leq \mathbb{P}_{xy}\left(\tau_{\mathrm{couple}} > t\right)$;

(ii) $d(t) \leq \max_{x,y\in\Omega}\mathbb{P}_{xy}\left(\tau_{\mathrm{couple}} > t\right)$.

Proof. (i) For any $t\in\mathbb{N}$, note that (X_t, Y_t) is a coupling of $P^t(x,\cdot)$ and $P^t(y,\cdot)$. Recall that any coupling remains together after the coupling time from (3.1). Therefore, the event $\{X_t \neq Y_t\}$ is equal to $\{\tau_{\mathrm{couple}} > t\}$. By Proposition 3.2,

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

(ii) This follows from combining $d(t)\leq\overline{d}(t)$ from Lemma 2.23 with (i). □

Recall that Lemma 2.26 asserted that $\overline{d}(t) = \max_{x,y\in\Omega}\left\|P^t(x,\cdot) - P^t(y,\cdot)\right\|_{\mathrm{TV}}$ is submultiplicative, i.e. $\overline{d}(s+t)\leq\overline{d}(s)\overline{d}(t)$. We can now prove this using coupling.

Proof of Lemma 2.26. Let $s,t\in\mathbb{N}$. Fix $x,y\in\Omega$. By Proposition 3.2, we can consider an optimal coupling of $P^s(x,\cdot)$ and $P^s(y,\cdot)$, such that

$$\|P^s(x,\cdot) - P^s(y,\cdot)\|_{\mathrm{TV}} = \mathbb{P}_{xy}\left(X_s\neq Y_s\right). \tag{3.5}$$

From time s onwards, run the two copies of the Markov chain independently, until the two chains coalesce. The distribution of $(X_t)_{t\in\mathbb{N}}$ at time $s+t$ is, for all $w\in\Omega$,

$$P^{s+t}(x,w) = \sum_{z\in\Omega}P^s(x,z)P^t(z,w) = \sum_{z\in\Omega}\mathbb{P}_x\left(X_s=z\right)P^t(z,w) = \mathbb{E}_{xy}\left[P^t(X_s,w)\right].$$

Similarly, the distribution of $(Y_t)_{t\in\mathbb{N}}$ at time $s+t$ is $P^{s+t}(y,w) = \mathbb{E}_{xy}\left[P^t(X_s,w)\right]$. Therefore, we can use the linearity of the expectation to write

$$P^{s+t}(x,w) - P^{s+t}(y,w) = \mathbb{E}_{xy}\left[P^t(X_s,w) - P^t(Y_s,w)\right].$$

By summing over all states and using (2.27), it follows that

$$\left\|P^{s+t}(x,\cdot) - P^{s+t}(y,\cdot)\right\|_{\mathrm{TV}} = \frac{1}{2}\sum_{w\in\Omega} \left|\mathbb{E}_{xy}\left[P^t(X_s,w) - P^t(Y_s,w)\right]\right|.$$

Since the absolute value of the expectation of a random variable is less than or equal to the expectation of the absolute value of the random variable, this is upper bounded by

$$\mathbb{E}_{xy}\left[\frac{1}{2}\sum_{w\in\Omega}\left|P^t(X_s,w) - P^t(Y_s,w)\right|\right] = \mathbb{E}_{xy}\left[\left\|P^t(X_s,\cdot) - P^t(Y_s,\cdot)\right\|_{\mathrm{TV}}\right].$$

If $X_s = Y_s$, then the two chains move together by the coupling condition (3.1), and thus the right hand side is zero. If $X_s \neq Y_s$, the total variation distance inside the expectation is bounded from above by its maximum over all possible X_s, Y_s, which is $\overline{d}(t)$. Hence,

$$\left\|P^{s+t}(x,\cdot) - P^{s+t}(y,\cdot)\right\|_{\mathrm{TV}} \leq \mathbb{E}_{xy}\left[\overline{d}(t)\,\mathbb{1}_{\{X_s\neq Y_s\}}\right] = \overline{d}(t)\,\mathbb{P}\left(X_s \neq Y_s\right)$$

Since (X_s, Y_s) is an optimal coupling, by (3.5),

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

Taking the maximum over $x, y \in \Omega$ shows that $\overline{d}(s+t) \leq \overline{d}(s)\,\overline{d}(t)$, as desired. □

Next, we will provide some examples of how the coupling technique can be used to obtain upper bounds for mixing times.

Example 3.4 (Random walk on the cycle)**.** Let $n \in \mathbb{Z}_+$ with $n \geq 2$. Consider a random walk on the group $\mathbb{Z}/n\mathbb{Z}$, or n-cycle, where $\Omega = \{0, 1, \ldots, n-1\}$, as illustrated in Figure 3.1.1. At each step, the chain moves to one of its neighbours uniformly at random.

Note that if n is even, then the chain has period two. Hence, we will consider the lazy version instead, where self-loops are added such that $P(x,x) = 1/2$ for all $x \in \Omega$. The other transition probabilities are given by $P(x,y) = 1/4$ if $|x-y| \equiv 1 \,(\mathrm{mod}\, n)$, and zero otherwise. Since this is a (reversible) random walk on a group, the uniform distribution π given by $\pi(g) = \frac{1}{n}$, $g \in G$ is the stationary distribution, by Remark 2.7.

Consider the following coupling $(X_t, Y_t)_{t\in\mathbb{N}}$ of two copies of the chain started at x and y respectively. With probability $1/2$, X_t chooses to move to the left or right with uniform probability, and Y_t remains at its current state. Otherwise, Y_t moves and X_t remains.

Consider the process $(D_t)_{t\in\mathbb{N}}$, such that D_t counts the clockwise distance between X_t and Y_t. This is also a Markov chain that takes values in $\{0, 1, 2, \ldots, n\}$. By the coupling

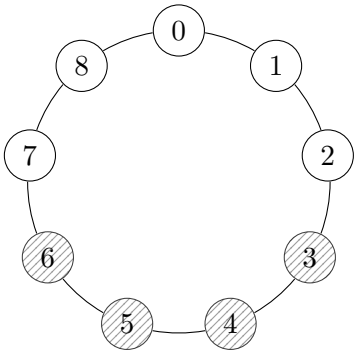

Figure 3.1.1: A random walk on the n-cycle (here $n = 9$). The shaded vertices correspond to the set A in Example 3.5.

condition (3.1), D_t is absorbed when it reaches 0 or n. Moreover, the coupling time τ_{couple} is exactly the first time that D_t is absorbed. If $0 < D_t < n$, then it increases or decreases by one at each step with equal probability $1/2$.

The coupling argument allows for D_t to be analysed instead. This reduces to the well-known Gambler's Ruin problem: starting with "wealth" $d := |x - y|$ (assuming $X_t \geq Y_t$ without loss of generality), a gambler wins or loses with equal probability, until they are either ruined ($D_t = 0$), or leave ($D_t = n$). This problem has been well-studied by the method of linear recurrence relations in [24, Chapter XIV, Section 3]. In particular, the expected duration of the game is $d(n - d)$. Note that $\max_{x,y\in\Omega} d(n-d) \leq \frac{n^2}{4}$.

Thus, combining the coupling theorem (Theorem 3.3 (ii)) and Markov's inequality (Theorem 1.1) shows that

$$d(t) \leq \max_{x,y\in\Omega} \mathbb{P}_{xy}\left(\tau_{\text{couple}} > t\right) \leq \max_{x,y\in\Omega} \frac{\mathbb{E}_{xy}\left[\tau_{\text{couple}}\right]}{t} = \frac{n^2}{4t}.$$

Hence $t_{\text{mix}} \leq n^2$, so that $t_{\text{mix}} = O(n^2)$.

Example 3.5 (Random walk on the cycle II)**.** We will provide a matching lower bound for Example 3.4, by showing that the mixing time of a lazy random walk on the n-cycle satisfies $t_{\text{mix}} \geq \lfloor cn^2 \rfloor$ for some constant $c > 0$. Assume that $n \geq 4$ to avoid trivialities.

Coupling is not used to find this lower bound. Instead, the idea is to use the definition of the total variation distance (2.24), and to choose a particular set $A \subset \Omega$ and starting state $x_0 \in \Omega$ such that $|P^t(x_0, A) - \pi(A)|$ is large. Define $A = \{\lceil \frac{n}{4} \rceil, \lceil \frac{n}{4} \rceil + 1, \ldots, \lfloor \frac{3n}{4} \rfloor\}$, which has at least $\lfloor \frac{n}{2} \rfloor$ elements (see the previous Figure 3.1.1). Hence, under the stationary distribution, the probability that the random walk is in A is

$$\pi(A) = \sum_{x\in A} \frac{1}{n} \geq \frac{1}{n} \cdot \left\lfloor \frac{n}{2} \right\rfloor \geq \frac{1}{n} \cdot \frac{n-1}{2} = \frac{1}{2} - \frac{1}{2n}.$$

Let $(X_t)_{t\in\mathbb{N}}$ be a lazy random walk on the n-cycle starting at 0. Consider the process $(Z_t)_{t\in\mathbb{N}}$, which counts the net clockwise steps made by X_t. Both chains remain still together, and when X_t moves clockwise (resp. anti-clockwise), Z_t increases (resp. decreases)

by 1. Thus, Z_t is a lazy random walk on $\mathbb{Z}$, which can be written $Z_t = \sum_{s=1}^{t} Y_s$, where each Y_s is an independent and identically distributed increment, taking value 0 w.p. $1/2$, and values -1 or 1 w.p. $1/4$ each. Hence, $\mathbb{E}_0[Z_t] = 0$, and $\mathrm{Var}_0(Z_t) = \frac{t}{2}$.

Note that $\lfloor \frac{3n}{4} \rfloor - n = -\lceil \frac{n}{4} \rceil$. Therefore, if X_t is in A, then the net number of clockwise (or anti-clockwise) steps it has taken must necessarily be at least $\lceil \frac{n}{4} \rceil$. Hence,

$$P^t(0, A) \leq \mathbb{P}_0\left(|Z_t| \geq \left\lceil \frac{n}{4} \right\rceil\right). \tag{3.6}$$

By Chebyshev's inequality (Theorem 1.2),

$$\mathbb{P}_0\left(|Z_t| \geq \left\lceil \frac{n}{4} \right\rceil\right) \leq \frac{t}{2\lceil n/4 \rceil^2} \leq \frac{8t}{n^2}. \tag{3.7}$$

Therefore, if $t \leq \lfloor \frac{n^2}{32} - \frac{n}{16} \rfloor$, (3.6) and (3.7) show that $P^t(0, A) \leq \frac{1}{4} - \frac{1}{2n}$. Hence,

$$d(t) \geq \pi(A) - P^t(0, A) \geq \frac{1}{2} - \frac{1}{2n} - \frac{1}{4} + \frac{1}{2n} = \frac{1}{4}.$$

Hence, $t_{\mathrm{mix}} \geq \lfloor \frac{n^2}{32} - \frac{n}{16} \rfloor$. Since $n \geq 4$, $\frac{n}{16} \leq \frac{n^2}{64}$, and so this bound can be simplified to $t_{\mathrm{mix}} \geq \lfloor \frac{n^2}{64} \rfloor$ (which, in fact, holds for all $n \in \mathbb{Z}_+$).

Example 3.6 (Random walk on the hypercube)**.** Let $n \in \mathbb{Z}_+$ with $n \geq 2$. Consider a random walk on $\Omega = (\mathbb{Z}/2\mathbb{Z})^n$, which is the set of all bitstrings (i.e. sequences of 0s and 1s) of length n. The chain moves between the 2^n vertices of the n-dimensional hypercube by choosing uniformly at random from its neighbours at each step (see Figure 3.1.2).

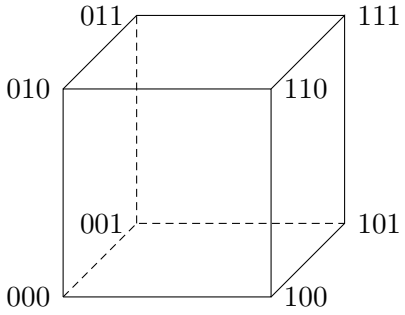

Figure 3.1.2: A random walk on the n-dimensional hypercube (here $n = 3$).

Note that this chain has period two, since each move flips the parity of the sum of the bits of the current state. Hence, we will consider the lazy version instead, with $P(\mathbf{x}, \mathbf{x}) = 1/2$, $\mathbf{x} \in \Omega$. The other transition probabilities are given by $P(\mathbf{x}, \mathbf{y}) = 1/(2n)$ if $\mathbf{x}$ and $\mathbf{y}$ differ by exactly one bit, and zero otherwise. As a random walk on a group, the stationary distribution is the uniform distribution $\pi(\mathbf{x}) = 2^{-n}$, $\mathbf{x} \in \Omega$, by Remark 2.7.

Consider the following coupling $(X_t, Y_t)_{t \in \mathbb{N}}$ of two copies of the chain, started at $\mathbf{x}$ and $\mathbf{y}$ respectively. Choose one of the n bits uniformly at random, and change it to either 0 or 1 with equal probability for both chains. Individually, each chain moves to an adjacent

vertex if the selected bit is changed with probability $\frac{1}{2n}$, and otherwise remains still with probability $n \times \frac{1}{2n} = \frac{1}{2}$. Hence, the marginals are correct.

Observe that once a bit is selected, both chains will agree on that particular bit going forwards by the coupling condition (3.1). This reduces to the classical Coupon Collector's problem, where selecting a bit is the same as "collecting a coupon". (This is extensively studied in Feller [24].) Let τ_{coupon} be the time it takes for all n bits to have been selected at least once. Since coalescence is guaranteed at time τ_{coupon}, it follows that $\tau_{\text{couple}} \leq \tau_{\text{coupon}}$, and hence $\mathbb{P}_{\mathbf{xy}}(\tau_{\text{couple}} > t) \leq \mathbb{P}_{\mathbf{xy}}(\tau_{\text{coupon}} > t)$.

Let B_i denote the event that the ith bit has not been selected by time t. Since the time for the ith bit to be selected is geometrically distributed, $\mathbb{P}(B_i) = \left(1 - \frac{1}{n}\right)^t$. Let $t = \lceil n \log n + cn \rceil$, where c is a fixed real number. Then by the simple union bound,

$$\mathbb{P}_{\mathbf{xy}}(\tau_{\text{coupon}} > t) = \mathbb{P}\left(\bigcup_{i=1}^{n} B_i\right) \leq \sum_{i=1}^{n} \mathbb{P}(B_i) = n\left(1 - \frac{1}{n}\right)^t \leq ne^{-\frac{t}{n}} = e^{-c}. \tag{3.8}$$

Therefore, $d(t) \leq e^{-c}$ by the coupling theorem (Theorem 3.3 (ii)). If $c = \log 4$, then $d(t) \leq \frac{1}{4}$. Hence, $t_{\text{mix}} \leq \lceil n \log n + n \log 4 \rceil$, so that $t_{\text{mix}} = O(n \log n)$.

It is shown in [23] that the asymptotic distribution of τ_{coupon} is

$$\lim_{n \to \infty} \mathbb{P}(\tau_{\text{couple}} > n \log n + cn) = 1 - \exp(-e^{-c}), \qquad c \in \mathbb{R}. \tag{3.9}$$

If c is not small, then $\exp(-e^{-c}) \approx 1 - e^{-c}$. Thus, the upper bound in (3.8) is quite sharp.

3.2 Path coupling

In this section, we discuss an extension of the coupling technique. The idea is to abstract away from considering all pairs of states in achieving a coupling bound, and instead only consider pairs of states that are adjacent in some path. This minimises the combinatorial complexity, and has led to better bounds in some cases.

Path coupling was first introduced by Bubley and Dyer [7] to show that several chains involved in hard combinatorial problems are rapidly mixing. The main result of this section is the path coupling theorem (Theorem 3.12). A fascinating connection to the transportation metric will be uncovered in developing the theory, following [38]

Definition 3.7. Suppose that Ω is equipped with a metric ρ. The **transportation metric** (also known as the **Wasserstein distance**) between two distributions μ and ν on Ω is

$$W_\rho(\mu, \nu) := \inf\left\{\mathbb{E}[\rho(X, Y)] : (X, Y) \text{ is a coupling of } \mu \text{ and } \nu\right\}. \tag{3.10}$$

This allows a metric on Ω to be lifted to distributions on Ω. Equivalently, it can be written in terms of the distribution functions of the associated random variables:

$$W_\rho(\mu, \nu) = \inf_{\phi}\left\{\sum_{(x,y) \in \Omega \times \Omega} \rho(x, y)\phi(x, y) : q_1(\phi) = \mu,\ q_2(\phi) = \nu\right\}. \tag{3.11}$$

The infimum is taken over all joint distributions ϕ on $\Omega \times \Omega$. The function q_1 (resp. q_2), where $q_1(\phi) = \phi(\cdot, \Omega) = \sum_{y \in \Omega} \phi(\cdot, y)$, is the projection onto the first (resp. second) coordinate. (These correspond to the marginals of X and Y.)

Remark 3.8. (i) Why is this known as the transportation metric? Suppose that Ω represents "locations", and ρ represents the "costs" of moving resources from one site to another. The distributions of the starting and ending resources in each site are represented by μ and ν respectively. Thus, a coupling is a particular strategy of transporting the resources, and the transportation metric gives the minimal cost.

(ii) Suppose that ρ is the discrete metric: $\rho(x, y) = \mathbb{1}_{\{x \neq y\}}$. Then, recalling (3.2), $W_\rho(\mu, \nu) = \inf \left\{ \mathbb{P}(X \neq Y) : (X, Y) \text{ is a coupling of } \mu \text{ and } \nu \right\} = \|\mu - \nu\|_{\mathrm{TV}}$. Thus, the transportation metric generalises the total variation distance.

Lemma 3.9. (i) *Let μ, ν be distributions on Ω. Then there exists an optimal coupling (X_*, Y_*) of μ and ν such that $\mathbb{E}[\rho(X_*, Y_*)] = W_\rho(\mu, \nu)$.*

(ii) *The transportation metric is a metric on the space of probability distributions on Ω.*

Proof. (i) Some classical results from analysis will be assumed for this proof (see [51] for a classical reference). In this part, we will identify each distribution on $\Omega \times \Omega$ with a point in the following probability simplex in $\mathbb{R}^{|\Omega|^2}$ (a $(|\Omega|^2 - 1)$-dimensional polytope):

$$A = \left\{ (x_{ij})_{i,j=1}^{|\Omega|} \in \mathbb{R}^{|\Omega|^2} : \sum_{i,j=1}^{|\Omega|} x_{ij} = 1, \ x_{ij} \geq 0 \text{ for all } 1 \leq i, j \leq |\Omega| \right\}.$$

This is a closed and bounded subset of $\mathbb{R}^{|\Omega|^2}$, and hence it is compact by the Heine-Borel Theorem. The projection onto the first coordinate corresponds to $q_1 : \mathbb{R}^{|\Omega|^2} \to \mathbb{R}^{|\Omega|}$, $(x_{ij})_{i,j=1}^{|\Omega|} \mapsto (\sum_{j=1}^{|\Omega|} x_{ij})_{i=1}^{|\Omega|}$. As the sum of continuous projections from $\mathbb{R}^{|\Omega|^2}$ to $\mathbb{R}^{|\Omega|}$, q_1 is continuous. Similarly, the projection onto the second coordinate q_2 is continuous.

Let B be the set of distributions on $\Omega \times \Omega$ that project down to μ and ν on the first and second coordinates. Then B is a closed subset of A, since it is the intersection of the closed sets $q_1^{-1}(\{\mu\})$ and $q_2^{-1}(\{\nu\})$), and hence it is also compact. Note that the following function is continuous (as the sum of projections from $\mathbb{R}^{|\Omega|^2}$ to $\mathbb{R}$ multiplied by scalars):

$$B \to \mathbb{R}, \quad \phi \mapsto \sum_{(x,y) \in \Omega \times \Omega} \rho(x, y) \phi(x, y).$$

Therefore, by the Extreme Value Theorem, there exists a distribution ϕ_* that achieves the infimum in (3.11). This corresponds to an optimal coupling (X_*, Y_*).

(ii) The symmetry of W_ρ follows clearly from the symmetry of ρ. Next, we show that W_ρ is positive definite. If $\mu = \nu$, then we can construct the coupling with distribution function $\phi(x, x) = \mu(x) = \nu(x)$ for all $x \in \Omega$, and $\phi(x, y) = 0$ for all $x \neq y$. Thus $\mathbb{E}[\rho(X, Y)] = \sum_{x \in \Omega} \rho(x, x) \phi(x, x) = 0$, and so $W_\rho(\mu, \nu) = 0$.

Conversely, suppose that $W_\rho(\mu,\nu)=0$. From (i), we can choose an optimal coupling ϕ_* of μ and ν. Then $\sum_{x\in\Omega}\sum_{y\in\Omega}\rho(x,y)\phi_*(x,y)=0$, which implies that $\rho(x,y)\phi_*(x,y)=0$ for all $x,y\in\Omega$. If $\phi_*(x,y)>0$, then $\rho(x,y)=0$ and hence $x=y$. Thus, $\phi_*(x,y)=0$ for all $x\neq y$. Therefore, since ϕ_* projects to μ and ν on its first and seconds coordinates respectively, $\mu(x)=\sum_{y\in\Omega}\phi(x,y)=\phi(x,x)=\sum_{z\in\Omega}\phi(z,x)=\nu(x)$, and so $\mu=\nu$.

Finally, it remains to show that the triangle inequality holds. Let μ, ν and η be distributions on Ω. From (i), let (X,Z) be an optimal coupling of μ and η with joint distribution ϕ, and let (Y,Z) be an optimal coupling of η and ν with joint distribution ψ. Let

$$r(x,y,z) \coloneqq \frac{\phi(x,z)\psi(y,z)}{\eta(z)}.$$

This is in fact a joint distribution function of (X,Y,Z). Moreover, r projects onto ϕ on its first and third coordinates (i.e. (X,Z)), which can be checked by calculating

$$\sum_{y\in\Omega} r(x,y,z) = \frac{\phi(x,z)}{\eta(z)}\sum_{y\in\Omega}\xi(y,z) = \frac{\phi(x,z)}{\eta(z)}\,\eta(z) = \phi(x,z).$$

Similarly, r projects onto ψ on its second and third coordinates (i.e. (Y,Z)). Its projection onto its first and second coordinates is a coupling of μ and ν, which we denote by (X,Y).

Recall that W_ρ is defined as an infimum (3.10). Hence, taking expectations with respect to r implies that $W_\rho(\mu,\nu)\le\mathbb{E}\left[\rho(X,Y)\right]$. Moreover, recall that (X,Z) and (Y,Z) are optimal couplings. By using the triangle inequality for ρ, and the linearity of expectations,

$$\mathbb{E}\left[\rho(X,Y)\right]\le\mathbb{E}\left[\rho(X,Z)\right]+\mathbb{E}\left[\rho(Y,Z)\right]=W_\rho(\mu,\eta)+W_\rho(\eta,\nu). \tag{3.12}$$

Hence, we conclude that $W_\rho(\mu,\nu)\le W_\rho(\mu,\eta)+W_\rho(\eta,\nu)$, as desired. □

Let $G=(\Omega,E)$ be a connected graph on the state space Ω of a Markov chain. The adjacent states in this graph are not necessarily the same as the permissible transitions of the chain, however they usually are. Let ℓ be a length function that assigns length $\ell(x,y)\ge 1$ to each edge $\{x,y\}\in E$, and $\ell(x,x)\coloneqq 0$.

Definition 3.10. Define a **path** ξ from x to y to be a sequence $x=x_0,\,x_1,\,\ldots,\,x_r=y$ of states, such that $\{x_{i-1},x_i\}$ is an edge for $i=1,2,\ldots,r$. The **length** of the path is defined to be $\sum_{i=1}^{r}\ell(x_{i-1},x_i)$. (The length of a path from x to x is defined to be $\ell(x,x)=0$.) Then the **path metric** on Ω is defined by

$$\rho(x,y)=\min\big\{\text{length}(\xi):\xi\text{ is a path from }x\text{ to }y\big\}. \tag{3.13}$$

It is not too difficult to check that this is indeed a metric on Ω.

The transportation metric allows the path metric ρ to be lifted as a metric W_ρ on the space of distributions on Ω. The next lemma connects W_ρ to the total variation distance.

Lemma 3.11. *Let ρ be the path metric on Ω. Then for any distributions μ and ν on Ω,*

$$\|\mu - \nu\|_{\mathrm{TV}} \leq W_\rho(\mu, \nu). \tag{3.14}$$

Proof. Since $\ell(x, y) \geq 1$ for $x \neq y$, it follows that $\rho(x, y) \geq \mathbb{1}_{\{x \neq y\}}$ for all $x, y \in \Omega$. Let (X, Y) be an optimal coupling of μ and ν for W_ρ by Lemma 3.9. Then

$$\mathbb{P}(X \neq Y) = \mathbb{E}\left[\mathbb{1}_{\{X \neq Y\}}\right] \leq \mathbb{E}\left[\rho(X, Y)\right] = W_\rho(\mu, \nu).$$

Since $\|\mu - \nu\|_{\mathrm{TV}} \leq \mathbb{P}(X \neq Y)$ from (3.2), we are done. □

Theorem 3.12 (Path coupling theorem)**.** *Consider a Markov chain with transition matrix P, and finite state space Ω with associated graph $G = (\Omega, E)$. Let ρ be the path metric defined in (3.13). Suppose that for each edge $\{x, y\} \in E$, there exists a coupling (X, Y) of the distributions $P(x, \cdot)$ and $P(y, \cdot)$ that satisfies the "contraction condition"*

$$\mathbb{E}_{xy}\left[\rho(X, Y)\right] \leq e^{-\alpha}\rho(x, y), \tag{3.15}$$

for some $\alpha > 0$. Then for any two probability distributions μ and ν on Ω,

$$W_\rho(\mu P, \nu P) \leq e^{-\alpha} W_\rho(\mu, \nu). \tag{3.16}$$

Proof. Fix arbitrary states $x, y \in \Omega$. Let $x = x_0, x_1, \ldots, x_r = y$ be a path of minimum length joining x and y. Then $\{x_{i-1}, x_i\} \in E$ for $i = 1, 2, \ldots, r$, and $\sum_{i=1}^r \ell(x_{i-1}, x_i) = \rho(x, y)$. By the triangle inequality (Lemma 3.9) and the contraction assumption (3.15),

$$W_\rho(P(x, \cdot), P(y, \cdot)) \leq \sum_{i=1}^r W_\rho(P(x_{i-1}, \cdot), P(x_i, \cdot)) \leq e^{-\alpha} \sum_{i=1}^r \ell(x_{i-1}, x_i) = e^{-\alpha}\rho(x, y). \tag{3.17}$$

Thus, (3.16) holds for the special case of $\mu = \delta_x$, $\nu = \delta_y$. For each $x, y \in \Omega$, by Lemma 3.9, we can find an optimal coupling θ_{xy} of $P(x, \cdot)$ and $P(y, \cdot)$, such that

$$W_\rho(P(x, \cdot), P(y, \cdot)) = \sum_{u, v \in \Omega} \rho(u, v)\theta_{xy}(u, v). \tag{3.18}$$

Similarly, we can find an optimal coupling ϕ of the two distributions μ and ν, such that

$$W_\rho(\mu, \nu) = \sum_{x, y \in \Omega} \rho(x, y)\phi(x, y). \tag{3.19}$$

Now, we claim that the following is a coupling of μP and νP:

$$\theta = \sum_{x, y \in \Omega} \phi(x, y)\theta_{xy}. \tag{3.20}$$

We can check that the projection of θ onto the first coordinate is μP. Recall that the projection of θ_{xy} onto the first coordinate is $P(x,\cdot)$, and that the projection of ϕ onto the first coordinate is μ. Hence, interchanging the order of summation shows that

$$\sum_{v\in\Omega}\theta(u,v) = \sum_{x,y\in\Omega}\phi(x,y)\left(\sum_{v\in\Omega}\theta_{xy}(u,v)\right) = \sum_{x,y\in\Omega}\phi(x,y)P(x,u)$$
$$= \sum_{x\in\Omega}\left(\sum_{y\in\Omega}\phi(x,y)\right)P(x,u) = \sum_{x\in\Omega}\mu(x)P(x,u) = \mu P(u).$$

A similar calculation shows that the projection of θ onto the second coordinate is νP. Thus, θ is indeed a coupling of μP and νP. Therefore, by (3.10),

$$W_\rho(\mu P, \nu P) \le \sum_{u,v\in\Omega}\rho(u,v)\theta(u,v). \tag{3.21}$$

By putting in the definition of θ from (3.20), interchanging the order of summation, and using (3.18), the right hand side of (3.21) is equal to

$$\sum_{x,y\in\Omega}\left(\sum_{u,v\in\Omega}\rho(u,v)\theta_{xy}(u,v)\right)\phi(x,y) = \sum_{x,y\in\Omega}W_\rho(P(x,\cdot),P(y,\cdot))\,\phi(x,y).$$

Finally, using (3.17), and then (3.19), shows that this upper bound for $W_\rho(\mu P,\nu P)$ is less than or equal to

$$e^{-\alpha}\sum_{x,y\in\Omega}\rho(x,y)\phi(x,y) = e^{-\alpha}W_\rho(\mu,\nu).$$

Hence, $W_\rho(\mu P,\nu P)\le e^{-\alpha}W_\rho(\mu,\nu)$, as desired. □

Corollary 3.13. *Assume that the same conditions as Theorem 3.12 hold. Suppose that the Markov chain has stationary distribution π. Let the diameter of the associated graph be* $\operatorname{diam}(\Omega) := \max_{x,y\in\Omega}\rho(x,y)$. *Then*

$$d(t) = \max_{x\in\Omega}\left\|P^t(x,\cdot)-\pi\right\|_{\mathrm{TV}} \le e^{-\alpha t}\operatorname{diam}(\Omega). \tag{3.22}$$

Consequently,

$$t_{\mathrm{mix}}(\epsilon) \le \left\lceil \frac{1}{\alpha}\left(\log(\operatorname{diam}(\Omega)) + \log(\epsilon^{-1})\right)\right\rceil. \tag{3.23}$$

Proof. By iterating the path coupling theorem (Theorem 3.12),

$$W_\rho(\mu P^t,\nu P^t) \le e^{-\alpha}W_\rho(\mu P^{t-1},\nu P^{t-1}) \le \cdots \le e^{-\alpha t}W_\rho(\mu,\nu).$$

Note that by (3.11), $W_\rho(\mu,\nu)$ is a weighted average of $\rho(x,y)$ over $x,y\in\Omega$. Thus, replacing each $\rho(x,y)$ with its maximum shows that $W_\rho(\mu,\nu)\le\max_{x,y\in\Omega}\rho(x,y)$. Setting

$\mu = \delta_x$ and $\nu = \pi$, and then using Lemma 3.11, shows that

$$\left\|P^t(x,\cdot) - \pi\right\|_{\mathrm{TV}} \leq e^{-\alpha t}\mathrm{diam}(\Omega),$$

uniformly over $x \in \Omega$, which is (3.22). Next, setting t equal to the right hand side of (3.23) shows that $d(t) \leq \epsilon$, as desired. □

Remark 3.14. The path coupling theorem as stated only applies when the contraction constant $\alpha > 0$ in (3.15). However, if $\alpha = 0$ (i.e. the expected path metric is non-increasing), then we may still be able to say something useful. In particular, if the probability that $\rho(X, Y)$ is non-zero is bounded below by some $\beta > 0$, then β can be used to upper bound the mixing time by a coupling argument (see [21, Theorem 2.1] or [7]).

An example of path coupling will be presented in Lemma 5.4 of Chapter 5, to show that the Glauber dynamics for sampling colourings of certain graphs is rapidly mixing.

3.3 Strong stationary times

In this section, we will describe how the method of strong stationary times can be used to upper bound mixing times. The idea of strong stationary times was introduced by Aldous and Diaconis in [2], and is further related to coupling and Fourier analysis in [3].

Recall from Section 2.3 that a stopping time for the Markov chain $(X_t)_{t\in\mathbb{N}}$ is a random variable τ taking values in $\mathbb{N} \cup \{+\infty\}$, such that the event $\{\tau = t\}$ is determined by $X_0, X_1, \ldots, X_t$ for every $t \in \mathbb{N}$.

Definition 3.15. A **stationary time** τ is a stopping time, possibly depending on the starting state x, such that X_τ is distributed as π:

$$\mathbb{P}_x\left(X_\tau = y\right) = \pi(y), \qquad \text{for all } y \in \Omega. \tag{3.24}$$

If we further require that X_τ is independent of τ, such that

$$\mathbb{P}_x\left(X_\tau = y, \tau = t\right) = \pi(y)\,\mathbb{P}_x\left(\tau = t\right), \qquad \text{for all } y \in \Omega,\, t \in \mathbb{N}, \tag{3.25}$$

then we say that τ is a **strong stationary time**.

Recall the definition of the separation distance (2.38):

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

Let $s(t) \coloneqq \max_{x\in\Omega}\mathrm{sep}(P^t(x,\cdot),\pi)$.

The following lemma shows that strong stationary times are connected with the separation distance, analogous to how coupling times are connected to the total variation distance (Theorem 3.3 (ii)).

Lemma 3.16. *Consider a Markov chain $\mathcal{M}$ with finite state space Ω and transition matrix P, starting at $x \in \Omega$. If τ is a strong stationary time for $\mathcal{M}$, then*

$$\operatorname{sep}(P^t(x,\cdot),\pi) \leq \mathbb{P}_x(\tau > t). \tag{3.26}$$

Proof. First, observe that for any $y \in \Omega$ and $t \in \mathbb{N}$,

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

using the strong stationary property. Since π is stationary for P, it follows that $\pi P^{t-s} = \pi$, and so $\sum_{z\in\Omega}\pi(z)P^{(t-s)}(z,y) = \pi(y)$. Hence, this shows that

$$\mathbb{P}_x(X_t = y, \tau \leq t) = \pi(y)\mathbb{P}_x(\tau \leq t). \tag{3.27}$$

Fix $y \in \Omega$. Note that $P^t(x,y) = \mathbb{P}_x(X_t = y) \geq \mathbb{P}_x(X_t = y, \tau \leq t)$, which is equal to $\pi(y)\mathbb{P}_x(\tau \leq t)$ by (3.27). Hence,

$$1 - \frac{P^t(x,y)}{\pi(y)} \leq 1 - \frac{\pi(y)\mathbb{P}_x(\tau \leq t)}{\pi(y)} = \mathbb{P}_x(\tau > t). \tag{3.28}$$

By taking the maximum over all $y \in \Omega$, the left hand side is equal to $\operatorname{sep}(P^t(x,\cdot),\pi)$. □

Remark 3.17. If the event $\{X_t = y\}$ implies that $\{\tau \leq t\}$, then the inequality in (3.28) is actually an equality. In that case, we say that y is a halting state for τ, and $\operatorname{sep}(P^t(x,\cdot), y) = \mathbb{P}_x(\tau > t)$. (This is analogous to an optimal coupling.)

The separation distance provides an upper bound for the total variation distance.

Lemma 3.18. *Continuing the notation used in Lemma 3.16, we have*

$$\left\|P^t(x,\cdot) - \pi\right\|_{\mathrm{TV}} \leq \operatorname{sep}(P^t(x,\cdot),\pi). \tag{3.29}$$

In particular, taking the maximum over all $x \in \Omega$ implies that $d(t) \leq s(t)$.

Proof. Using (2.31),

$$\left\|P^t(x,\cdot) - \pi\right\|_{\mathrm{TV}} = \sum_{\substack{y\in\Omega \\ \pi(y)>P^t(x,y)}} [\pi(y) - P^t(x,y)] = \sum_{\substack{y\in\Omega \\ \pi(y)>P^t(x,y)}} \pi(y)\left[1 - \frac{P^t(x,y)}{\pi(y)}\right]$$

Since $\pi(y)$ sums to one or less, $\left\|P^t(x,\cdot) - \pi\right\|_{\mathrm{TV}}$ is upper bounded by $\max_{y\in\Omega}\left[1 - \frac{P^t(x,y)}{\pi(y)}\right]$. □

Theorem 3.19. *Let τ be a strong stationary time for a Markov chain with finite state space Ω, and an irreducible and aperiodic transition matrix P. Then*

$$\left\|P^t(x,\cdot)-\pi\right\|_{\mathrm{TV}} \leq \mathbb{P}_x\left(\tau > t\right). \tag{3.30}$$

In particular, taking the maximum over all $x \in \Omega$ implies that $d(t) \leq \max_{x\in\Omega} \mathbb{P}_x\left(\tau > t\right)$.

Proof. This immediately follows from Lemma 3.18 and Lemma 3.16. □

We will use the method of strong stationary times to provide an upper bound on the mixing time of the well-known riffle shuffle. First, we will show, in a precise sense, that a random walk on a group has the same mixing time if it is "run backwards".

Definition 3.20. (i) The **time reversal** of an irreducible Markov chain with transition matrix P and stationary distribution π is the Markov chain with transition matrix $\widehat{P}$, where $\widehat{P}(x,y) = \frac{\pi(y)}{\pi(x)}P(y,x)$.

(ii) The time reversal of a random walk on a group G with increment distribution μ (defined in Remark 2.7) is the random walk on G with increment distribution $\widehat{\mu}$, where $\widehat{\mu}(g) = \mu(g^{-1})$.

Lemma 3.21. *Consider a random walk on a group G with transition matrix P induced by the increment distribution μ, and with uniform stationary distribution π. Let $\widehat{P}$ be the transition matrix of its time reversal induced by the increment distribution $\widehat{\mu}$. Then*

$$\left\|P^t(\mathrm{id},\cdot)-\pi\right\|_{\mathrm{TV}} = \left\|\widehat{P}^t(\mathrm{id},\cdot)-\pi\right\|_{\mathrm{TV}}, \qquad \text{for any } t \in \mathbb{N}.$$

In particular, the random walk on a group and its time reversal have the same mixing time (recall that the starting state does not matter by Lemma 2.29).

Proof. Fix $g \in G$. For the original chain to get from id to g in t steps means that we can write $z_t z_{t-1} \cdots z_1 = g$, where $z_i \in G$. The probability of this event is $\mu(z_1)\cdots\mu(z_t)$. By taking the inverse of both sides, this is equivalent to $z_1^{-1}z_2^{-1}\cdots z_t^{-1} = g^{-1}$. This describes the time reversal moving from id to g^{-1} in t steps, with probability $\widehat{\mu}(z_t^{-1})\cdots\widehat{\mu}(z_1^{-1})$, which is equal to $\mu(z_1)\cdots\mu(z_t)$. Summing the probabilities of all such events by varying the symbols $z_1,\ldots,z_t$ shows that $P^t(\mathrm{id},g) = \widehat{P}^t(\mathrm{id},g^{-1})$. Hence,

$$\frac{1}{2}\sum_{g\in G}\left|P^t(\mathrm{id},g)-\frac{1}{|\Omega|}\right| = \frac{1}{2}\sum_{g\in G}\left|\widehat{P}^t(\mathrm{id},g^{-1})-\frac{1}{|\Omega|}\right| = \frac{1}{2}\sum_{\widetilde{g}\in G}\left|\widehat{P}^t(\mathrm{id},\widetilde{g})-\frac{1}{|\Omega|}\right|,$$

relabelling $\widetilde{g} = g^{-1}$ in the last sum. Hence, the total variation distances are equal. □

Example 3.22 (Riffle shuffle)**.** The Gilbert-Shannon-Reeds model describes a mathematically precise way of riffle shuffling a deck of cards (see [6] for more details). This can be viewed as a random walk on the symmetric group S_n.

Cut the deck into two packets according to a Binomial$(n, \frac{1}{2})$ distribution. Interleave the packets, where a card is dropped from one of the packets with probability proportional to the number of cards remaining, until there are no cards remaining (see Figure 3.3.1). That is, if there are a cards in the left packet and b cards in the right packet, then the next card is dropped from the left packet with probability $\frac{a}{a+b}$, and the right packet otherwise.

$$[1]\ [2]\ [3]\ [4]\ [5]\ [6]\ [7]\ [8] \longrightarrow \begin{matrix}[1]\ [2]\ [3] \\ [4]\ [5]\ [6]\ [7]\ [8]\end{matrix} \longrightarrow [1]\ [4]\ [5]\ [2]\ [6]\ [7]\ [3]\ [8]$$

Figure 3.3.1: A riffle shuffle – the deck is cut into two packets and then interleaved.

A rising sequence is defined to be a maximal subset of an arrangement of cards consisting of successive face values when read in order. For example, reading the arrangement 14526738 from left to right shows that there are two rising sequences 123 and 45678 interleaved together. The increment distribution μ generates permutations characterised by the number of rising sequences, and is given by

$$\mu(\sigma) = \begin{cases} \frac{n+1}{2^n} & \text{if } \sigma = \mathrm{id}, \\ \frac{1}{2^n} & \text{if } \sigma \text{ has exactly two rising sequences}, \\ 0 & \text{otherwise.} \end{cases}$$

This chain is aperiodic, since the identity can be generated. The chain is also irreducible, since any transposition $(i \;\; i+1)$, $1 \leq i < n$ can be produced, and recalling that S_n can be generated by adjacent transpositions. (More concretely, we can cut the deck at i, drop the right packet except for $i+1$, and then drop i, $i+1$, and the rest of the left packet.)

We will bound the mixing time using a creative strong stationary time described by Aldous and Diaconis [2]. First, we will describe how to perform an inverse riffle shuffle (see Figure 3.3.2). Label the cards with n independent bits selected uniformly at random. Pull all the cards labelled 0 to the top of the deck, and append the remaining cards labelled 1 to the end, preserving their relative order.

$$\begin{matrix} & [4]\ [5] & & [6]\ [7] & & [8] \\ [1] & & [2] & & [3] & \end{matrix} \longrightarrow [1]\ [2]\ [3]\ [4]\ [5]\ [6]\ [7]\ [8]$$

$$0\ \ 1\ \ 1\ \ 0\ \ 1\ \ 1\ \ 0\ \ 1$$

Figure 3.3.2: An inverse riffle shuffle – after independent and uniformly random bits are assigned to each card, the cards labelled 0 are pulled to the front of the deck, and the remaining cards labelled 1 are appended to the end, preserving their relative order. (This reverses the riffle shuffle in Figure 3.3.1 by identifying the two rising sequences.)

In other words, if a riffle shuffle generates permutation σ, then the bits identify the two rising sequences and inverts the interleaving procedure, so that σ^{-1} is generated with the same probability. Thus, the inverse riffle shuffle is the time reversal of the riffle shuffle. By Lemma 3.21, it will suffice to bound the mixing time of the inverse riffle shuffle.

Note that each step of the inverse riffle shuffle assigns an independent and uniformly random bit to each card. After t inverse shuffles, each card can be associated with a bitstring of length t. Let τ be the first time that all n cards have a distinct bitstring. Then we claim that τ is a strong stationary time.

Indeed, when each card has a distinct bitstring, the resulting permutation is completely determined by sorting the cards in reverse lexicographic order. This is independent of the starting arrangement and τ. Furthermore, since each bitstring sequence arises independently with the same probability, the n distinct bitstrings at τ can be permuted to obtain all $n!$ equally likely arrangements.

Analysing τ reduces the problem to the well-known Birthday Problem. Observe that $\{\tau > t\}$ corresponds to the event that selecting n bitstrings ("birthdays") from 2^t equally likely choices ("number of days") results in at least one duplication (i.e. two people have the same birthday). Hence, its complement $\{\tau \leq t\}$, the event that all n bitstrings are distinct, has probability $\prod_{k=0}^{n-1}\left(1-\frac{k}{2^t}\right)$.

By applying Theorem 3.19 and Lemma 3.21, we conclude that

$$d(t) = \left\|P^t(\mathrm{id},\cdot) - \pi\right\|_{\mathrm{TV}} = \left\|\widehat{P}^t(\mathrm{id},\cdot) - \pi\right\|_{\mathrm{TV}} \leq \mathbb{P}\left(\tau > t\right) = 1 - \prod_{k=0}^{n-1}\left(1-\frac{k}{2^t}\right). \quad (3.31)$$

Let $t = 2\log_2(n/c)$, where c is a fixed real number. Asymptotically, $d(t) = 1 - e^{-\frac{c^2}{2}+O\left(\frac{1}{n}\right)}$ from (3.31). For sufficiently large n, choosing $c = \frac{3}{4}$ implies that $d(t) \leq \epsilon$, and hence the mixing time satisfies $t_{\mathrm{mix}} \leq 2\log_2\left(\frac{4n}{3}\right)$.

The Gilbert-Shannon-Reeds model of the riffle shuffle is further analysed in Bayer and Diaconis [6]. In their paper, the exact distribution of the chain after t steps is obtained, which leads to even more precise expressions for the distance to stationarity. A summary of some of their key results is given in Section 6.1 of Chapter 6.

Chapter 4

Analytical Techniques for Analysing Mixing Times

In this chapter, we will describe some analytical techniques that can be used to bound mixing times. In Section 4.1, the class of reversible chains will be shown to permit a highly useful spectral representation, which connects the eigenvalues of the transition matrix to mixing times. This will then be used to develop the geometrical methods of conductance (Section 4.2), and canonical paths (Section 4.3).

4.1 Spectral representation

Consider an irreducible Markov chain with finite state space Ω, irreducible transition matrix P, and stationary distribution π. Let $L^2(\pi)$ denote the space of all functions from Ω to $\mathbb{R}$, equipped with the following inner product induced by π:

$$\langle f, g \rangle_\pi := \sum_{x \in \Omega} f(x) g(x) \pi(x), \quad f, g \in L^2(\pi).$$

Suppose that P is reversible with π (i.e. the detailed balance conditions (2.6) are satisfied). Then P is a self-adjoint operator on $L^2(\pi)$, since

$$\begin{aligned} \langle Pf, g \rangle_\pi &= \sum_{x \in \Omega} Pf(x) g(x) \pi(x) = \sum_{x \in \Omega} \sum_{y \in \Omega} P(x, y) f(y) g(x) \pi(x) \\ &= \sum_{y \in \Omega} \sum_{x \in \Omega} P(y, x) f(y) g(x) \pi(y) = \sum_{y \in \Omega} f(y) Pg(y) \pi(y) = \langle f, Pg \rangle_\pi . \end{aligned} \tag{4.1}$$

For the rest of this chapter, we will assume that all Markov chains considered are reversible. This naturally leads to a more elegant theory and tighter bounds. For a greater treatment of the non-reversible case and the analytical tools that can be used to obtain similar bounds (such as Dirichlet forms and log-Sobolev inequalities), see [46, 52].

Theorem 4.1. *Let P be the transition matrix of a reversible and irreducible Markov chain with finite state space Ω and stationary distribution π. Then*

(i) *There is an orthonormal basis for $L^2(\pi)$ consisting of eigenfunctions $\{f_j\}_{j=1}^{|\Omega|}$ of P corresponding to real eigenvalues $\{\lambda_j\}_{j=1}^{|\Omega|}$. The eigenfunction f_1 corresponding to the eigenvalue 1 can be taken to be the function $\mathbf{1}$ that maps every $x \in \Omega$ to one.*

(ii) *Given any function* $f : \Omega \to \mathbb{R}$ *and* $t \in \mathbb{N}$*, we can write*

$$P^t f = \langle f, \mathbf{1} \rangle_\pi \mathbf{1} + \sum_{j=2}^{|\Omega|} \lambda_j^t \langle f, f_j \rangle_\pi f_j. \tag{4.2}$$

(Note that $\langle f, \mathbf{1} \rangle_\pi = \mathbb{E}_\pi[f]$*.) In particular, for any* $x, y \in \Omega$*,*

$$\frac{P^t(x,y)}{\pi(y)} = 1 + \sum_{j=2}^{|\Omega|} \lambda_j^t f_j(x) f_j(y). \tag{4.3}$$

Proof. (i) Recall from (4.1) that P is a self-adjoint operator on $L^2(\pi)$. Hence, the Spectral Theorem (e.g. see [27, p.154, Theorem 6]) implies that there exists a orthonormal basis for $L^2(\pi)$ of real-valued eigenfunctions of P, associated with real eigenvalues. Note that $P\mathbf{1} = \mathbf{1}$ because P is stochastic. Since the eigenspace associated with the eigenvalue 1 is one-dimensional by the Perron-Frobenius Theorem (Theorem 2.14), we can take $f_1 = \mathbf{1}$.

(ii) Observe that for each j, $\langle P^t f, f_j \rangle_\pi = \langle f, P^t f_j \rangle_\pi = \lambda_j^t \langle f, f_j \rangle_\pi$. Therefore, expanding $P^t f$ with respect to the orthonormal basis $\{f_j\}_{j=1}^{|\Omega|}$ shows that

$$P^t f = \langle P^t f, \mathbf{1} \rangle_\pi \mathbf{1} + \sum_{j=2}^{|\Omega|} \langle P^t f, f_j \rangle_\pi f_j = \langle f, \mathbf{1} \rangle_\pi \mathbf{1} + \sum_{j=2}^{|\Omega|} \lambda_j^t \langle f, f_j \rangle_\pi f_j,$$

as desired. For the second statement, recall the delta function $\delta_y(x) = \mathbb{1}_{\{y=x\}}$. Note that for each j, $\langle \delta_y, f_j \rangle_\pi = f_j(y)\pi(y)$. Since $P^t(x,y) = (P^t \delta_y)(x)$, it follows from (4.2) that

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

Moving $\pi(y)$ to the other side completes the proof. □

Definition 4.2. Let P be a reversible and irreducible transition matrix. Then its eigenvalues can be written in decreasing order: $1 = \lambda_1 > \lambda_2 \geq \cdots \geq \lambda_{|\Omega|} \geq -1$.

We call λ_2 the **second eigenvalue**, and $\lambda_{|\Omega|}$ the smallest eigenvalue. We define the absolute second eigenvalue to be $\lambda_* := \max\{\lambda_2, |\lambda_{|\Omega|}|\}$. The **spectral gap** and **absolute spectral gap** are defined by $\gamma = 1 - \lambda_2$ and $\gamma_* = 1 - \lambda_*$ respectively.

The following relates the eigenvalues of an irreducible transition matrix P to its period. In particular, if P is aperiodic, then -1 is not an eigenvalue (so that $\lambda_{|\Omega|} > -1$).

Proposition 4.3 ([38, Exercise 12.1])**.** *Let* P *be an irreducible transition matrix, and fix any state* $x_0 \in \Omega$*. Recall that* $\mathcal{T}(x_0) = \{t \in \mathbb{Z}_+ : P^t(x,x) > 0\}$ *from Definition 2.1, and that the period of* P *is the greatest common divisor of* $\mathcal{T}(x_0)$*.*

If ω *is the* a*-th root of unity, then* a *is a common divisor of* $\mathcal{T}(x_0)$ *if and only if* ω *is an eigenvalue of* P*.*

Proof. Let b be the greatest common divisor of $\mathcal{T}(x_0)$. If a divides $\mathcal{T}(x_0)$, then a divides b. Recall that from Proposition 2.12, Ω can be partitioned into $\{\mathcal{C}_k\}_{k=0}^{b-1}$. For any $x \in \Omega$, $x \in \mathcal{C}_{j(x)}$ for a certain $j(x) \in \{0, 1, \ldots, b-1\}$. Consider the function $f(x) = \omega^{j(x)}$. If $P(x, y) > 0$, then $y \in \mathcal{C}_{j(x)+1}$ (where the indices are taken modulo b). Hence, f is an eigenfunction since $Pf(x) = \sum_{y \in \Omega} P(x, y)\omega^{j(y)} = \omega^{j(x)+1} = \omega f(x)$.

Conversely, suppose that ω is an a-th root of unity, and $Pf = \omega f$ for some f. Then it can be shown that if $P(x, y) > 0$, then $f(y) = \omega f(x)$ (however, we will omit the details). Hence, we can define $\mathcal{D}_j = \{x \in \Omega : f(x) = \omega^j f(x_0)\}$. Thus, if $x \in \mathcal{D}_j$ and $P(x, y) > 0$, then $y \in \mathcal{D}_{j+1}$ (where the indices are taken modulo b). Hence, a must divide $\mathcal{T}(x_0)$. □

Lemma 4.4. *Consider a Markov chain with transition matrix P. Let $\widetilde{P} = \frac{P+I}{2}$ be the transition matrix of the lazy version of the chain, such that the holding probabilities satisfy $\widetilde{P}(x, x) \geq \frac{1}{2}$ for all x. If λ is an eigenvalue of P, then $\frac{1}{2}(\lambda + 1)$ is an eigenvalue of $\widetilde{P}$. In particular, all the eigenvalues of $\widetilde{P}$ are non-negative.*

Proof. Suppose that f is an eigenfunction of P, corresponding to eigenvalue λ. Then we have $\widetilde{P}f = \frac{Pf+If}{2} = \frac{\lambda f+f}{2} = \frac{1}{2}(\lambda + 1)f$, so that f is an eigenfunction of $\widetilde{P}$, corresponding to eigenvalue $\frac{1}{2}(\lambda + 1)$. Since $-1 \leq \lambda \leq 1$, it follows that $0 \leq \frac{1}{2}(\lambda + 1) \leq 1$. □

Remark 4.5.
(i) Making a chain lazy addresses the potential problem of periodicity. Moreover, it also ensures that all the eigenvalues are non-negative by Lemma 4.4. This may lead to more convenient analysis, since the spectral gap γ and absolute spectral gap γ_* coincide. However, laziness will slow down the chain by at most a factor of two. For more discussion on lazy chains in combinatorial problems, see [25].
(ii) Certain Markov chains may naturally have non-negative eigenvalues, such as heat-bath chains [22] (for example, this includes the single-site Glauber dynamics discussed in Chapter 7). Thus, lazy versions of such chains are unnecessary.

The following theorem shows that the absolute second eigenvalue λ_* effectively determines the mixing time. The further away λ_* is from 1, the faster the chain mixes.

Theorem 4.6. *Let P be the transition matrix of a reversible, irreducible, and aperiodic chain, with finite state space Ω, and stationary distribution π. Let $\pi_{\min} \coloneqq \min_{x \in \Omega} \pi(x)$. Then*

$$\left\lfloor \frac{\lambda_*}{1 - \lambda_*} \log\left(\frac{1}{2\epsilon}\right) \right\rfloor \leq t_{\text{mix}}(\epsilon) \leq \left\lceil \frac{1}{1 - \lambda_*} \log\left(\frac{1}{\epsilon \pi_{\min}}\right) \right\rceil. \tag{4.4}$$

Proof. We first show the upper bound of (4.4). Recall that $|\lambda_j| \leq \lambda_*$ for all $2 \leq j \leq |\Omega|$. Fix $x, y \in \Omega$. Starting from (4.3), using the triangle inequality, and then the Cauchy-Schwarz inequality, shows that

$$\left| \frac{P^t(x, y)}{\pi(y)} - 1 \right| \leq \lambda_*^t \sum_{j=2}^{|\Omega|} |f_j(x)|\,|f_j(y)| \leq \lambda_*^t \left[\sum_{j=2}^{|\Omega|} f_j^2(x)\right]^{1/2} \left[\sum_{j=2}^{|\Omega|} f_j^2(y)\right]^{1/2}. \tag{4.5}$$

Using (4.2), we can expand δ_x with respect to the orthonormal basis $\{f_j\}_{j=1}^{|\Omega|}$ to show that

$$\pi(x) = \langle \delta_x, \delta_x \rangle_\pi = \left\langle \sum_{j=1}^{|\Omega|} f_j(x)\pi(x) f_j, \sum_{j=1}^{|\Omega|} f_j(x)\pi(x) f_j \right\rangle_\pi = \pi(x)^2 \sum_{j=1}^{|\Omega|} f_j(x)^2. \tag{4.6}$$

Since $f_1(x) = 1$, this shows that $\sum_{j=2}^{|\Omega|} f(x)^2 \leq \frac{1}{\pi(x)}$. Thus, (4.5) implies that

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

using the inequality $1 - x \leq e^{-x}$ for $x \geq 0$. Since this is independent of x and y, this provides an upper bound for the separation distance $s(t)$. Furthermore, since $d(t) \leq s(t)$ by Lemma 3.18, this shows that $d(t) \leq \frac{e^{-\gamma_* t}}{\pi_{\min}}$. Hence, setting t equal to the upper bound in (4.4) shows that $d(t) \leq \epsilon$, as desired.

Next, we will show the lower bound of (4.4). Let f be an eigenfunction of P corresponding to eigenvalue $\lambda \neq 1$, so that $Pf = \lambda f$. Recalling that the eigenfunctions of a self-adjoint operator corresponding to distinct eigenvalues are orthogonal (e.g. see [27, p.154, Theorem 4]), we have $\langle \mathbf{1}, f \rangle_\pi = \sum_{y \in \Omega} \pi(y) f(y) = 0$. Let $\|f\|_\infty := \max_{x \in \Omega} |f(x)|$. Then for any $x \in \Omega$,

$$|\lambda^t f(x)| = |P^t f(x)| = \left| \sum_{y \in \Omega} \left([P^t(x,y) - \pi(y)] f(y) \right) \right| \leq \|f\|_\infty \cdot 2d(t). \tag{4.7}$$

Choosing x such that $|f(x)| = \|f\|_\infty$, (4.7) implies that $|\lambda|^t \leq 2d(t)$. Hence $|\lambda|^{t_{\text{mix}}(\epsilon)} \leq 2\epsilon$, which is equivalent to $t_{\text{mix}}(\epsilon) \log\left(\frac{1}{|\lambda|}\right) \geq \log\left(\frac{1}{2\epsilon}\right)$. Using the inequality $x - 1 \geq \log x$ for $x > 0$ then implies that $t_{\text{mix}}(\epsilon)\left(\frac{1}{|\lambda|} - 1\right) \geq \log\left(\frac{1}{2\epsilon}\right)$. Since this holds with $\lambda = \lambda_*$ in particular, the lower bound is obtained by rearranging. □

Example 4.7 (Random walk on the cycle)**.** Recall the random walk on the n-cycle from Example 3.4. We will explicitly find the eigenvalues of the associated non-lazy transition matrix P. By Lemma 4.4, this will give the eigenvalues of the lazy walk.

It will be algebraically convenient to consider the state space $\Omega = \{1, \omega, \omega^2, \ldots, \omega^{n-1}\}$, where ω is the nth root of unity in $\mathbb{C}$. At each step, the chain moves from ω^k to ω^{k-1} or ω^{k+1} with equal probability. Let $\phi_j : \Omega \to \mathbb{C}$ be defined by $\phi_j(\omega^k) = \omega^{jk}$. Then we can check that for all $0 \leq j, k \leq n-1$,

$$P\phi_j(\omega^k) = \frac{\phi_j(\omega^{k-1}) + \phi_j(\omega^{k+1})}{2} = \frac{\omega^{jk-j} + \omega^{jk+j}}{2} = \left(\frac{\omega^{-j} + \omega^j}{2} \right) \phi_j(\omega^k).$$

Similarly, its complex conjugate $\overline{\phi}_j : \Omega \to \mathbb{C}$ defined by $\overline{\phi}_j(\omega^k) = \omega^{-jk}$ also satisfies $P\overline{\phi}_j = \left(\frac{\omega^{-j}+\omega^j}{2}\right)\overline{\phi}_j$. Hence, the real-valued function $f_j := \text{Re}(\phi_j) = \frac{1}{2}(\phi_j + \overline{\phi}_j)$ is an

eigenfunction of P, with eigenvalue $\lambda_j = \cos\left(\frac{2\pi j}{n}\right)$. This gives the n distinct eigenvalues of P. The second eigenvalue is $\lambda_2 = \cos\left(\frac{2\pi}{n}\right) = 1 - \frac{4\pi^2}{n^2} + O(n^{-4})$, using Taylor series.

Hence, the lazy chain has second eigenvalue $\frac{1+\lambda_2}{2} = 1 - \frac{2\pi^2}{n^2} + O(n^{-4})$, and spectral gap $\frac{2\pi^2}{n^2} + O(n^{-4})$. Since $\pi_{\min} = \frac{1}{n}$, Theorem 4.6 implies an $O(n^2 \log n)$ upper bound for the mixing time, which has an extra factor of $\log n$ compared to the coupling bound.

Example 4.8 (Random walk on the hypercube)**.** Recall the lazy random walk on the n-dimensional hypercube from Example 3.6. Denote its state space by $\widetilde{\Omega}$. We will now find the eigenvalues of the associated (lazy) transition matrix $\widetilde{P}$.

First, consider the case of $n = 1$. This is a random walk on $\Omega = \{0, 1\}$ with 2×2-transition matrix P, where each element is equal to $\frac{1}{2}$. Consider the functions $\mathbf{1}$ and g, where $\mathbf{1}(x) = 1$ and $g(x) = 2x - 1$ for $x = 0, 1$. Then $P\mathbf{1} = (\frac{1}{2} + \frac{1}{2})\mathbf{1} = \mathbf{1}$, and $Pg = (\frac{1}{2} - \frac{1}{2})\mathbf{1} = \mathbf{0}$. Hence $\mathbf{1}$ and g are eigenfunctions of P, corresponding to eigenvalues 1 and 0 respectively.

Now let $n \geq 1$. Let P_i, Ω_i be n copies of the transition matrix and state space where $n = 1$. Then for the lazy random walk on the n-dimensional hypercube, $\widetilde{\Omega} = \Omega_1 \times \dots \Omega_n$, and

$$\widetilde{P}(\mathbf{x}, \mathbf{y}) = \sum_{i=1}^{n} \left[\frac{1}{n} P_i(x_i, y_i) \prod_{j \neq i} \mathbb{1}_{\{y_j = x_j\}} \right],$$

where $\mathbf{x} = (x_i)_{i=1}^n$, $x_i \in \Omega_i$, and similarly for $\mathbf{y}$. In other words, a bit is selected uniformly at random, and a transition is made in that coordinate. The product signifies that $\mathbf{y}$ is accessible from $\mathbf{x}$ if and only if they differ by at most one bit.

For each $1 \leq i \leq n$, let f_i be an eigenfunction of P_i corresponding to eigenvalue λ_i. Then $P_i f_i = \lambda_i f_i$, or equivalently, $\sum_{y_i \in \Omega_i} P_i(x_i, y_i) f_i(y_i) = \lambda_i f_i(x_i)$ for all $x_i \in \Omega_i$. We claim that the function $\widetilde{f}$ defined by $\widetilde{f}(\mathbf{x}) = f_1(x_1) f_2(x_2) \dots f_n(x_n)$ is an eigenfunction of $\widetilde{P}$ (this is actually the tensor product $\widetilde{f} = f_1 \otimes \dots \otimes f_n$). Indeed, since only one bit can be changed at a time, we can write

$$\begin{aligned} \widetilde{P}\widetilde{f}(\mathbf{x}) = \sum_{\mathbf{y} \in \widetilde{\Omega}} \widetilde{P}(\mathbf{x}, \mathbf{y}) \widetilde{f}(\mathbf{y}) &= \sum_{i=1}^{n} \sum_{y_i \in \Omega_i} \left[\frac{1}{n} P_i(x_i, y_i) f_i(y_i) \prod_{j \neq i} f_j(x_j) \right] \\ &= \sum_{i=1}^{n} \frac{1}{n} \lambda_i f_1(x_1) f_2(x_2) \dots f_n(x_n) = \left(\sum_{i=1}^{n} \frac{\lambda_i}{n} \right) \widetilde{f}(\mathbf{x}). \end{aligned}$$

This shows that $\widetilde{f}$ is an eigenfunction of $\widetilde{P}$, corresponding to eigenvalue $\sum_{i=1}^n \frac{\lambda_i}{n}$.

Therefore, varying the choices of whether each f_i corresponds to eigenvalue 0 or 1 gives all the eigenvalues of $\widetilde{P}$. Setting $n-1$ of the f_i to correspond to 1, and the remaining one to 0 implies that the (absolute) second eigenvalue of $\widetilde{P}$ is $\widetilde{\lambda}_2 = \frac{n-1}{n} = \widetilde{\lambda}_*$. Since $\pi_{\min} = \frac{1}{n}$, Theorem 4.6 implies that the mixing time t_{mix} is upper bounded by $\lceil n \log(4n) \rceil$. This is the same as the coupling bound.

4.2 Conductance

In the next two sections, we will use the spectral representation of a reversible Markov chain to develop some geometrical methods for bounding the mixing time. While probabilistic methods, such as coupling and strong stationary times, can yield tight bounds for simple chains, they are less amenable to more irregular processes, which lack a high degree of symmetry [31]. The following methods, based on the underlying weighted graph of a chain, have been particularly successful for some of these problems.

In this section, we will discuss the method of conductance, which is typically used to lower bound the mixing time. Intuitively, the conductance measures whether there are any "bottlenecks" in the underlying geometry of a chain that inhibit mixing. This is related to the Cheeger constant in differential geometry (see [33] for further discussion).

Definition 4.9. Let P be a transition matrix with finite state space Ω and stationary distribution π. Then the **edge measure** Q is defined by

$$\begin{aligned} Q(x,y) &= \pi(x)P(x,y), \quad x,y \in \Omega, \\ Q(A,B) &= \sum_{x\in A, y\in B} Q(x,y), \quad A,B \subseteq \Omega. \end{aligned} \tag{4.8}$$

The **conductance** (also known as the bottleneck ratio) of a set $S \subseteq \Omega$, and of the whole chain, is defined by, respectively,

$$\Phi(S) = \frac{Q(S,S^c)}{\pi(S)}, \qquad \text{and} \qquad \Phi_* = \min_{\substack{S\subseteq\Omega \\ \pi(S)\leq\frac{1}{2}}} \Phi(S). \tag{4.9}$$

Lemma 4.10. *Continuing the notation used in Definition 4.9,*

(i) $0 \leq \Phi(S) \leq 1$ *for any* $S \subseteq \Omega$*; and*

(ii) $Q(S,S^c) = Q(S^c,S)$ *for any* $S \subseteq \Omega$.

Proof. (i) The lower bound is clear. For the upper bound, observe that

$$\Phi(S) = \sum_{x\in S} \left[\frac{\pi(x)}{\pi(S)} \sum_{y\in S^c} P(x,y) \right].$$

Since $\sum_{y\in S^c} P(x,y) \leq 1$ for all $x \in S$, it follows that $\Phi(S) \leq 1$.

(ii) Recall that $\pi = \pi P$ is equivalent to $\pi(y) = \sum_{x\in\Omega} \pi(x)P(x,y)$ for all $y \in \Omega$. Thus, by interchanging the order of summation,

$$\begin{aligned} Q(S^c,S) = \sum_{x\in S^c, y\in S} \pi(x)P(x,y) &= \sum_{y\in S} \left[\sum_{x\in\Omega} \pi(x)P(x,y) - \sum_{x\in S} \pi(x)P(x,y) \right] \\ &= \sum_{y\in S} \pi(y) - \sum_{y\in S, x\in S} \pi(x)P(x,y). \end{aligned}$$

Since $\sum_{y\in\Omega} P(x,y) = 1$, by interchanging the order of summation again,

$$Q(S^c, S) = \sum_{y\in S} \pi(y) - \sum_{x\in S} \pi(x) \left[\sum_{y\in\Omega} P(x,y) - \sum_{y\in S^c} P(x,y) \right]$$
$$= \sum_{y\in S} \pi(y) - \sum_{x\in S} \pi(x) + \sum_{x\in S, y\in S^c} \pi(x)P(x,y) = Q(S, S^c),$$

as desired. □

Remark 4.11. The conductance of a set S can be interpreted as the conditional probability of leaving S, given that the chain starts in S. A set with high conductance is conducive for the chain to leave and explore the rest of the state space. The conductance of the whole chain looks for the global bottleneck, which has the lowest conductance.

The conductance provides a particularly simple lower bound on the mixing time of irreducible (and not necessarily reversible) chains.

Theorem 4.12. *Consider an irreducible Markov chain with transition matrix P and stationary distribution π. Then*

$$t_{\text{mix}}(\epsilon) \geq \left\lfloor \left(\frac{1}{2} - \epsilon \right) \frac{1}{\Phi_*} \right\rfloor. \tag{4.10}$$

In particular, $t_{\text{mix}} \geq \left\lfloor \frac{1}{4\Phi_*} \right\rfloor$.

Proof. Fix $S \subseteq \Omega$ with $\pi(S) \leq \frac{1}{2}$. Define π_S by $\pi_S(A) = \pi(A \cap S)$ for any $A \subseteq \Omega$, which is π restricted to S. Define μ_S by $\mu_S(A) = \frac{\pi(A\cap S)}{\pi(S)}$, which is π conditioned on S. Recall that from (2.31), the total variation distance between $\mu_S P$ and μ_S can be written

$$\pi(S) \left\| \mu_S P - \mu_S \right\|_{\text{TV}} = \pi(S) \sum_{\substack{y\in\Omega \\ \mu_S P(y) \geq \mu_(y)}} [\mu_S P(y) - \mu_S(y)] = \sum_{\substack{y\in\Omega \\ \pi_S P(y) \geq \pi_(y)}} [\pi_S P(y) - \pi_S(y)]. \tag{4.11}$$

Note that $\pi_S(x) = \pi(x)\mathbb{1}_{\{x\in S\}}$. Thus $\pi_S P(y) = \sum_{x\in\Omega} \pi_S(x)P(x,y) = \sum_{x\in S} \pi(x)P(x,y)$, which is less than or equal to $\sum_{x\in\Omega} \pi(x)P(x,y) = \pi(y)$. If $y \in S$, then $\pi_S P(y) \leq \pi_S(y)$.

On the other hand, if $y \notin S$, then $\pi_S P(y) \geq 0 = \pi_S(y)$. Thus, the sum in the last term of (4.11) can be taken over all $y \in S^c$ (which satisfies $\pi_S(y) = 0$), so that

$$\pi(S) \left\| \mu_S P - \mu_S \right\|_{\text{TV}} = \sum_{y\in S^c} \pi_S P(y) = \sum_{y\in S^c, x\in S} \pi(x)P(x,y) = Q(S, S^c).$$

Hence, $\left\| \mu_S P - \mu_S \right\|_{\text{TV}} = \Phi(S)$. Recall that the total variation distance is non-decreasing when advanced by P from (2.33). Thus, $\left\| \mu_S P^{u+1} - \mu_S P^u \right\|_{\text{TV}} \leq \left\| \mu_S P - \mu_S \right\|_{\text{TV}} = \Phi(S)$

for any $u \in \mathbb{N}$. By the triangle inequality,

$$\left\|\mu_S P^t - \mu_S\right\|_{\mathrm{TV}} = \left\|\sum_{u=0}^{t-1}(\mu_S P^{u+1} - \mu_S P^u)\right\|_{\mathrm{TV}} \leq \sum_{u=0}^{t-1}\left\|\mu_S P^{u+1} - \mu_S P^u\right\|_{\mathrm{TV}} \leq t\Phi(S). \tag{4.12}$$

Since $\pi(S) \leq \frac{1}{2}$, $\|\mu_S - \pi\|_{\mathrm{TV}} \geq \pi(S^c) - \mu_S(S^c) = \pi(S^c) \geq \frac{1}{2}$. Using the triangle inequality again shows that

$$\frac{1}{2} \leq \|\mu_S - \pi\|_{\mathrm{TV}} \leq \left\|\mu_S P^t - \mu_S\right\|_{\mathrm{TV}} + \left\|\mu_S P^t - \pi\right\|_{\mathrm{TV}}. \tag{4.13}$$

Therefore, if $t = t_{\mathrm{mix}}(\epsilon)$, then the last term in (4.13) is at most ϵ, by definition of the total variation distance (recall that the starting distribution μ_S does not matter, by Corollary 2.22 (ii)). Combining this with (4.12) shows that $\frac{1}{2} \leq t_{\mathrm{mix}}(\epsilon)\Phi(S) + \epsilon$. Rearranging and then minimising over all subsets S with $\pi(S) \leq \frac{1}{2}$ leads to the desired bound. □

We will now discuss how the conductance of a reversible chain characterises its mixing.

Definition 4.13. Let P be a reversible transition matrix with stationary distribution π. The **Dirichlet form** on P is defined on functions f and g on Ω by

$$\mathcal{E}(f, g) \coloneqq \langle (I - P)f, g\rangle_\pi, \tag{4.14}$$

where I is the identity operator. We also define $\mathcal{E}(f) \coloneqq \mathcal{E}(f, f)$. By expanding the square and using the reversibility assumption, it is not too difficult to check that

$$\mathcal{E}(f) = \frac{1}{2}\sum_{x,y\in\Omega}[f(x) - f(y)]^2 Q(x, y). \tag{4.15}$$

The following variational characterisation of the spectral gap will be key.

Proposition 4.14. *Suppose that $f \in L^2(\pi)$. Then the spectral gap $\gamma = 1 - \lambda_2$ satisfies*

$$\gamma = \inf_{\mathbb{E}_\pi[f]=0, \|f\|_\pi=1} \mathcal{E}(f) = \inf_{\mathbb{E}_\pi[f]=0, f\neq 0} \frac{\mathcal{E}(f)}{\|f\|_\pi^2} = \inf_{\mathrm{Var}_\pi(f)\neq 0} \frac{\mathcal{E}(f)}{\mathrm{Var}_\pi(f)}. \tag{4.16}$$

Here $\|f\|_\pi^2 = \langle f, f\rangle_\pi$ and $\mathrm{Var}_\pi(f) = \|f - \mathbb{E}_\pi[f]\|_\pi^2$. Also, note that the condition that $\mathbb{E}_\pi[f] = 0$ is equivalent to $\langle f, \mathbf{1}\rangle_\pi = 0$, that is $f \perp_\pi \mathbf{1}$.

Proof. Note that this is essentially the min-max theorem from linear algebra for the second eigenvalue (for example, see [27, p.181]). We will provide a nice, brief proof using the spectral representation.

Let $f \in L^2(\pi)$. Consider the orthonormal basis $\{f_j\}_{j=1}^{|\Omega|}$ of $L^2(\pi)$ from Theorem 4.1. If $\|f\|_\pi = 1$ and $f \perp_\pi \mathbf{1}$, then $f = \sum_{j=2}^{|\Omega|} c_j f_j$, where $c_j = \langle f, f_j \rangle$ and $\sum_{j=2}^{|\Omega|} c_j^2 = 1$. Thus,

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

Hence, the infimum of $\mathcal{E}(f)$ is greater than or equal to $1 - \lambda_2$. Since f_2 satisfies $\|f_2\|_\pi = 1$ and $\langle f_2, \mathbf{1} \rangle_\pi = 0$, the infimum of $\mathcal{E}(f)$ is at most $\mathcal{E}(f_2) = 1 - \lambda_2$. This shows the first equality. For the second equality, observe that $\mathcal{E}(cf) = c^2 \mathcal{E}(f)$ for any real c. Hence, for any $f \neq 0$ with $\mathbb{E}_\pi[f] = 0$,

$$\frac{\mathcal{E}(f)}{\|f\|_\pi^2} = \mathcal{E}\left(\frac{f}{\|f\|_\pi}\right) \geq \inf_{\mathbb{E}_\pi[f]=0, \|f\|_\pi = 1} \mathcal{E}(f).$$

Therefore, $\inf_{\mathbb{E}_\pi[f]=0, f \neq 0} \frac{\mathcal{E}(f)}{\|f\|_\pi^2} \geq \inf_{\mathbb{E}_\pi[f]=0, \|f\|_\pi = 1} \mathcal{E}(f)$. The opposite inequality follows from subset inclusion. Finally, the third equality follows similarly, by observing that $\mathcal{E}(f + c) = \mathcal{E}(f)$ for any real c, and using the standard result $\mathrm{Var}_\pi(f + c) = \mathrm{Var}_\pi(f)$. □

Remark 4.15. A useful form of the variational characterisation for the spectral gap is

$$\gamma = \inf_{\mathbb{E}_\pi[f]=0, f \neq 0} \frac{\sum_{x,y \in \Omega} [f(x) - f(y)]^2 Q(x, y)}{\sum_{x,y \in \Omega} [f(x) - f(y)]^2 \pi(x) \pi(y)}. \tag{4.17}$$

This follows from Proposition 4.14, by verifying that for any non-zero $f \in L^2(\pi)$ satisfying $\mathbb{E}_\pi[f] = \sum_{x \in \Omega} f(x)\pi(x) = 0$, the denominator of (4.17) equals $\|f\|_\pi^2 = \sum_{x \in \Omega} f(x)^2 \pi(x)$. (This can be done by expanding the square, and using $\sum_{x \in \Omega} \pi(x) = 1$.)

The variational characterisation is used to prove the next theorem, which connects the conductance of a chain and the spectral gap. Since the details are somewhat beyond the scope of this thesis, the proof will be omitted for brevity.

Theorem 4.16 ([38, Theorem 13.14])**.** *Let P be the transition matrix of a reversible and irreducible Markov chain, with finite state space Ω and stationary distribution π. Then the spectral gap $\gamma = 1 - \lambda_2$ is related to the conductance of the chain by*

$$\frac{\Phi_*^2}{2} \leq 1 - \lambda_2 \leq 2\Phi_*. \tag{4.18}$$

By combining Theorem 4.16 with the bounds on the mixing time from the spectral gap in Theorem 4.6, we deduce the following.

Corollary 4.17. *Suppose that $\lambda_* = \lambda_2$ (for example, if the chain is lazy). Then the conductance provides the following bounds on the mixing time:*

$$\left\lfloor \left(\frac{1}{2\Phi_*} - 1\right) \log\left(\frac{1}{2\epsilon}\right) \right\rfloor \leq t_{\mathrm{mix}}(\epsilon) \leq \left\lceil \frac{2}{\Phi_*^2} \log\left(\frac{1}{\epsilon \pi_{\min}}\right) \right\rceil. \tag{4.19}$$

Remark 4.18. Comparing the lower bound on the mixing time in Corollary 4.17 with the lower bound in Theorem 4.12, it is interesting to see that for $\epsilon = 1/4$, the latter bound actually performs better, even though it does not assume reversibility. The bound obtained using the spectral representation is superior only for much smaller ϵ.

Example 4.19 (Random walk on two glued complete graphs)**.** Consider a random walk on two complete graphs K_n glued at a particular vertex v (see Figure 4.2.1). Suppose that $\Omega = \{v, v_2, \ldots, v_n, w_2, \ldots, w_n\}$, where we identify the common vertex $v = v_1 = w_1$.

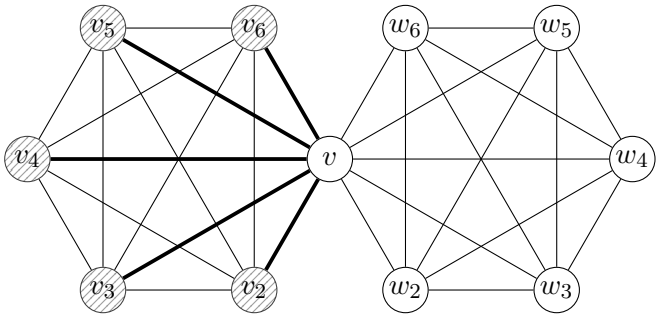

Figure 4.2.1: A random walk on two copies of the complete graph K_n glued together at a common vertex (here $n = 6$). The vertex v is a bottleneck that inhibits mixing, since the only path out of the shaded set is via the bold edges.

The random walk selects a neighbouring vertex (including itself) to move to, uniformly at random. More precisely, the transition probabilities are $P(v_i, v_j) = P(w_i, w_j) = \frac{1}{n}$ for $1 < i \leq n$, $1 \leq j \leq n$. For the common vertex, $P(v, v_i) = P(v, w_i) = \frac{1}{2n-1}$ for $1 \leq i \leq n$.

Consider the distribution π given by $\pi(v_i) = \pi(w_i) = \frac{C}{2n-1}$ for $1 < i \leq n$, and $\pi(v) = \frac{C}{n}$, where C is the normalising constant. It can be checked that P is reversible with respect to π, and hence π is the stationary distribution by Proposition 2.6.

Let $A = \{v_2, \ldots, v_n\}$ be the vertices of one of the complete graphs, excluding the common vertex v. The conductance of this set is $\Phi(A) = \frac{Q(A, A^c)}{\pi(A)}$, where $\pi(A) = \frac{C(n-1)}{2n-1}$ and $Q(A, A^c) = \frac{C(n-1)}{n(2n-1)}$, since the only way out of A is through v. Therefore $\Phi(A) = \frac{1}{n}$, and Theorem 4.12 implies that the mixing time of the chain has a lower bound of $\left\lfloor \frac{n}{4} \right\rfloor$.

4.3 Canonical paths

The idea of the canonical paths method is to see whether a set of canonical paths between each state can be constructed, such that "hot spots" carrying a particularly large burden are avoided. The method has been used to obtain upper bounds on complicated combinatorial chains (such as sampling matchings of a graph [33]).

The canonical paths method was first used to bound the conductance by Jerrum and Sinclair [31]. Subsequent papers by Diaconis and Stroock [17] and Sinclair [54] showed that the method could be used to directly bound the mixing time. For more detailed applications, see [30, 31]. This has also been generalised to the method of multicommodity flows, which allows for multiple paths between states (for example, see [54, 47]).

In this section, we will continue to assume that all chains considered are reversible. (For non-reversible versions of the canonical paths bound, see [30, Corollary 5.9].)

Definition 4.20. Suppose that P is the transition matrix of a reversible and irreducible Markov chain, with finite state space Ω and stationary distribution π. We can view the chain as a directed graph (Ω, E), where $(x, y) \in E$ if and only if $P(x, y) > 0$. We define the edge measure Q (4.8) applied to an edge $e = (x, y)$ by $Q(e) = Q(x, y) = \pi(x)P(x, y)$.

A **canonical path** γ_{xy} between a pair $(x, y) \in \Omega^2$ is a sequence of edges (x_{i-1}, x_i), $1 \leq i \leq r$, such that $x_0 = x$ and $x_r = y$. The length of the path is $|\gamma_{xy}| = r$. Let $\Gamma = \{\gamma_{xy} : x, y \in \Omega\}$ denote a chosen set of canonical paths between all pairs $(x, y) \in \Omega^2$. The **maximum edge loading** of Γ is defined by

$$\rho(\Gamma) = \max_{e \in E} \frac{1}{Q(e)} \sum_{\gamma_{xy} \ni e} \pi(x)\pi(y)|\gamma_{xy}|. \tag{4.20}$$

This measures the maximum "flow" in the chosen set of canonical paths along any edge e, as a fraction of its capacity $Q(e)$. Intuitively, the chain is rapidly mixing if a good set of paths can be selected, such that no particular edge is over-utilised (i.e. no bottlenecks).

Theorem 4.21 (Sinclair [54])**.** *Let P be the transition matrix of a reversible and irreducible Markov chain with stationary distribution π. For any choice of canonical paths Γ, the spectral gap $\gamma = 1 - \lambda_2$ satisfies*

$$\gamma \geq \frac{1}{\rho(\Gamma)}. \tag{4.21}$$

Proof. Recall the functional characterisation of the spectral gap from Proposition 4.14. Let $f \in L^2(\pi)$ be any non-zero function with $\mathbb{E}_\pi[f] = 0$. Since the edges in γ_{xy}, which we shall denote by $e = (e^-, e^+)$, ultimately start at x and end at y, the denominator of (4.17), $\sum_{x,y\in\Omega}[f(x) - f(y)]^2\pi(x)\pi(y)$, can be written using a telescoping sum as

$$\sum_{x,y\in\Omega} \pi(x)\pi(y)\Big(\sum_{e\in\gamma_{xy}} [f(e^+) - f(e^-)]\Big)^2 \leq \sum_{x,y\in\Omega} \pi(x)\pi(y)\,|\gamma_{xy}|\Big(\sum_{x,y\in\Omega} [f(e^+) - f(e^-)]^2\Big).$$

By summing over all edges instead of the vertices, the upper bound above is equal to

$$\begin{aligned}\sum_{e\in E}[f(e^+) - f(e^-)]^2\Big(\sum_{\gamma_{xy}\ni e} \pi(x)\pi(y)\,|\gamma_{xy}|\Big) &\leq \sum_{e\in E}[f(e^+) - f(e^-)]^2 Q(e)\rho(\Gamma) \\ &= \rho(\Gamma)\sum_{x,y\in\Omega}[f(x) - f(y)]^2 Q(x, y).\end{aligned}$$

Rearranging shows that $\frac{1}{\rho(\Gamma)} \leq \frac{\sum_{x,y\in\Omega}[f(x)-f(y)]^2 Q(x,y)}{\sum_{x,y\in\Omega}[f(x)-f(y)]^2\pi(x)\pi(y)}$. Since f was arbitrary, and γ is the infimum by (4.17), this implies that $\gamma \geq \frac{1}{\rho(\Gamma)}$, as desired. □

By combining Theorem 4.21 with the bounds on the mixing time in terms of the spectral gap from Theorem 4.6, we can deduce the following upper bound on the mixing time in terms of the maximum edge loading.

Corollary 4.22. *Suppose that $\lambda_* = \lambda_2$ (for example, if the chain is lazy). Let Γ be a set of canonical paths with maximum edge loading $\rho(\Gamma)$. Then*

$$t_{\mathrm{mix}}(\epsilon) \leq \left\lceil \rho(\Gamma) \log\left(\frac{1}{\epsilon \pi_{\min}}\right) \right\rceil. \tag{4.22}$$

Example 4.23 (Random walk on the hypercube)**.** Recall the lazy random walk on the n-dimensional hypercube from Example 3.6. We will now obtain an upper bound of the mixing time using a canonical paths argument.

Let $\mathbf{x} = (x_1, x_2, \ldots, x_n)$ and $\mathbf{y} = (y_1, y_2, \ldots, y_n)$ be arbitrary states (i.e. bitstrings) in Ω. Define the canonical path $\gamma_{\mathbf{xy}}$ between $\mathbf{x}$ and $\mathbf{y}$ by $\mathbf{x}_0 = \mathbf{x}$, $\mathbf{x}_1$, $\ldots$, $\mathbf{x}_n = \mathbf{y}$, where $\mathbf{x}_i = (y_1, y_2, \ldots, y_i, x_{i+1}, \ldots, x_n)$ for $1 \leq i \leq n$. That is, we change the bits of $\mathbf{x}$ to the corresponding bits of $\mathbf{y}$ one at a time, moving from left to right.

The length of every canonical path is $|\gamma_{\mathbf{xy}}| = n$. An edge e in the underlying graph of the chain connects two states if and only if they differ by one bit. Since $\pi(\mathbf{x}) = 2^{-n}$ for any $\mathbf{x} \in \Omega$, it follows that $Q(e)$ is constant across any edge. Therefore, to obtain $\rho(\Gamma)$ from (4.20), it suffices to compute the number of canonical paths using any edge.

Suppose that $e = (\mathbf{w}, \mathbf{w}')$, where $\mathbf{w}$ and $\mathbf{w}'$ only differ in the ith bit for some $1 \leq i \leq n$. Since our canonical paths sweep from left to right, it follows that if $e \in \gamma_{\mathbf{xy}}$, then the last $n - i + 1$ bits of $\mathbf{x}$ must be the same as $\mathbf{w}$. This leaves the first $i - 1$ bits of $\mathbf{x}$ free to vary over 2^{i-1} choices. Similarly, the first i bits of $\mathbf{y}$ must be the same as $\mathbf{w}'$, which leaves the last $n - i$ bits of $\mathbf{y}$ free to vary over 2^{n-i} choices. Thus, the total number of canonical paths using any particular edge is 2^{n-1}.

Since $Q(e) = \pi(\mathbf{w})P(\mathbf{w}, \mathbf{w}') = 2^{-n}(2n)^{-1}$ (recall that the chain is lazy),

$$\rho(\Gamma) = 2^n(2n) \cdot (2^{-2n}n) \cdot 2^{n-1} = n^2.$$

Therefore, by Corollary 4.22,

$$t_{\mathrm{mix}}(\epsilon) \leq \lceil n^2(n \log 2 + \log \epsilon^{-1}) \rceil.$$

This bound is worse than the $O(n \log n + n \log \epsilon^{-1})$ bound from coupling. The extra slack is typical of this method. However, as noted, its flexibility allows it to tackle particularly complicated combinatorial problems. (References to papers on such problems, some of which have used the canonical paths or multicommodity flows method, are given in Chapter 5.)

Chapter 5

Connection Between Sampling and Counting

In this chapter, we will describe an interesting application of rapidly mixing Markov chains to the counting of combinatorial objects. The majority of these counting problems are "hard", and belong to the complexity class #P [55], which is analogous to the class NP of "hard" decision problems. It is an open question whether all such counting problems are efficiently computable (i.e. in polynomial time). However, randomised algorithms have been developed that can approximately count (within arbitrarily small error), by using the MCMC method.

Definition 5.1. A randomised approximation scheme for a counting problem is a randomised algorithm that, given any error tolerance $\epsilon > 0$, outputs an estimate $\widehat{N}$ of the number of instances N, such that

$$\mathbb{P}\left((1-\epsilon)N \leq \widehat{N} \leq (1+\epsilon)N\right) \geq \frac{3}{4}. \tag{5.1}$$

If the runtime of the algorithm is bounded by a polynomial in the size of the input length n (e.g. number of vertices of a graph) and ϵ^{-1}, then we call it a **fully polynomial randomised approximation scheme** (**FPRAS**).

Remark 5.2. The tolerance $\frac{3}{4}$ in the definition is mostly arbitrary. To achieve any desired tolerance of $1-\eta > 0$ instead, it suffices to run an algorithm satisfying (5.1) $12\lceil\log(\eta^{-1})\rceil+1$ times, and then taking the median. (The proof of this relies on Chernoff's inequality and can be found in [35, Lemma 6.1].)

A FPRAS has been shown to exist for some of the most difficult counting problems, notably approximating the permanent of a 0-1 matrix (which is equivalent to counting the number of perfect matchings in a bipartite graph) [31, 34], and approximating the volume of a convex body in $\mathbb{R}^n$ [20]. Other examples include counting matchings [33], counting 0-1 knapsack solutions [47], and approximating intractable quantities in models used in statistical physics [32]. For a more comprehensive treatment of the intimate connection between approximate counting and sampling, see Jerrum [30].

At the core of these MCMC algorithms is a rapidly mixing Markov chain. The remainder of this chapter will focus on the problem of counting the number of q-colourings of a graph with maximum degree Δ. It is well-known that a colouring is always possible if there are $q \geq \Delta+1$ colours. However, the exact counting problem is in #P. The existence of a FPRAS when $q \geq 2\Delta+1$ was shown by Jerrum [29]. A faster algorithm is exhibited by Dyer and Greenhill [21]. A FPRAS for $q > \frac{11}{6}\Delta$ was shown by Vigoda [56], and a

slight improvement to this bound was obtained more recently in [11]. It remains an open question whether a FPRAS always exists when $q \geq \Delta + 1$.

We will provide a proof of the result when $q \geq 2\Delta + 1$ using path coupling, which was first shown by Bubley and Dyer [7]. We will consider the (heat-bath) Glauber dynamics, instead of the usual Metropolis chain used in the literature. The only difference is that we will choose conditional on the set of permissible colours, avoiding the need to reject invalid colours. Similar results are obtained.

Let $G = (V, E)$ be a graph with n vertices, m edges, and maximum degree Δ. For $q \in \mathbb{Z}_+$, let $\mathcal{C} = \{1, 2, \ldots, q\}$ be a set of q colours, and $\Omega \subseteq \mathcal{C}^V$ be the set of all proper vertex q-colourings of G (each colouring $\sigma \in \Omega$ is precisely a function from V to $\mathcal{C}$). For a colouring σ, let $\mathcal{N}_v(\sigma) = \{\sigma(w) : \{v, w\} \in E\} \subseteq \mathcal{C}$ be the set of colours of the neighbours of v. The main result of this chapter is the following theorem.

Theorem 5.3. *If $q \geq 2\Delta + 1$, then there exists a FPRAS for estimating the number of proper vertex q-colourings of G, with runtime bounded by*

$$t_{\text{mix}}(\epsilon) \leq \left\lceil \frac{1 - \Delta/q}{1 - 2\Delta/q} n(\log n + \log \epsilon^{-1}) \right\rceil \cdot \left\lceil \frac{37m}{\epsilon^2} \right\rceil. \tag{5.2}$$

This will follow from combining Lemma 5.4 and Lemma 5.5. The term on the left of (5.2) corresponds to the mixing time of a Markov chain to obtain an "almost uniform" sample, and the term on the right gives the number of independent samples needed.

Consider the Glauber dynamics on Ω. Given the current state σ, a vertex v is selected uniformly at random (with probability $1/n$). A colour is then selected from the set of permissible colours $\mathcal{C}_v(\sigma) = \{c \in \mathcal{C} : c \notin \mathcal{N}_v(\sigma)\}$ uniformly at random (with probability $1/|\mathcal{C}_v(\sigma)|$) to recolour $\sigma(v)$.

This is aperiodic, since we can always leave the colouring unchanged. This is also irreducible, since we can move from σ to τ by sequentially recolouring each vertex (in lexicographic order). For example, to recolour $\sigma(v)$ to $\tau(v)$, we first recolour all neighbouring vertices $u > v$ with $\sigma(u) = \tau(v)$. (If $q \geq \Delta + 2$, we have enough colours for this to always be possible.) Furthermore, it can easily be verified that the chain is reversible with respect to the uniform distribution $\pi(\sigma) = |\Omega|^{-1}$ for all $\sigma \in \Omega$. Hence, this is the stationary distribution by Proposition 2.6.

Lemma 5.4. *The Glauber dynamics for sampling random proper q-colourings of a graph G with n vertices and maximum degree Δ satisfying $q \geq 2\Delta + 1$ is rapidly mixing, with*

$$t_{\text{mix}}(\epsilon) \leq \left\lceil \frac{1 - \Delta/q}{1 - 2\Delta/q} n(\log n + \log \epsilon^{-1}) \right\rceil. \tag{5.3}$$

Proof. We will apply the path coupling method from Section 3.2. Consider the Glauber dynamics, described above, on the extended state space $\widetilde{\Omega} = \mathcal{C}^V$ of *all* q-colourings of the graph G. Note that this is no longer irreducible. However, since all the extra non-proper

colourings are transient states, the extended chain still converges to a unique stationary distribution π, which is uniform on the set of all proper q-colourings Ω, and satisfies $\pi(\sigma) = 0$ for any non-proper colouring σ (see [38, Proposition 1.26]).

For the underlying graph, connect two colourings $\sigma, \tau \in \widetilde{\Omega}$ with an edge (of length 1) if and only if they differ by one vertex. Thus, the path metric $p(\sigma, \tau) = \sum_{v \in V} \mathbb{1}_{\{\sigma(v) = \tau(v)\}}$ counts the number of vertices that disagree.

Let $\sigma, \tau \in \widetilde{\Omega}$ be two colourings that agree everywhere except at one vertex, say v. Consider the following coupling (X, Y) of one step of the chain, started at σ and τ respectively. Choose a vertex w uniformly at random, and consider cases:

- If $w = v$, or w is not a neighbour of v: recolour w for both chains with the same colour, chosen uniformly at random from the set of permissible colours $\mathcal{C}_w(\sigma) = \mathcal{C}_w(\tau)$.
- If w is a neighbour of v: let $\mathcal{C}_w^{-(v,w)}(\sigma)$ (resp. $\mathcal{C}_w^{-(v,w)}(\tau)$) be the set of permissible colours for σ (resp. τ) at w, ignoring the edge from w to v. Since σ and τ agree everywhere except at v, $\mathcal{C}_w^{-(v,w)}(\sigma) = \mathcal{C}_w^{-(v,w)}(\tau)$. Consider four subcases:
  (i) $\sigma(v), \tau(v) \notin \mathcal{C}_w^{-(v,w)}(\sigma)$: in this case, adding in the edge $\{w, v\}$ will not change the permissible colours at w. Hence, we can recolour w for both chains with the same colour, chosen uniformly at random from $\mathcal{C}_w(\sigma) = \mathcal{C}_w(\tau)$.
  (ii) $\sigma(v), \tau(v) \in \mathcal{C}_w^{-(v,w)}(\sigma)$: here $\mathcal{C}_w(\sigma) = \mathcal{C}_w^{-(v,w)}(\sigma) \setminus \{\sigma(v)\}$, and by symmetry, $\mathcal{C}_w(\tau) = \mathcal{C}_w^{-(v,w)}(\sigma) \setminus \{\tau(v)\}$. Hence, we can choose a colour c uniformly at random from $\mathcal{C}_w(\sigma)$ to recolour $\sigma(w)$. Use c to recolour $\tau(w)$ unless $c = \tau(v)$, in which case use $\sigma(v)$ to recolour $\tau(w)$ instead.
  (iii) $\sigma(v) \notin \mathcal{C}_w^{-(v,w)}(\sigma)$, $\tau(v) \in \mathcal{C}_w^{-(v,w)}(\sigma)$: after adding back in the edge $\{w, v\}$, $\mathcal{C}_w(\sigma)$ will have the extra colour $\tau(v)$ compared to $\mathcal{C}_w(\tau)$. In this case, select a colour c uniformly at random from $\mathcal{C}_w(\sigma)$ to recolour $\sigma(w)$. Use c to recolour $\tau(w)$ unless $c = \tau(v)$, in which case select another colour c', independently and uniformly at random, from $\mathcal{C}_w(\tau)$ to recolour $\tau(w)$ instead.
  Since $|\mathcal{C}_w(\sigma)| = |\mathcal{C}_w(\tau)| + 1$, we can check that each permissible colour in $\mathcal{C}_w(\tau)$ is indeed selected uniformly at random, with probability

$$\frac{1}{|\mathcal{C}_w(\sigma)|} + \frac{1}{|\mathcal{C}_w(\sigma)|} \cdot \frac{1}{|\mathcal{C}_w(\tau)|} = \frac{1}{|\mathcal{C}_w(\sigma)|} \cdot \frac{|\mathcal{C}_w(\tau)| + 1}{|\mathcal{C}_w(\tau)|} = \frac{1}{|\mathcal{C}_w(\tau)|}.$$

  (iv) $\sigma(v) \in \mathcal{C}_w^{-(v,w)}(\sigma)$, $\tau(v) \notin \mathcal{C}_w^{-(v,w)}(\sigma)$: this is the same as the previous case (iii) after swapping the roles of σ and τ.

  After one step of the chain, the number of vertices that disagree either:

  - Increases by 1 if w is a neighbour of v, and different colours are picked for $\tau(w)$ and $\sigma(w)$. In case (ii), this occurs with probability $\frac{\deg(v)}{n}$. In case (iii), this occurs with probability $\frac{\deg(v)}{n} \cdot \frac{1}{|\mathcal{C}_w(\sigma)|}$ (this is the same for case (iv), swapping σ for τ).
  - Decreases by 1 if $w = v$, with probability $\frac{1}{n}$.
  - Remains unchanged otherwise if $w \neq v$, and w is recoloured for both chains with the same colour.

Since $\deg(v) \leq \Delta$, and $|\mathcal{C}_w(\sigma)| \geq q - \Delta$, using the inequality $1 - x \leq e^{-x}$ for $x \geq 0$ shows that the expected path metric after one step is

$$\mathbb{E}_{\sigma,\tau}\left[\rho(X,Y)\right] \leq 1 - \frac{1}{n} + \frac{\Delta}{n} \cdot \frac{1}{q-\Delta} = 1 - \frac{1}{n}\left(1 - \frac{\Delta}{q-\Delta}\right) \leq \exp\left(-\frac{1}{n} \cdot \frac{q-2\Delta}{q-\Delta}\right).$$

Since $q > 2\Delta$, the chain is contracting, with $\alpha = \frac{1}{n} \cdot \frac{1-2\Delta/q}{1-\Delta/q}$. Therefore, Corollary 3.13 of the path coupling theorem implies (5.3), as desired. □

Lemma 5.5. *Suppose that we have an almost uniform sampler p for proper q-colourings of a graph G with runtime bounded by $T(n,\delta)$, where n is the number of edges of G and $\delta > 0$ is the tolerance. (That is, p satisfies $\|p - \pi\|_{\mathrm{TV}} < \delta$, where π is uniform on the set of all proper q-colourings of G).*

If G is a graph with n vertices, m edges, and maximum degree Δ satisfying $q \geq \Delta + 1$, then for any $\epsilon > 0$, there exists a randomised approximation scheme (5.1) for the number of proper q-colourings of G, with runtime bounded by $\lceil 37m\epsilon^{-2} \rceil\, T(n, \epsilon/(6m))$.

Proof. Let $(V, \emptyset) = G_0 < G_1 < \cdots < G_{m-1} < G_m = G$ be any sequence of graphs, in which G_{i-1} is obtained from G_i by removing a single edge. Let $\Omega(G_i)$ denote the set of all proper q-colourings of G_i. We can express the number of proper q-colourings of G as

$$|\Omega(G)| = \frac{|\Omega(G_m)|}{|\Omega(G_{m-1})|} \times \frac{|\Omega(G_{m-1})|}{|\Omega(G_{m-2})|} \times \cdots \times \frac{|\Omega(G_1)|}{|\Omega(G_0)|} \times |\Omega(G_0)|. \tag{5.4}$$

Note that $|\Omega(G_0)| = q^n$. By the self-reducibility of the problem, we will be able to estimate $|\Omega(G)|$ by estimating each ratio in (5.4) with a Monte Carlo-style estimator. Let

$$\rho_i = \frac{|\Omega(G_i)|}{|\Omega(G_{i-1})|}, \quad 1 \leq i \leq m. \tag{5.5}$$

Then $|\Omega(G)| = q^n \rho_1 \rho_2 \ldots \rho_m$.

Observe that $\Omega(G_i) \subseteq \Omega(G_{i-1})$. Suppose that G_{i-1} is obtained from G_i by removing the edge $\{u, v\}$. Consider a colouring in $\Omega(G_{i-1}) \setminus \Omega(G_i)$, which necessarily assigns the same colour to u and v. This can be perturbed into a colouring in $\Omega(G_i)$ by recolouring u with one of at least $q - \Delta \geq 1$ colours. Moreover, each colouring in $\Omega(G_i)$ can be obtained in at most one way by such a perturbation. Thus, $|\Omega(G_{i-1}) \setminus \Omega(G_i)| \leq |\Omega(G_i)|$. Since $|\Omega(G_{i-1})| = |\Omega(G_i)| + |\Omega(G_{i-1}) \setminus \Omega(G_i)|$, we deduce that $\frac{1}{2} \leq \rho_i \leq 1$.

Consider the following procedure to estimate ρ_i. Generate an almost uniform sample σ_i from $\Omega(G_{i-1})$ by running p with tolerance $\delta = \epsilon/(6m)$. Let Z_i be the Bernoulli random variable $\mathbb{1}_{\{\sigma_i \in \Omega(G_i)\}}$, which has mean $\mu_i := \mathbb{E}\left[Z_i\right]$, and variance $\mathrm{Var}(Z_i) = \mu_i(1 - \mu_i)$. If π_{i-1} is uniform on $\Omega(G_{i-1})$, then by definition of the total variation distance,

$$|\mu_i - \rho_i| = \left| p(\Omega_i) - \frac{|\Omega(G_i)|}{|\Omega(G_{i-1})|} \right| = |p(\Omega_i) - \pi_{i-1}(\Omega_i)| \leq \frac{\epsilon}{6m}. \tag{5.6}$$

Recalling that $\rho_i \geq \frac{1}{2}$, this implies that

$$\left(1 - \frac{\epsilon}{3m}\right)\rho_i \leq \mu_i \leq \left(1 + \frac{\epsilon}{3m}\right)\rho_i. \tag{5.7}$$

Thus, the sample mean $\overline{Z}_i = \frac{1}{s}\sum_{k=1}^{s} Z_i^{(k)}$ of sufficiently many independent copies $Z_i^{(1)}, Z_i^{(2)}, \ldots, Z_i^{(s)}$ of Z_i should provide a good estimate of ρ_i. More specifically, let $s = \lceil 37m\epsilon^{-2} \rceil$. Since the $Z_i^{(k)}$ are i.i.d. Bernoulli(μ_i) random variables,

$$\frac{\mathrm{Var}(\overline{Z}_i)}{\mu_i^2} = \frac{\mu_i(1-\mu_i)}{s\mu_i^2} \leq \frac{2}{s} - \frac{1}{s} = \frac{1}{s}. \tag{5.8}$$

Consider the estimator $\widehat{N} = q^n \overline{Z}_1 \overline{Z}_2 \cdots \overline{Z}_m$ for $|\Omega(G)|$. Since the $Z_i^{(k)}$ are independent, $\mathbb{E}\left[\widehat{N}\right] = q^n \mu_1 \mu_2 \cdots \mu_m$, which is close to $|\Omega(G)|$ by (5.7). Hence, its performance will largely depend upon its variance. Using $\mathrm{Var}(\overline{Z}_i) = \mathbb{E}\left[\overline{Z}_i^2\right] - \mu_i^2$,

$$\frac{\mathrm{Var}(\overline{Z}_1\overline{Z}_2\cdots\overline{Z}_m)}{(\mu_1\mu_2\cdots\mu_m)^2} = \frac{\mathbb{E}\left[\overline{Z}_1\overline{Z}_2\cdots\overline{Z}_m\right]}{\mu_1^2\mu_2^2\cdots\mu_m^2} - 1 = \prod_{i=1}^{m}\left(1 + \frac{\mathrm{Var}(\overline{Z}_i)}{\mu_i^2}\right) - 1$$

By (5.8), and recalling that we chose $s \geq 37m\epsilon^{-2}$ samples, this is upper bounded by $\left(1 + \frac{1}{s}\right)^m - 1 \leq \exp\left(\frac{\epsilon^2}{37}\right) - 1 \leq \frac{\epsilon^2}{36}$, using the inequality $e^{x/(k+1)} \leq 1 + \frac{x}{k}$ for any $0 \leq x \leq 1$ and $k \in \mathbb{Z}_+$. Thus, by Chebyshev's inequality (Theorem 1.2),

$$\mathbb{P}\left(\left|\frac{\overline{Z}_1\overline{Z}_2\cdots\overline{Z}_m}{\mu_1\mu_2\cdots\mu_m} - 1\right| > \frac{\epsilon}{3}\right) \leq \frac{\mathrm{Var}(\overline{Z}_1\overline{Z}_2\cdots\overline{Z}_m)}{\mu_1^2\mu_2^2\cdots\mu_m^2} \cdot \frac{9}{\epsilon^2} \leq \frac{1}{4}.$$

Hence, with probability at least $\frac{3}{4}$, the following inequality holds:

$$\left(1 - \frac{\epsilon}{3}\right) q^n \mu_1\mu_2\cdots\mu_m \leq q^n \overline{Z}_1\overline{Z}_2\cdots\overline{Z}_m \leq \left(1 + \frac{\epsilon}{3}\right) q^n \mu_1\mu_2\cdots\mu_m. \tag{5.9}$$

Furthermore, using (5.7) and the inequalities $\left(1 + \frac{\epsilon}{3m}\right)^m \leq e^{\epsilon/3} \leq 1 + \frac{\epsilon}{2}$ again, also shows that

$$\left(1 - \frac{\epsilon}{2}\right)\rho_1\rho_2\cdots\rho_m \leq \mu_1\mu_2\cdots\mu_m \leq \left(1 + \frac{\epsilon}{2}\right)\rho_1\rho_2\cdots\rho_m. \tag{5.10}$$

Since $\left(1 + \frac{\epsilon}{3}\right)\left(1 + \frac{\epsilon}{2}\right) \leq 1 + \epsilon$, and $\left(1 - \frac{\epsilon}{3}\right)\left(1 - \frac{\epsilon}{2}\right) \geq 1 - \epsilon$, combining the final two inequalities (5.9) and (5.10) shows that with probability at least $\frac{3}{4}$, the estimator $\widehat{N}$ for $|\Omega(G)|$ satisfies $(1-\epsilon)|\Omega(G)| \leq \widehat{N} \leq (1+\epsilon)|\Omega(G)|$.

Therefore, we have a FPRAS for the number of proper q-colourings. The total runtime is bounded by the runtime to generate each almost uniform sample, which is at most $T(n, \delta)$, multiplied by the number of independent samples $s = \lceil 37m\epsilon^{-2} \rceil$ needed. □

Chapter 6

The Cutoff Phenomenon

In this chapter, we will describe the cutoff phenomenon, which describes how certain Markov chains have been observed to show a very sharp transition to stationarity. In other words, mixing occurs abruptly at, and not before, a certain point (akin to a "probabilistic phase transition"), as illustrated in Figure 6.1.1. Proving that a Markov chain exhibits a cutoff provides a rigorous stopping rule for MCMC samplers. It is also of theoretical interest as a seemingly general phenomenon that is not yet well understood.

Section 6.1 will describe a motivating example of riffle shuffling, and give other examples of cutoff in the literature. Section 6.2 will provide and prove the equivalences of various precise definitions of cutoff. In this chapter, we will denote a sequence of Markov chains by $(\Omega_n, P_n, \pi_n)_{n\in\mathbb{Z}_+}$. The standard notation will be used alongside an additional index n, which is usually a parameter of increasing size or complexity of the chain.

6.1 Motivation

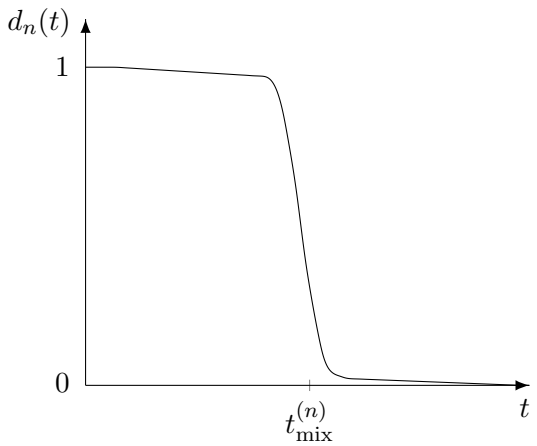

Figure 6.1.1: The total variation distance of a sequence of chains that exhibits a cutoff drops precipitously after a certain point. Prior to that, it is poorly mixed.

Recall the Gilbert-Shannon-Reeds model of the riffle shuffle of n cards in Example 3.22. The method of strong stationary times was used to find that $t_{\text{mix}}^{(n)} = O(\log_2 n)$. The paper of Bayer and Diaconis [6] provides a far more precise description of this riffle shuffle. We will summarise some of their key results that allow sharp bounds on the distance to stationarity to be obtained and used to demonstrate a cutoff.

The paper generalises the riffle shuffle to an a-shuffle (the normal riffle shuffle is a 2-shuffle). Several equivalent representations of the a-shuffle are given, which are used to

show that an a-shuffle followed by a b-shuffle has the same distribution as an ab-shuffle. Using a very nice combinatoric argument, the exact probability of any arrangement of cards after t shuffles (equivalent to a 2^t-shuffle) is obtained.

Theorem 6.1 ([6, Theorem 1]). *If a pack of n cards is riffle shuffled t times, then the probability that the deck is in arrangement σ is $\frac{1}{2^{tn}}\binom{2^t+n-r}{n}$, where r is the number of rising sequences in σ.*

(Recall from Example 3.22 that a rising sequence is a maximal subset of an arrangement of cards consisting of successive face values when read in order.) By Theorem 6.1, the number of rising sequences is a sufficient statistic for the permutation distribution of the riffle shuffle. Since these are counted by the Eulerian numbers, which have well-known asymptotics, an asymptotic expression for the distance to stationarity can be computed.

Theorem 6.2 ([6, Theorem 4]). *Let $t = \frac{3}{2}\log_2(n)+\alpha$, where $\alpha > 0$ is a fixed real number. Let $c = 2^\alpha$. Then, as $n \to \infty$,*

$$d_n(t) = \left\|P_n^t(\mathrm{id}, \cdot) - \pi_n\right\|_{\mathrm{TV}} = 1 - 2\Phi\left(\frac{-1}{4c\sqrt{3}}\right) + O_c\left(\frac{1}{n^{1/4}}\right).$$

Here π_n is the uniform distribution on S_n, and $\Phi(x) = \int_{-\infty}^{x} \frac{1}{\sqrt{2\pi}} e^{-t^2/2}\,\mathrm{d}t$ is the standard normal cumulative distribution function.

Therefore, for sufficiently large n, $d_n(\frac{3}{2}\log_2 n + \alpha) \to 0$ as $\alpha \to \infty$, which shows an abrupt transition to stationarity after $\frac{3}{2}\log_2 n$ shuffles. Before this point, the deck is far from uniform, since $d_n(\frac{3}{2}\log_2 n - \alpha) \to 1$ as $\alpha \to \infty$. Hence, it appears that $\frac{3}{2}\log_2 n$ riffle shuffles are, in this sense, necessary and sufficient to mix up a pack of n cards.

We will conclude this section by describing other instances of cutoff that have been found in the literature. The cutoff phenomenon was first identified for the random transpositions chain (shuffling n cards by repeatedly swapping two cards) by Diaconis and Shahshahani [16] using group representations, where a cutoff at $\frac{1}{2}n\log n$ with window n was found. Aldous and Diaconis [2] found that the top-to-random shuffle (repeatedly inserting the top card into the deck randomly) also has the same cutoff point and window (and also coined the term cutoff).

Diaconis' survey [13] describes a wealth of other examples for random walks on finite groups, including shuffling, the Ehrenfests' urn, and the n-dimensional hypercube. For example, the simple (lazy) random walk on the hypercube (also described in [38]) has a cutoff at $\frac{1}{2}n\log n$ with window n. It is suggested that the cutoff phenomenon may be related to high multiplicity of the second eigenvalue.

Lubetzky and Sly [40] found that, with high probability, the simple random walk on the random regular graph $G \sim \mathcal{G}(n,d)$ (i.e. G is uniformly distributed over the set of all d-regular graphs with n vertices) exhibits a cutoff at $\frac{d}{d-2}\log_{d-1} n$ with window $\sqrt{\log n}$. Chatterjee et al. [9] also found that the repeated averages chain (which, given n real numbers, repeatedly replaces two numbers with their average) exhibits a cutoff at

$\frac{1}{2}n\log_2 n$ with window $n\sqrt{\log n}$. A cutoff has also been found for the Glauber dynamics for the Ising model from statistical physics [37, 42], which is the focus of the next chapter.

As a relatively recent discovery, there are still many open questions about what causes cutoff. The product condition (Theorem 6.13) has been found to be necessary and sufficient for cutoff for the class of birth-and-death chains [19]. Cutoff has also been proved in less common distances (such as the separation and L^p distances) [15, 10]. More recently, cutoff has been characterised for reversible chains in terms of the concentration of some hitting time [5].

Rigorously proving cutoff is a difficult, delicate affair that requires precise upper and lower bounds on the distance to stationarity. Many of the examples given above require more sophisticated techniques, beyond that described in this thesis. The next chapter will focus on proving cutoff for the Glauber dynamics for the mean-field Ising model.

6.2 Definitions of cutoff

In this section, we will give some precise definitions of cutoff that have appeared in the literature. We will also provide some original proofs of their equivalences in order to unify the various characterisations.

Definition 6.3. We say that a sequence of Markov chains $(\Omega_n, P_n, \pi_n)_{n\in\mathbb{Z}_+}$ exhibits a **cutoff** if it satisfies either of the following:

(i) Let $c > 0$ be a fixed constant. Then

$$\begin{aligned} \lim_{n\to\infty} d_n(\lceil ct_{\text{mix}}^{(n)}\rceil) &= 0 \quad \text{if } c > 1, \\ \lim_{n\to\infty} d_n(\lfloor ct_{\text{mix}}^{(n)}\rfloor) &= 1 \quad \text{if } c < 1. \end{aligned} \tag{6.1}$$

(ii) For any $0 < \epsilon < 1/2$,

$$\lim_{n\to\infty} \frac{t_{\text{mix}}^{(n)}(\epsilon)}{t_{\text{mix}}^{(n)}(1-\epsilon)} = 1. \tag{6.2}$$

Proposition 6.4. *If $t_{\text{mix}}^{(n)} \to \infty$ as $n \to \infty$, then conditions (i) and (ii) for cutoff given in Definition 6.3 are equivalent.*

Proof. Suppose that (i) holds. Let $0 < \epsilon < 1/2$ and $\delta > 0$. Then for sufficiently large n, $d_n(\lceil(1+\delta)t_{\text{mix}}^{(n)}\rceil) < \epsilon$, and hence $t_{\text{mix}}^{(n)}(\epsilon) \le (1+\delta)t_{\text{mix}}^{(n)}+1$. Similarly, $d_n(\lfloor(1-\delta)t_{\text{mix}}^{(n)}\rfloor) > 1-\epsilon$, which implies that $t_{\text{mix}}^{(n)}(1-\epsilon) \ge (1-\delta)t_{\text{mix}}^{(n)} - 1$. Therefore, since $t_{\text{mix}}^{(n)}(\epsilon) \ge t_{\text{mix}}^{(n)}(1-\epsilon)$ for $0 < \epsilon < 1/2$, this shows that

$$1 \le \frac{t_{\text{mix}}^{(n)}(\epsilon)}{t_{\text{mix}}^{(n)}(1-\epsilon)} \le \frac{1+\delta+(1/t_{\text{mix}}^{(n)})}{1-\delta-(1/t_{\text{mix}}^{(n)})}.$$

By assumption, $1/t_{\text{mix}}^{(n)}$ tends to zero as n tends to infinity. Thus, taking the limit as $\delta \to 0$ shows that $\lim_{n\to\infty} \frac{t_{\text{mix}}^{(n)}(\epsilon)}{t_{\text{mix}}^{(n)}(1-\epsilon)} = 1$ by the sandwich theorem.

Conversely, suppose that (ii) holds. Let $0 < \epsilon < 1/2$. Then for any $\delta > 0$ and sufficiently large n, $t_{\text{mix}}^{(n)}(\epsilon) < (1+\delta)t_{\text{mix}}^{(n)}(1-\epsilon)$. Since $t_{\text{mix}}^{(n)}(1-\epsilon) \leq t_{\text{mix}}^{(n)}$, by definition of the mixing time, this shows that $\lim_{n\to\infty} d_n(\lceil(1+\delta)t_{\text{mix}}^{(n)}\rceil) \leq \epsilon$. Since this holds for all $0 < \epsilon < 1/2$, $\lim_{n\to\infty} d_n(\lceil(1+\delta)t_{\text{mix}}^{(n)}\rceil) = 0$. The proof for the other side of the cutoff proceeds analogously using $t_{\text{mix}}^{(n)}(1-\epsilon) > (1-\delta)t_{\text{mix}}^{(n)}(\epsilon)$ for sufficiently large n. □

Remark 6.5. The definition of cutoff in condition (ii) may be slightly relaxed. We say that a sequence of Markov chains $(\Omega_n, P_n, \pi_n)_{n\in\mathbb{Z}_+}$ exhibits a **pre-cutoff** if, for any $0 < \epsilon < 1/2$,

$$\limsup_{n\to\infty} \frac{t_{\text{mix}}^{(n)}(\epsilon)}{t_{\text{mix}}^{(n)}(1-\epsilon)} < \infty. \tag{6.3}$$

It is clear that having a pre-cutoff is a necessary condition for cutoff. However, an example of a chain that exhibits a pre-cutoff but not a cutoff has been constructed by Aldous (see Figure 18.2 of [38]). Hence, having a pre-cutoff is not sufficient for having a cutoff.

The next definition allows us to specify the size of the window at which cutoff occurs.

Definition 6.6. Let $(w_n)_{n\in\mathbb{Z}_+}$ be a sequence of positive numbers. Then we say that the sequence of Markov chains $(\Omega_n, P_n, \pi_n)_{n\in\mathbb{Z}_+}$ exhibits a **cutoff with window** w_n if $w_n = o(t_{\text{mix}}^{(n)})$, and it satisfies either of the following:

(i)
$$\begin{aligned} \lim_{\alpha\to\infty} \limsup_{n\to\infty} d_n(\lceil t_{\text{mix}}^{(n)} + \alpha w_n\rceil) &= 0, \\ \lim_{\alpha\to\infty} \liminf_{n\to\infty} d_n(\lfloor t_{\text{mix}}^{(n)} - \alpha w_n\rfloor) &= 1. \end{aligned} \tag{6.4}$$

(ii) $t_{\text{mix}}^{(n)}(\epsilon) = t_{\text{mix}}^{(n)}(1-\epsilon) + O(w_n)$: that is, for any $0 < \epsilon < 1/2$, there exists a constant $c_\epsilon > 0$ such that for all $n \in \mathbb{Z}_+$,

$$t_{\text{mix}}^{(n)}(\epsilon) - t_{\text{mix}}^{(n)}(1-\epsilon) \leq c_\epsilon w_n. \tag{6.5}$$

Proposition 6.7. *If $t_{\text{mix}}^{(n)} \to \infty$ and $w_n \to \infty$ as $n \to \infty$, then conditions (i) and (ii) for cutoff with window w_n given in Definition 6.6 are equivalent.*

Proof. Suppose that (i) holds. Let $0 < \epsilon < 1/2$. Then there exists a real A and positive integer N (depending only on ϵ) such that $d_n(\lceil t_{\text{mix}}^{(n)} + \alpha w_n\rceil) < \epsilon$ for all $\alpha \geq A$ and $n > N$. Hence, $t_{\text{mix}}^{(n)}(\epsilon) \leq t_{\text{mix}}^{(n)} + Aw_n + 1$. Similarly, there exists a real $B \leq A$ such that $t_{\text{mix}}^{(n)}(1-\epsilon) \geq t_{\text{mix}}^{(n)} + Bw_n - 1$. Therefore, for sufficiently large n,

$$t_{\text{mix}}^{(n)}(\epsilon) - t_{\text{mix}}^{(n)}(1-\epsilon) \leq (A-B)w_n + 2 \leq (A-B+2)w_n.$$

The last inequality follows from the assumption that $w_n \to \infty$. Hence, we can choose a constant c_ϵ such that $t_{\text{mix}}^{(n)}(\epsilon) - t_{\text{mix}}^{(n)}(1-\epsilon) \leq c_\epsilon w_n$ for all $n \in \mathbb{Z}_+$, by setting c_ϵ to be the greater of $(A-B+2)$ and $\max_{n=1,2,\ldots,N}\left\{\frac{t_{\text{mix}}^{(n)}(\epsilon) - t_{\text{mix}}^{(n)}(1-\epsilon)}{w_n}\right\}$.

Conversely, suppose that (ii) holds. Let $0 < \epsilon < 1/2$. Then for some constant $c_\epsilon > 0$,

$$t_{\text{mix}}^{(n)}(\epsilon) \leq t_{\text{mix}}^{(n)}(1-\epsilon) + c_\epsilon w_n \leq t_{\text{mix}}^{(n)} + c_\epsilon w_n, \quad \text{for all } n \in \mathbb{Z}_+.$$

This shows that $d(\lceil t_{\text{mix}}^{(n)} + c_\epsilon w_n \rceil) \leq \epsilon$. Since $d(\lceil t_{\text{mix}}^{(n)} + \alpha w_n \rceil) \leq d(\lceil t_{\text{mix}}^{(n)} + c_\epsilon w_n \rceil)$ for all $\alpha > c_\epsilon$, it follows that $\lim_{\alpha\to\infty} \limsup_{n\to\infty} d(\lceil t_{\text{mix}}^{(n)} + \alpha w_n \rceil) \leq \epsilon$. Since this holds for all $0 < \epsilon < 1/2$, the first part of (i) holds. The second part of (i) follows analogously. □

Proposition 6.8. *If a sequence of Markov chains $(\Omega_n, P_n, \pi_n)_{n\in\mathbb{Z}_+}$ exhibits a cutoff with window w_n as in condition (i) (resp. (ii)) of Definition 6.6, then it also exhibits a cutoff as in condition (i) (resp. (ii)) of Definition 6.3.*

Proof. Suppose that we have a cutoff with window $w_n = o(t_{\text{mix}}^{(n)})$ as in condition (ii). Then for any $0 < \epsilon < 1/2$, there exists a $c_\epsilon > 0$ such that for all $n \in \mathbb{Z}_+$,

$$0 \leq 1 - \frac{t_{\text{mix}}^{(n)}(1-\epsilon)}{t_{\text{mix}}^{(n)}(\epsilon)} \leq c_\epsilon \frac{w_n}{t_{\text{mix}}^{(n)}(\epsilon)}.$$

First, suppose that $0 < \epsilon < 1/4$. Since $t_{\text{mix}}^{(n)} \leq t_{\text{mix}}^{(n)}(\epsilon)$, it follows that $w_n = o(t_{\text{mix}}^{(n)}(\epsilon))$, and hence $\lim_{n\to\infty} \frac{w_n}{t_{\text{mix}}^{(n)}(\epsilon)} = 0$. Therefore, $\lim_{n\to\infty} \frac{t_{\text{mix}}^{(n)}(\epsilon)}{t_{\text{mix}}^{(n)}(1-\epsilon)} = 1$ by sandwiching. Next, for $1/4 \leq \epsilon < 1/2$, this follows from the first part, since $0 < \frac{\epsilon}{2} < 1/4$, and

$$1 \leq \frac{t_{\text{mix}}^{(n)}(\epsilon)}{t_{\text{mix}}^{(n)}(1-\epsilon)} \leq \frac{t_{\text{mix}}^{(n)}(\epsilon/2)}{t_{\text{mix}}^{(n)}(1-\epsilon/2)}.$$

Therefore $\lim_{n\to\infty} \frac{t_{\text{mix}}^{(n)}(\epsilon)}{t_{\text{mix}}^{(n)}(1-\epsilon)} = 1$ for all $0 < \epsilon < 1/2$, so we have cutoff as in condition (ii).

If $t_{\text{mix}}^{(n)} \to \infty$ and $w_n \to \infty$, then conditions (i) and (ii) for both definitions are equivalent. Otherwise, it can also be shown directly that condition (i) for a cutoff with window implies condition (i) for a cutoff, however we will omit the details. □

A cutoff has been defined quite conceptually so far in terms of the mixing time, for which we usually do not have an explicit expression. The following definitions allow us to specify the point at which cutoff occurs, and hence are more useful in practice.

Definition 6.9. Consider a sequence of Markov chains $(\Omega_n, P_n, \pi_n)_{n\in\mathbb{Z}_+}$. Let $(a_n)_{n\in\mathbb{Z}_+}$ and $(w_n)_{n\in\mathbb{Z}_+}$ be sequences of positive numbers.

(i) The sequence of chains exhibits a cutoff at a_n if, given any fixed constant $c > 0$,

$$\begin{aligned} \lim_{n\to\infty} d_n(\lceil ca_n \rceil) &= 0 \quad \text{if } c > 1, \\ \lim_{n\to\infty} d_n(\lfloor ca_n \rfloor) &= 1 \quad \text{if } c < 1. \end{aligned} \tag{6.6}$$

(ii) The sequence of chains exhibits a cutoff at a_n with window w_n if $w_n = o(a_n)$, and

$$\begin{aligned} \lim_{\alpha\to\infty} \limsup_{n\to\infty} d_n(\lceil a_n + \alpha w_n \rceil) = 0, \\ \lim_{\alpha\to\infty} \liminf_{n\to\infty} d_n(\lfloor a_n - \alpha w_n \rfloor) = 1. \end{aligned} \tag{6.7}$$

Proposition 6.10. *Let $(\Omega_n, P_n, \pi_n)_{n\in\mathbb{Z}_+}$ be a sequence of Markov chains. Let $(a_n)_{n\in\mathbb{Z}_+}$ and $(w_n)_{n\in\mathbb{Z}_+}$ be sequences of positive numbers. Assume that $a_n \to \infty$ and $w_n \to \infty$ as $n \to \infty$.*

(i) *If the sequence exhibits a cutoff at a_n as in Definition 6.3, then it exhibits a cutoff as in Definition 6.9.*

(ii) *If the sequence exhibits a cutoff at a_n with window w_n as in Definition 6.6, then it exhibits a cutoff with window w_n as in Definition 6.9. Furthermore, it exhibits a cutoff at a_n.*

In both cases, $t_{\text{mix}}^{(n)} = (1 + o(1))a_n$.

Proof. (i) Let $\delta > 0$ and $0 < \epsilon < 1$. By using a similar argument as for the proof of condition (i) of Proposition 6.4, we deduce that for sufficiently large n (depending on ϵ),

$$t_{\text{mix}}^{(n)}(\epsilon) \le (1+\delta)a_n + 1, \quad \text{and} \quad t_{\text{mix}}^{(n)}(1-\epsilon) \ge (1-\delta)a_n - 1. \tag{6.8}$$

In particular, this implies that for all $0 < \epsilon < 1/2$,

$$1 \le \frac{t_{\text{mix}}^{(n)}(\epsilon)}{t_{\text{mix}}^{(n)}(1-\epsilon)} \le \frac{1+\delta+(1/a_n)}{1-\delta-(1/a_n)}.$$

By assumption, $\lim_{n\to\infty} \frac{1}{a_n} = 0$. Thus, sending $\delta \to 0$ shows that $\lim_{n\to\infty} \frac{t_{\text{mix}}^{(n)}(\epsilon)}{t_{\text{mix}}^{(n)}(1-\epsilon)} = 1$ by sandwiching. Hence, we have cutoff as in (6.2), or equivalently (6.1). In particular, upper and lower bounds on $\frac{t_{\text{mix}}^{(n)}}{a_n}$ can be obtained from (6.8) by taking $\epsilon = 1/4$ and $\epsilon = 3/4$ respectively. This shows that for sufficiently large n,

$$1 - \delta - \frac{1}{a_n} \le \frac{t_{\text{mix}}^{(n)}}{a_n} \le 1 + \delta + \frac{1}{a_n}.$$

Taking the limit as $\delta \to 0$ shows that $\lim_{n\to\infty} \frac{t_{\text{mix}}^{(n)}}{a_n} = 1$, or equivalently $t_{\text{mix}}^{(n)} = (1+o(1))a_n$.

(ii) A similar argument as for the proof of condition (i) for Proposition 6.7 implies that there exist real constants A and B such that, for sufficiently large n,

$$a_n + Bw_n - 1 \le t_{\text{mix}}^{(n)} \le a_n + Aw_n + 1, \tag{6.9}$$

By rearranging, this is equivalent to $B\frac{w_n}{a_n} \le \frac{t_{\text{mix}}^{(n)}}{a_n} - 1 \le A\frac{w_n}{a_n}$. Since $w_n = o(a_n)$, $\lim_{n\to\infty}\frac{w_n}{a_n} = 0$. Therefore, $\lim_{n\to\infty}\frac{t_{\text{mix}}^{(n)}}{a_n} = 1$, or equivalently $t_{\text{mix}}^{(n)} = (1+o(1))a_n$, by sandwiching. Moreover, $\frac{w_n}{t_{\text{mix}}^{(n)}} = \frac{w_n}{(1+o(1))a_n}$, which tends to zero as $n\to\infty$, and so $w_n = o(t_{\text{mix}}^{(n)})$. Next, (6.9) also implies that

$$d(\lceil t_{\text{mix}}^{(n)} + \alpha w_n\rceil) \le d(\lceil a_n + (\alpha + B - 1)w_n\rceil), \qquad d(\lfloor t_{\text{mix}}^{(n)} + \alpha w_n\rfloor) \ge d(\lfloor a_n + (\alpha + A + 1)w_n\rfloor).$$

Taking limits and using (6.7) shows that we have a cutoff with window w_n, as in (6.4).

Finally, we will show that a cutoff at a_n with window w_n implies a cutoff at a_n. We have shown that $t_{\text{mix}}^{(n)} = (1+o(1))a_n$. Suppose that $c > 1$. Then for sufficiently large n, $ct_{\text{mix}}^{(n)} = c(1+o(1))a_n > \widetilde{c}a_n$, where $\widetilde{c} = c\left(1 - \frac{c-1}{2}\right) > 1$. Thus, $d_n(\lceil ct_{\text{mix}}^{(n)}\rceil) \le d_n(\lceil \widetilde{c}a_n\rceil)$. Since the right hand side tends to zero as $n\to\infty$ by (6.6), $\lim_{n\to\infty} d_n(\lceil ct_{\text{mix}}^{(n)}\rceil) = 0$. The other side of the cutoff when $c < 1$ follows analogously. □

Remark 6.11. (i) Proposition 6.10 shows that if a sequence exhibits a cutoff at a_n and at b_n, then the two cutoff points are asymptotically equivalent: $a_n = (1+o(1))b_n$.

(ii) However, (6.5) suggests that two cutoff windows do not necessarily have to be of the same order. Any cutoff window can be made wider as long as it is still $o(t_{\text{mix}}^{(n)})$. Thus, it might interesting to ask what the widest or narrowest a window can be.

(iii) While the assumptions that $t_{\text{mix}}^{(n)}$, a_n, and w_n tend to infinity are not necessary to define cutoff, they were needed to prove the equivalences of the various definitions. This is an artefact of working in discrete time (i.e. because of rounding). Without these assumptions, the most that can be said in general about two cutoff points a_n and b_n is that the limit points of $|a_n - b_n|$ all lie in the interval $[0,1]$ (see [15, p.4]).

We will describe a simple demonstration of cutoff where the mixing time is bounded.

Example 6.12. Let $(\Omega_n, P_n, \pi_n)_{n\in\mathbb{Z}_+}$ denote a random walk (with self-loops) on the complete graph with n vertices, where $P_n(i,j) = \frac{1}{n}$ for $1 \le i, j \le n$. Since P_n is symmetric, by Remark 2.7, this has uniform stationary distribution $\pi_n(i) = \frac{1}{n}$ for $1 \le i \le n$.

Here $d_n(0) = \max_{x\in\Omega_n}\|\delta_x - \pi\|_{\text{TV}} = 1 - \frac{1}{n}$, and $d_n(1) = 0$. Therefore, $t_{\text{mix}}^{(n)} = 1$ for all $n \in \mathbb{Z}_+$, which does not tend to infinity. However, there is still a sharp cutoff since $d_n(\lceil c\rceil) \to 0$ for $c > 1$, and $d_n(\lfloor c\rfloor) \to 1$ for $c < 1$, so condition (i) of Definition 6.3 holds.

Furthermore, there is also a cutoff with window $w_n = \frac{1}{n}$, since $d(\lceil 1 + \frac{\alpha}{n}\rceil) = 0$ for $\alpha > 0$, and $d(\lfloor 1 + \frac{\alpha}{n}\rfloor) = 1$ for $\alpha < 0$, so condition (i) of Definition 6.6 holds. Note that any other window tending to zero will also work, such as $v_n = \frac{1}{\log n}$ (which is wider).

In this case, it is easy to check that condition (ii) of both cutoff definitions also hold.

We will conclude this section by showing that a necessary condition for a reversible chain to have a cutoff is that the product of the mixing time and spectral gap (from Definition 4.2) tends to infinity (the "product condition"). Peres conjectured that this would

be sufficient in many settings (but not all; see [38, Example 18.7] for a counterexample). So far, this conjecture has been confirmed for birth-and-death chains [19, 15].

Theorem 6.13. *Suppose that $(\Omega_n, P_n, \pi_n)_{n\in\mathbb{Z}_+}$ is a sequence of reversible, irreducible, and aperiodic chains with mixing times $(t_{\text{mix}}^{(n)})_{n\in\mathbb{Z}_+}$ and (absolute) spectral gaps $(\gamma_*^{(n)})_{n\in\mathbb{Z}_+}$. Then the chains exhibit a pre-cutoff (and hence cutoff) only if $t_{\text{mix}}^{(n)} \cdot \gamma_*^{(n)} \to \infty$ as $n \to \infty$.*

Proof. We show the contrapositive. Suppose that $t_{\text{mix}}^{(n)} \cdot \gamma_*^{(n)}$ is asymptotically bounded from above by some real $C > 0$. Fix $0 < \epsilon < \frac{1}{2}$. Since $t_{\text{mix}}^{(n)}(\epsilon) \geq \left(\frac{1}{\gamma_*^{(n)}} - 1\right) \log(2\epsilon)^{-1}$ from Theorem 4.6, and $t_{\text{mix}}^{(n)}(1-\epsilon) \leq t_{\text{mix}}^{(n)}$, this implies that for sufficiently large n,

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

As ϵ tends to zero, the right hand side tends to infinity. Therefore, the ratio of the mixing times is unbounded, and there is no pre-cutoff (and hence there is no cutoff). □

Remark 6.14.

(i) It is useful to use Theorem 6.13 to prove that there is no cutoff by showing that the product condition fails to hold (i.e. if $t_{\text{mix}}^{(n)} \cdot \gamma_*^{(n)} = O(1)$).

(ii) For the (lazy) random walk on the n-cycle, Examples 3.4 and 3.5 showed that for some constants c_1, $c_2 > 0$, we have $c_1 n^2 \leq t_{\text{mix}}^{(n)} \leq c_2 n^2$. Moreover, Example 4.7 showed that $\gamma_*^{(n)} = O(n^{-2})$. Hence, $t_{\text{mix}}^{(n)} \cdot \gamma_*^{(n)} = O(1)$, which is bounded. Therefore, there is no cutoff, and we conclude that the transition to stationarity occurs gradually around order n^2 steps.

(iii) For the (lazy) random walk on the n-dimensional hypercube, Example 3.6 showed that $t_{\text{mix}}^{(n)} = O(n \log n)$, and Example 4.8 showed that $\gamma_*^{(n)} = O(n^{-1})$. Hence, their product tends to infinity. Indeed, it can be shown that this random walk exhibits a cutoff, using a slightly more sophisticated coupling (see [38, Theorem 18.3]).

Chapter 7

The Mean-Field Ising Model

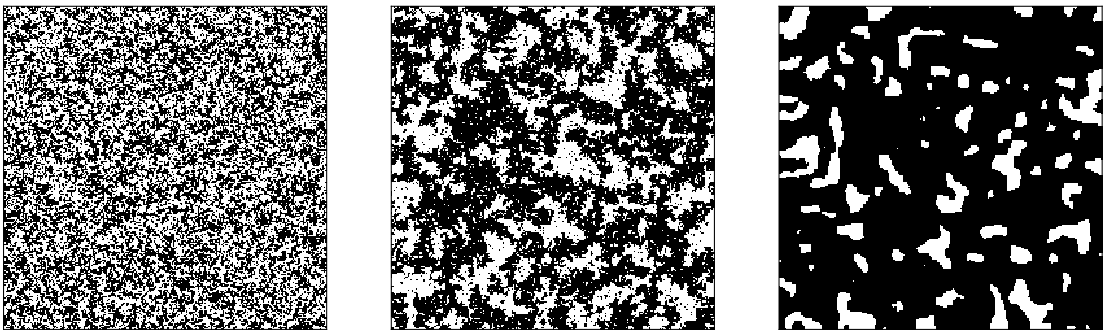

Figure 7.0.1: Ising model on the 250×250 grid at high, critical, and low temperatures.

In this chapter, we will analyse the Ising model, which is a widely-studied model of magnetism and phase transitions from statistical physics. Figure 7.0.1 displays illustrative samples at three distinct temperatures, which I generated by Glauber dynamics using Python. The normalisation constant of the Ising model, known as the partition function, can be used to compute almost all physical properties related to the system [32]. However, its intractability means that the Ising model is, in general, not exactly solvable.

Hence, MCMC techniques (see Chapter 5) have been found to be extremely useful. We will focus on analysing the mixing time of the Glauber dynamics, which is one of the most popular and widely used methods for sampling the Ising model. Apart from providing practical guarantees of efficiency, this topic presents incredibly interesting connections to fields such as probability theory, statistical physics, and complexity theory.

Section 7.1 will describe the background of the Glauber dynamics. The remaining sections will focus on the mean-field Ising model, which enjoys a particularly high degree of symmetry since the underlying geometry is effectively removed. The aim will be to describe its distinct behaviour, closely following Levin, Luczak, and Peres [37].

At high temperatures, the Glauber dynamics for the mean-field Ising model is rapidly mixing and exhibits a cutoff (Section 7.2). At the critical temperature, the chain is rapidly mixing and does not exhibit a cutoff (Section 7.3). At low temperatures, the chain mixes exponentially slowly and does not exhibit a cutoff (Section 7.4).

Other related work includes a more precise analysis around the critical window by Ding, Lubetzky and Peres [18]. More recently, Lubetzky and Sly have resolved fundamental open problems for the Glauber dynamics for the Ising model on the lattice (where

the underlying geometry is non-trivial) by establishing a cutoff in the high temperature regime [42], and determining the critical mixing rate [41]. More recently, they have also shown that cutoff occurs on any geometry at high enough temperatures [43]. This supports a conjecture of Peres [38, Section 23.2] on the universality of the cutoff phenomenon.

7.1 Glauber dynamics for the Ising model

The Ising model on an underlying graph $G = (V, E)$ is defined as follows. Its set of possible **configurations** is $\Omega = \{\pm 1\}^V$, where each state is an assignment of a spin of $+1$ or -1 to each vertex in V. The probability that the system is in a configuration $\sigma \in \Omega$ is given by the **Gibbs distribution**

$$\mu(\sigma) := \frac{1}{Z(\beta)} \exp\left(\beta J \sum_{\{v,w\}\in E} \sigma(v)\sigma(w) + h \sum_{v\in V} \sigma(v)\right). \tag{7.1}$$

The parameter $\beta > 0$ represents inverse temperature, h is the external field, and J is the interaction strength. The **partition function** $Z(\beta)$ is the normalising constant. We will consider the ferromagnetic (i.e. attractive) case $J = 1$, and with no external field $h = 0$.

We will consider the (single-site, heat-bath) Glauber dynamics for the Ising model. At each step, a vertex $v \in V$ is chosen uniformly at random, and its spin is updated according to the Gibbs distribution μ, conditioned on the spins of all the other vertices.

Let $\Omega(\sigma, v) = \{\tau \in \Omega : \sigma(w) = \tau(w) \ \ \forall w \neq v\}$ be the set of configurations agreeing with σ everywhere except at v, $\mathcal{N}(v) = \{w \in V : \{v, w\} \in E\}$ be the set of neighbours of v, and $S(\sigma, v) = \sum_{w\in\mathcal{N}(v)} \sigma(w)$. Given that v is selected, the probability that its spin is updated to $+1$ and -1 is given by $p_+(S(\sigma, v))$ and $p_-(S(\sigma, v))$ respectively, where

$$p_+(s) := \frac{e^{\beta s}}{e^{\beta s} + e^{-\beta s}} = \frac{1 + \tanh(\beta s)}{2}, \qquad p_-(s) := \frac{e^{-\beta s}}{e^{\beta s} + e^{-\beta s}} = \frac{1 + \tanh(-\beta s)}{2}. \tag{7.2}$$

This follows from a routine calculation using Bayes' formula. Hence, the transition matrix of the Glauber dynamics can be explicitly written as

$$P(\sigma, \tau) = \frac{1}{n} \sum_{v\in V} \frac{1 + \tanh(\beta\, \tau(v)\, S(\sigma, v))}{2} \cdot \mathbb{1}_{\{\sigma(w)=\tau(w)\ \forall w\neq v\}}. \tag{7.3}$$

(By [22], this heat-bath chain has non-negative eigenvalues.) The chain is clearly irreducible and aperiodic. Moreover, it can also be easily verified that the detailed balance equations $\mu(\sigma)P(\sigma, \tau) = \mu(\tau)P(\tau, \sigma)$ hold for all $\sigma, \tau \in \Omega$. Hence, the chain is reversible with respect to the Gibbs distribution μ (7.1), which is therefore the unique stationary distribution.

Remark 7.1. Much more can be said about Gibbs measures and the Glauber dynamics on spin systems on (countably infinite) lattices that is beyond the scope of this thesis. For example, refer to the survey of Martinelli [45].

Theorem 7.2. *Suppose that G has n vertices and maximum degree Δ. If $\Delta \tanh(\beta) < 1$, then the Glauber dynamics for the Ising model on G is rapidly mixing, with*

$$t_{\text{mix}}(\epsilon) \leq \left\lceil \frac{n(\log n + \log(\epsilon^{-1}))}{1 - \Delta \tanh(\beta)} \right\rceil. \tag{7.4}$$

In particular, since $\tanh(x) \leq x$ for $x > 0$, this holds whenever $\beta < \frac{1}{\Delta}$.

Proof. We will use the path coupling method from Section 3.2. By identifying each spin with a "colour", the procedure is similar to the one for sampling colourings (Lemma 5.4 in Chapter 5). Recall that the path metric $\text{dist}(\sigma, \tau) = \sum_{v \in V} \mathbb{1}_{\{\sigma(v) = \tau(v)\}}$ counts the vertices with different spins. We want to construct a coupling (X, Y) of the chain started at two configurations σ and τ that only differ at one vertex, say v, that is contracting.

Choose a vertex w uniformly at random. Let U be an independent Uniform$(0, 1)$ random variable, which we will use as a common source of randomness to update the spin at w by the following:

$$X(w) = \begin{cases} +1 & \text{if } 0 \leq U \leq p_+(S(\sigma, w)), \\ -1 & \text{if } p_+(S(\sigma, w)) < U \leq 1; \end{cases} \qquad Y(w) = \begin{cases} +1 & \text{if } 0 \leq U \leq p_+(S(\tau, w)), \\ -1 & \text{if } p_+(S(\tau, w)) < U \leq 1. \end{cases}$$

Thus, we can always update both chains with the same spin, except when w is adjacent to v. In this case, assume that $\sigma(v) = -1$ and $\tau(v) = +1$ without loss of generality, so that $0 \leq p_+(S(\sigma, w)) \leq p_+(S(\tau, w)) \leq 1$. Hence, the chains will have different spins with probability $\big[p_+(S(\tau, w)) - p_-(S(\sigma, w))\big] = \frac{1}{2}\big[\tanh(\beta(S(\sigma, w) + 2)) + \tanh(\beta S(\sigma, w))\big]$. Using calculus, this can be shown to be upper bounded by $\tanh(\beta)$. Since $\text{dist}(X, Y)$ decreases by one only if $w = v$, and can increase by one only if $w \in \mathcal{N}(v)$,

$$\mathbb{E}_{\sigma,\tau}\left[\text{dist}(X, Y)\right] \leq 1 - \frac{1}{n} + \frac{1}{n} \sum_{w \in \mathcal{N}(v)} \big[p_+(S(\tau, w)) - p_+(S(\sigma, w))\big] \leq 1 - \frac{1 - \Delta \tanh(\beta)}{n}. \tag{7.5}$$

Note that $1 - \frac{1 - \Delta \tanh(\beta)}{n} \leq \exp\left(-\frac{1 - \Delta \tanh(\beta)}{n}\right)$. If $\alpha = \frac{1 - \Delta \tanh(\beta)}{n} > 0$, then the chain is contracting for adjacent configurations (i.e. satisfies (3.15)). Therefore, Corollary 3.13 of the path coupling theorem implies (7.4), as desired. □

Remark 7.3. Establishing the contraction condition $\mathbb{E}_{\sigma,\tau}\left[\text{dist}(X, Y)\right] \leq e^{-\alpha}$ for some $\alpha > 0$ is the key part of the proof of Theorem 7.2. This condition is closely related to the Dobrushin-Shlosman condition from statistical physics (which implies the uniqueness of the Gibbs measure on a countably infinite state space). See [28, 8] for more details on the connection between path coupling and the Dobrushin-Shlosman condition.

7.2 Rapid mixing and cutoff at high temperatures

For the rest of this chapter, we will focus on the mean-field Ising model on the complete graph on $n \geq 2$ vertices K_n, also known as the Curie-Weiss model (see Figure 7.2.1). Let the vertex set be $V = \{1, 2, \ldots, n\}$.

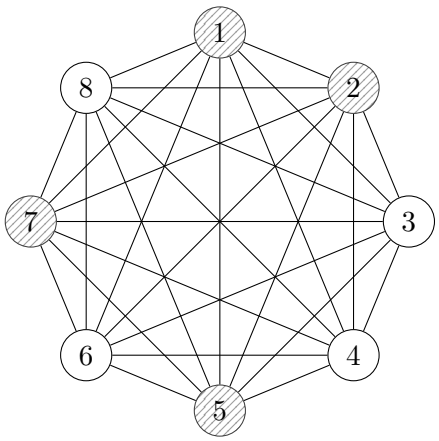

Figure 7.2.1: Mean-field Ising model on the complete graph K_n (here $n = 8$). Each vertex is associated a spin in $\{\pm 1\}$.

For convenience, we will rescale the inverse temperature parameter β, so that the Gibbs distribution μ_n takes the following form

$$\mu_n(\sigma) = \frac{1}{Z(\beta)} \exp\left(\frac{\beta}{n} \sum_{1 \leq i < j \leq n} \sigma(i)\sigma(j)\right). \tag{7.6}$$

In other words, β now corresponds to an inverse temperature parameter of $\widetilde{\beta} = \frac{\beta}{n}$ in the original definition (7.1). Since $\Delta = n - 1$ for K_n, Theorem 7.2 immediately implies that the Glauber dynamics on K_n is rapidly mixing, with $t_{\text{mix}}^{(n)} = O(n \log n)$, whenever $\beta < \frac{n}{n-1}$. In particular, this holds if $\beta < 1$. The main result of this section will be show that there is actually a cutoff in this high temperature regime.

Theorem 7.4. *If $\beta < 1$, then the Glauber dynamics for the Ising model on K_n has a cutoff at $\frac{n \log n}{2(1-\beta)}$ with window n. That is, from Definition 6.9,*

$$\begin{aligned} \lim_{\alpha \to \infty} \limsup_{n \to \infty} d_n\left(\left\lceil \frac{n \log n}{2(1-\beta)} + \alpha n \right\rceil\right) &= 0, \\ \lim_{\alpha \to \infty} \liminf_{n \to \infty} d_n\left(\left\lfloor \frac{n \log n}{2(1-\beta)} - \alpha n \right\rfloor\right) &= 1. \end{aligned} \tag{7.7}$$

This was first proved by Levin et al. [37], and our aim will be to provide an exposition of their key ideas. The related paper of Ding et al. [18], and the talk of Levin [36] offer additional insights. The main technique used is coupling, which is described in Section 3.1.

Preliminaries for the proof of Theorem 7.4

Define the **normalised magnetisation** (or mean magnetisation) by

$$S(\sigma) \coloneqq \frac{1}{n}\sum_{i=1}^{n}\sigma(i). \tag{7.8}$$

Let $(X_t)_{t\in\mathbb{N}}$ be the Glauber dynamics on K_n with stationary distribution μ_n (7.6). Given that a vertex i has been selected, its spin is updated from $\sigma(i)$ to $+1$ or -1 with probabilities $p_+(S(\sigma) - \frac{\sigma(i)}{n})$ and $p_-(S(\sigma) - \frac{\sigma(i)}{n})$ respectively, by (7.2). Hence, the **magnetisation chain** $(S_t)_{t\in\mathbb{N}}$, where $S_t \coloneqq S(X_t)$, is also a Markov chain on the states $\Omega_M \coloneqq \{-1, -1+\frac{2}{n}, \dots, 1-\frac{2}{n}, 1\}$, with stationary distribution π_n, and transition probabilities

$$P_M(s,s') = \begin{cases} \frac{1-s}{2}p_+(s+\frac{1}{n}) & \text{if } s' = s+\frac{2}{n}, \\ \frac{1+s}{2}p_-(s-\frac{1}{n}) & \text{if } s' = s-\frac{2}{n}, \\ 1-\frac{1+s}{2}p_-(s-\frac{1}{n}) - \frac{1-s}{2}p_+(s-\frac{1}{n}) & \text{if } s' = s. \end{cases} \tag{7.9}$$

(By symmetry, the distributions of $(S_t)_{t\in\mathbb{N}}$ and $(-S_t)_{t\in\mathbb{N}}$ are the same.) This is a **projection** of the Glauber dynamics X_t (see [38, Section 2.3.1]) onto a birth-and-death chain, where we identify configurations with the same magnetisation. The associated probabilities are given by the pushforward measure (e.g. $\pi_n = \mu_n S^{-1}$). We will find that the mixing time of $(X_t)_{t\in\mathbb{N}}$ is mainly determined by the mixing time of $(S_t)_{t\in\mathbb{N}}$.

Using the mean value theorem for $\tanh(x)$, which has positive derivative $1/\cosh^2(x)$, shows that the transition probabilities of S_t (7.9) take on the following simplified forms:

$$P_M(s,s') = \begin{cases} \frac{1-s}{4}(1+\tanh(\beta s)) + O(n^{-1}) & \text{if } s' = s+\frac{2}{n} \\ \frac{1+s}{4}(1-\tanh(\beta s)) + O(n^{-1}) & \text{if } s' = s-\frac{2}{n} \\ \frac{1}{2}(1+s\tanh(\beta s)) - O(n^{-1}) & \text{if } s' = s. \end{cases} \tag{7.10}$$

Thus, the holding probability $P_M(s,s)$ is bounded from above by $\frac{1}{2}(1+\tanh(\beta)) < 1$, uniformly in s and n. It can also be deduced that $P_M(s,s)$ is uniformly bounded away from 0. Furthermore, from [37, (2.13)], an expression for the drift of $(S_t)_{t\in\mathbb{N}}$ is given by $\mathbb{E}[S_{t+1} \mid S_t = s] = (1-n^{-1})s + f_n(s) - \theta_n(s)$, where

$$\begin{aligned} f_n(s) &= \frac{1}{2n}[\tanh(\beta(s+n^{-1})) + \tanh(\beta(s-n^{-1}))], \\ \theta_n(s) &= \frac{s}{2n}[\tanh(\beta(s+n^{-1})) - \tanh(\beta(s-n^{-1}))]. \end{aligned} \tag{7.11}$$

The behaviour of the hyperbolic tangent function will be crucial for precisely tracking the moments of the magnetisation chain in order to establish cutoff. Using $\tanh(x) \le x$ for $x \ge 0$, and the symmetry of $(S_t)_{t\in\mathbb{N}}$ and $(-S_t)_{t\in\mathbb{N}}$, with (7.11) shows that

$$\mathbb{E}[\,|S_{t+1}| - |S_t| \mid S_t] \le \frac{|S_t|(\beta-1)}{n}, \quad |S_t| > 1/n. \tag{7.12}$$

(See [37, (2.15)–(2.18)]. Without the absolute values, (7.12) holds for all $S_t \geq 0$.) Therefore, if $\beta \leq 1$, the chain $(|S_t|)$ has non-positive drift whenever $|S_t| > 1/n$.

Next, we will define a **monotone coupling** $(X_t, \widetilde{X}_t)_{t\in\mathbb{N}}$ of two copies of the Glauber dynamics $(X_t)_{t\in\mathbb{N}}$ and $(\widetilde{X}_t)_{t\in\mathbb{N}}$, started at σ_0 and $\widetilde{\sigma}_0$ respectively. This is essentially the coupling used in the proof of Theorem 7.2, where a common source of randomness is used to update the same vertex for both chains with the correct marginal probabilities. Let $(S_t)_{t\in\mathbb{N}}$ and $(\widetilde{S}_t)_{t\in\mathbb{N}}$ denote their magnetisation chains, started at s and $\widetilde{s}$.

Consider the partial order on Ω where $\sigma \leq \widetilde{\sigma}$ if and only if $\sigma(i) \leq \sigma(i)$ for all i. If $X_t \leq \widetilde{X}_t$, then $X_u \leq \widetilde{X}_u$ for all $u \geq t$ (hence why this is called a monotone coupling). This is because the function p_+ (7.2) used to update the spins is monotone increasing.

We will collect some technical lemmas about the magnetisation chain that will be referred to later. (We will omit proofs, and these can be skipped for the main results.)

Lemma 7.5 ([37])**.** *Let* $\rho \coloneqq 1 - \frac{1-n\tanh(\beta/n)}{n}$.

(i) *(Lemma 2.2). For the monotone coupling,* $\mathbb{E}_{\sigma,\widetilde{\sigma}}\left[|S_t - \widetilde{S}_t|\right] \leq 2\rho^t$.

(ii) *(Lemma 2.3).* $0 \leq \mathbb{E}_s[S_1] - \mathbb{E}_{\widetilde{s}}[S_1] \leq \rho(s - \widetilde{s})$ *for any* $s, \widetilde{s} \in \Omega_M$ *with* $s \geq \widetilde{s}$.

(iii) *(Prop. 2.7). If* $\beta < 1$, *then* $\mathrm{Var}_s(S_t) = O(n^{-1})$. *If* $\beta = 1$, *then* $\mathrm{Var}_s(S_t) = O(tn^{-2})$.

(iv) *(Lemma 2.8). For any subset* A *of vertices, let* $M_t(A) \coloneqq \frac{1}{2}\sum_{i\in A} X_t(i)$. *If* $\beta < 1$, *then for all* A *and* $\sigma \in \Omega$,

  (a) $|\mathbb{E}_\sigma[S_t]| \leq 2e^{-(1-\beta)t/n}$.

  (b) $|\mathbb{E}_\sigma[M_t(A)]| \leq |A|e^{-(1-\beta)t/n}$ *and* $\mathrm{Var}(M_t(A)) = O(n)$.

  (c) $\mathbb{E}_\sigma[|M_t(A)|] \leq ne^{(1-\beta)t/n} + O(\sqrt{n})$.

Finally, the next result, used in [18], allows for stopping times for processes with non-positive drift (such as the magnetisation chain, if $S_t \geq 0$, from (7.12)) to be bounded.

Lemma 7.6 ([38, Proposition 17.20])**.** *A stochastic process* $(W_t)_{t\in\mathbb{N}}$ *is a supermartingale if* $\mathbb{E}[|W_t|] < \infty$ *and* $\mathbb{E}[W_{t+1} \mid W_0, \ldots, W_n] \leq W_n$. *Let* $(W_t)_{t\in\mathbb{N}}$ *be a non-negative supermartingale and* τ *be a stopping time such that*

(i) $W_0 = k$;

(ii) $|W_{t+1} - W_t| \leq B$ *for some constant* $B > 0$; *and*

(iii) $\mathrm{Var}(W_{t+1} \mid W_0, \ldots, W_t) > \sigma^2$ *for some constant* $\sigma^2 > 0$ *on the event* $\{\tau > t\}$.

If $u > \frac{4B^2}{3\sigma^2}$, *then* $\mathbb{P}_k(\tau > t) \leq \frac{4k}{\sigma\sqrt{t}}$.

For example, let $(W_t)_{t\in\mathbb{N}}$ be an unbiased simple random walk on $\mathbb{Z}$ with holding probability $b < 1$, starting at $k > 0$. This satisfies $|W_{t+1} - W_t| \leq 1$, and $\mathrm{Var}(X_t) = 1 - b > 0$. If τ is the first time that W_t hits zero, then Lemma 7.6 implies that $\mathbb{P}_k(\tau > t) = O(t^{-1/2})$. (In this case, the same bound can be obtained using less technical machinery, by modifying the argument of [38, Corollary 2.28] to account for the holding probability.)

In fact, the original argument of [37] relies on the comparison to such an unbiased simple random walk. However, this lemma makes explicit the dependence on a variance that is uniformly bounded away from zero.

Proof of Theorem 7.4 – upper bound of cutoff

The proof proceeds in three stages. In the first stage, we will show that the Glauber dynamics reaches a "nice" starting configuration after a short burn-in period of order n^{-1}. In the second stage, we will show that two copies of the Glauber dynamics can be coupled, such that their magnetisations will match with high probability in $\frac{n \log n}{2(1-\beta)} + O(n^{-1})$ steps. In the final stage, this will boosted into full mixing of the Glauber dynamics after an additional order n^{-1} steps, by coupling their associated two-coordinate chains.

Stage 1: Burn-in phase.

The following lemma allows us to consider the distance to stationarity starting from a "nice" set. The proof relies on the triangle inequality and the Markov property; however, it will be omitted for brevity.

Lemma 7.7 ([37, Lemma 3.3])**.** *For any subset $\Omega_0 \subseteq \Omega$,*

$$\begin{aligned} d_n(t_0 + t) &= \max_{\sigma \in \Omega} \|\mathbb{P}\left(X_{t_0+t} \in \cdot\right) - \mu_n\|_{\mathrm{TV}} \\ &\leq \max_{\sigma_0 \in \Omega_0} \|\mathbb{P}\left(X_t \in \cdot\right) - \mu_n\|_{\mathrm{TV}} + \max_{\sigma \in \Omega} \mathbb{P}_\sigma\left(X_{t_0} \notin \Omega_0\right). \end{aligned} \tag{7.13}$$

To prepare for the two-coordinate chain argument in Stage 3, we will want to start from a configuration in the following set, whose magnetisation is not too extreme:

$$\Omega_0 \coloneqq \left\{\sigma \in \Omega : |S(\sigma)| \leq \frac{1}{2}\right\}. \tag{7.14}$$

Observe that $\mathbb{P}_\sigma\left(X_{\theta_0 n} \notin \Omega_0\right) = \mathbb{P}_\sigma\left(|S_{\theta_0 n}| > 1/2\right)$. By Lemma 7.5 (iv) (a), there exists a constant $\theta_0 > 0$ such that $|\mathbb{E}_\sigma\left[S_{\theta_0 n}\right]| \leq \frac{1}{4}$. Therefore, by Chebyshev's inequality (Theorem 1.2), and Lemma 7.5 (iii),

$$\begin{aligned} \max_{\sigma \in \Omega} \mathbb{P}_\sigma\left(X_{\theta_0 n} \notin \Omega_0\right) &= \mathbb{P}_\sigma\left(|S_{\theta_0 n}| > 1/2\right) \\ &\leq \mathbb{P}_\sigma\left(\left|S_{\theta_0 n} - \mathbb{E}_\sigma\left[S_{\theta_0 n}\right]\right| > 1/4\right) \leq 16 \operatorname{Var}_\sigma(S_{\theta_0 n}) = O(n^{-1}). \end{aligned} \tag{7.15}$$

Stage 2: Magnetisation phase.

Let $t_n \coloneqq \frac{n \log n}{2(1-\beta)}$. The following shows that two Glauber dynamics can be coupled such that their magnetisations will agree with high probability in $t_n + O(n^{-1})$ steps.

Lemma 7.8. *Let $\tau_{\mathrm{mag}} \coloneqq \min\{t \in \mathbb{N} : S_t = \widetilde{S}_t\}$ be the coalescence time of two magnetisation chains. For any starting configurations σ and $\widetilde{\sigma}$, there exists a coupling $(X_t, \widetilde{X}_t)_{t \in \mathbb{N}}$ such that for some constant $c > 0$, independent of σ, $\widetilde{\sigma}$, and n,*

$$\mathbb{P}_{\sigma,\widetilde{\sigma}}\left(\tau_{\mathrm{mag}} > t_n + \alpha n\right) \leq \frac{c}{\sqrt{\alpha}}. \tag{7.16}$$

Proof. Let $(X_t, \widetilde{X}_t)_{t \in \mathbb{N}}$ be the monotone coupling of the Glauber dynamics started at σ and $\widetilde{\sigma}$ respectively, and $(S_t)_{t \in \mathbb{N}}$ and $(\widetilde{S}_t)_{t \in \mathbb{N}}$ be their corresponding magnetisation chains.

By Lemma 7.5 (i), there exists a constant $c_1 > 0$ such that

$$\mathbb{E}_{\sigma,\widetilde{\sigma}}\left[(n/2)\,|S_{t_n} - \widetilde{S}_{t_n}|\right] \le c_1\sqrt{n}. \tag{7.17}$$

(Thus, the dominant term in the cutoff t_n will be from driving the expected difference in magnetisation to order $\sqrt{n}$.) Suppose that the two magnetisation chains have not coalesced by t_n. To prevent S_t and $\widetilde{S}_t$ from "jumping over" each other, we will analyse up to $\tau_1 := \min\{t \ge t_n : (n/2)\,|S_t - \widetilde{S}_t| \le 1\}$, the first time after t_n that they are adjacent.

For $t_n \le t < \tau_1$, run the chains (X_t) and $(\widetilde{X}_t)$ independently. If we assume $S_{t_n} \ge \widetilde{S}_{t_n}$ without loss of generality, then $S_t \ge \widetilde{S}_t$ for $t \le \tau_1$. Hence, for $t_n \le t < \tau_1$, the process $(S_t - \widetilde{S}_t)$ is non-negative, and also has non-positive drift (i.e. it is a supermartingale), by Lemma 7.5 (ii). Moreover, from (7.10), the increments of $(S_t - \widetilde{S}_t)$ are non-zero with probability uniformly bounded away from zero (i.e. its variance is uniformly bounded from zero). Therefore, we can use Lemma 7.6 to deduce that for some constant $c_2 > 0$,

$$\mathbb{P}_{\sigma,\widetilde{\sigma}}\left(\tau_1 > t_n + \alpha n \mid X_{t_n}, \widetilde{X}_{t_n}\right) \le \frac{c_2 n|S_{t_n} - \widetilde{S}_{t_n}|}{\sqrt{\alpha n}}.$$

Taking expectation and using (7.17) implies that

$$\mathbb{P}_{\sigma,\widetilde{\sigma}}\left(\tau_1 > t_n + \alpha n\right) = O(\alpha^{-1/2}). \tag{7.18}$$

If $\tau_{\text{mag}} = \tau_1$, then we are done. Otherwise, X_{τ_1} has one more plus spin than $\widetilde{X}_{\tau_1}$. Then every vertex of X_{τ_1}, except for one, can be paired with a vertex of $\widetilde{X}_{\tau_1}$ with the same spin. The monotone coupling can then be used to update the paired vertices together. Since this corresponds to the case where the two starting configurations differ at only one vertex, the contraction property (7.5) holds (with a rescaled β), and so after iterating,

$$\begin{aligned}\mathbb{P}_{\sigma,\widetilde{\sigma}}\left(\tau_{\text{mag}} > \tau_1 + \alpha n \mid X_{\tau_1}, \widetilde{X}_{\tau_1}\right) &= \mathbb{P}_{\sigma,\widetilde{\sigma}}\left(\text{dist}(X_{\tau_1+\alpha n}, \widetilde{X}_{\tau_1+\alpha n}) \ge 1 \mid X_{\tau_1}, \widetilde{X}_{\tau_1}\right)\\ &\le \mathbb{E}_{\sigma,\widetilde{\sigma}}\left[\text{dist}(X_{\tau_1+\alpha n}, \widetilde{X}_{\tau_1+\alpha n}) \mid X_{\tau_1}, \widetilde{X}_{\tau_1}\right]\\ &\le \left(1 - \frac{(n-1)\tanh(\beta/n)}{n}\right)^{\alpha n} \le e^{-(1-\beta)\alpha}.\end{aligned} \tag{7.19}$$

Here, Markov's inequality and $\tanh(x) \le x$ for $x \ge 0$ are used in the first and last inequalities. Note that $e^{-(1-\beta)\alpha} = O(\alpha^{-1/2})$. To summarise,

$$\begin{aligned}\mathbb{P}_{\sigma,\widetilde{\sigma}}\left(\tau_{\text{mag}} > t_n + \alpha n\right) \le \mathbb{P}_{\sigma,\widetilde{\sigma}}\left(\tau_1 > t_n + (\alpha/2)n\right)\\ + \mathbb{P}_{\sigma,\widetilde{\sigma}}\left(\tau_{\text{mag}} > \tau_1 + (\alpha/2)n \mid \tau_1 \le t_n + (\alpha/2)n\right).\end{aligned}$$

Thus, (7.18) and (7.19) imply that $\mathbb{P}_{\sigma,\widetilde{\sigma}}\left(\tau_{\text{mag}} > t_n + \alpha n\right) = O(\alpha^{-1/2})$. □

Stage 3: Two-coordinate chain phase.

In this final stage, we will boost mixing of the magnetisation chain to the full mixing of the Glauber dynamics. After the burn-in stage, we can assume that the starting configuration comes from the "nice" set $\Omega_0 = \{\sigma \in \Omega : |S(\sigma)| \leq \frac{1}{2}\}$. For the rest of this stage, fix $\sigma_0 \in \Omega_0$, and denote its number of positive and negative spins by, respectively,

$$\overline{u}_0 \coloneqq |\{i : \sigma_0(i) = +1\}|, \qquad \overline{v}_0 \coloneqq |\{i : \sigma_0(i) = -1\}|. \tag{7.20}$$

For any configuration σ, we also want to consider its positive and negative spins separately. However, we will do so relative to the fixed σ_0. First, partition the vertices as follows:

$$\begin{aligned} A(\sigma) &\coloneqq \{i : \sigma(i) = +1,\, \sigma_0(i) = +1\}, \quad B(\sigma) \coloneqq \{i : \sigma(i) = -1,\, \sigma_0(i) = +1\} \\ C(\sigma) &\coloneqq \{i : \sigma(i) = +1,\, \sigma_0(i) = -1\}, \quad D(\sigma) \coloneqq \{i : \sigma(i) = -1,\, \sigma_0(i) = -1\}. \end{aligned} \tag{7.21}$$

Then define the following functions:

$$U(\sigma) \coloneqq |A(\sigma)|, \qquad V(\sigma) \coloneqq |D(\sigma)|. \tag{7.22}$$

Let $(X_t)_{t\in\mathbb{N}}$ be a copy of the Glauber dynamics. Then we can consider the process $(U_t, V_t)_{t\in\mathbb{N}}$, where $U_t \coloneqq U(X_t)$ and $V_t \coloneqq V(X_t)$. We call this the **two-coordinate chain** (see Figure 7.2.2), which is also a projection of the Glauber dynamics. This is a Markov chain on $\Lambda = \{0, 1, \ldots, \overline{u}_0\} \times \{0, 1, \ldots, \overline{v}_0\}$, which has stationary distribution π_2, and transition probabilities that depend on the choice of σ_0.

	$\overline{u}_0$						$\overline{v}_0$					
i	1	2	…			…	…				…	n
σ_0	+	+	+	+	+	+	−	−	−	−	−	−
X_t	+	+	+	+	−	−	+	+	−	−	−	−
	$A(X_t)$				$B(X_t)$		$C(X_t)$		$D(X_t)$			
$\widetilde{X}_t$	−	−	−	+	+	+	−	−	−	+	+	+
	$B(\widetilde{X}_t)$			$A(\widetilde{X}_t)$			$D(\widetilde{X}_t)$			$C(\widetilde{X}_t)$		

Figure 7.2.2: Given the Glauber dynamics $(X_t)_{t\in\mathbb{N}}$, the vertices are partitioned into four sets as defined in (7.21), with $|A(X_t)| = U_t$, $|B(X_t)| = \overline{u}_0 - U_t$, $|C(X_t)| = \overline{v}_0 - V_t$, and $|D(X_t)| = V_t$. This schematic represents the two-coordinate chains (U_t, V_t) and $(\widetilde{U}_t, \widetilde{V}_t)$ associated with the coupling $(X_t, \widetilde{X}_t)_{t\in\mathbb{N}}$ of the Glauber dynamics.

Observe that (U_t, V_t) determines the magnetisation by $S_t = \frac{2(U_t - V_t)}{n} - \frac{\overline{u}_0 - \overline{v}_0}{n}$. Define

$$\Lambda_0 \coloneqq \left\{(u, v) \in \Lambda : \frac{n}{4} \leq u, v \leq \frac{3n}{4}\right\}. \tag{7.23}$$

This connects the set of "nice" configurations to the two-coordinate chain. By using the identities $|S(\sigma_0)| = |\overline{u}_0 - \overline{v}_0|$ and $\overline{u}_0 + \overline{v}_0 = n$, we have $\sigma_0 \in \Omega_0$ if and only if $(\overline{u}_0, \overline{v}_0) \in \Lambda_0$.

By the following lemma, it will be sufficient to upper bound the distance to stationarity of the two-coordinate chain of the Glauber dynamics.

Lemma 7.9 ([37, Lemma 3.4])**.** *Let $(X_t)_{t\in\mathbb{N}}$ be the Glauber dynamics started from σ_0, and $(U_t, V_t)_{t\in\mathbb{N}}$ its corresponding two-coordinate chain started from $(\overline{u}_0, \overline{v}_0)$. Then*

$$\|P_{\sigma_0}(X_t \in \cdot) - \mu_n\|_{\mathrm{TV}} = \left\|P_{(\overline{u}_0,\overline{v}_0)}((U_t, V_t) \in \cdot) - \pi_2\right\|_{\mathrm{TV}}. \tag{7.24}$$

Proof. Let $\Omega(u, v) \coloneqq \{\sigma \in \Omega : (U(\sigma), V(\sigma)) = (u, v)\}$. Suppose that we condition on $\Omega(u, v)$. Since specifying $U(\sigma)$ and $V(\sigma)$ fixes the number of positive and negative spins, each arrangement under the Gibbs distribution μ_n is equally likely by symmetry. This is also true for the Glauber dynamics, since any path of t steps to get to a configuration in $\Omega(u, v)$ can be permuted. Hence, both the conditional distributions $\mathbb{P}_{\sigma_0}(X_t \in \cdot \mid (U_t, V_t) = (u, v))$ and $\mu_n(\cdot \mid \Omega(u, v))$ are uniform over $\Omega(u, v)$. Therefore,

$$\left|\mathbb{P}_{\sigma_0}(X_t = \tau) - \mu_n(\tau)\right| = \left|\sum_{(u,v)\in\Lambda} \frac{\mathbb{1}_{\{\tau\in\Omega(u,v)\}}}{|\Omega(u, v)|}\left[\mathbb{P}_{\sigma_0}((U_t, V_t) = (u, v)) - \mu_n(\Omega(u, v))\right]\right| \tag{7.25}$$

By using the triangle inequality, summing over all $\tau \in \Omega$ (recall the definition of the total variation distance (2.27)), and then interchanging the order of summation, we have

$$\|\mathbb{P}_{\sigma_0}(X_t \in \cdot) - \mu_n\|_{\mathrm{TV}} \le \left\|\mathbb{P}_{(\overline{u}_0,\overline{v}_0)}((U_t, V_t) \in \cdot) - \pi_2\right\|_{\mathrm{TV}}.$$

The reverse inequality follows since (U_t, V_t) is a projection of X_t (see [38, Lemma 7.9]). □

Recall that $d(t) \le \overline{d}(t)$ from Lemma 2.23. By taking the maximum over a smaller set in the last step of the proof of the inequality, we have the analogous bound

$$\begin{aligned}\max_{(\overline{u}_0,\overline{v}_0)\in\Lambda_0} &\left\|P_{(\overline{u}_0,\overline{v}_0)}((U_t, V_t) \in \cdot) - \pi_2\right\|_{\mathrm{TV}} \\ &\le \max_{\substack{(\overline{u}_0,\overline{v}_0)\in\Lambda_0\\(\widetilde{u},\widetilde{v})\in\Lambda}} \left\|P_{(\overline{u}_0,\overline{v}_0)}((U_t, V_t) \in \cdot) - P_{(\widetilde{u}_0,\widetilde{v}_0)}((\widetilde{U}_t, \widetilde{V}_t) \in \cdot)\right\|_{\mathrm{TV}}.\end{aligned} \tag{7.26}$$

Given a coupling of two-coordinate chains, let $\tau_2 \coloneqq \min\{t \in \mathbb{N} : (U_t, V_t) = (\widetilde{U}_t, \widetilde{V}_t)\}$ be the first time that they coalesce. By the coupling theorem (Theorem 3.3),

$$\left\|P_{(\overline{u}_0,\overline{v}_0)}((U_t, V_t) \in \cdot) - P_{(\widetilde{u}_0,\widetilde{v}_0)}((\widetilde{U}_t, \widetilde{V}_t) \in \cdot)\right\|_{\mathrm{TV}} \le \mathbb{P}_{(\overline{u}_0,\overline{v}_0),(\widetilde{u},\widetilde{v})}(\tau_2 > t). \tag{7.27}$$

Therefore, we want to construct a "good" coupling of a pair of two-coordinate chains (U_t, V_t) and $(\widetilde{U}_t, \widetilde{V}_t)$, with one starting at a "nice" configuration $\sigma_0 \in \Omega_0$ and the other at an arbitrary $\widetilde{\sigma} \in \Omega$. After the magnetisation phase, we can assume that the magnetisations of the two chains are the same. Thus, if $U_t = \widetilde{U}_t$, then this forces $V_t = \widetilde{V}_t$. Hence, it suffices to show that $U_t - \widetilde{U}_t$ hits zero to prove that the two chains have coalesced. The following lemma describes such a coupling.

Lemma 7.10. *Let σ and $\widetilde{\sigma}$ be two configurations with $S(\sigma) = S(\widetilde{\sigma})$. Define*

$$\Xi \coloneqq \left\{\sigma \in \Omega : \min\{U(\sigma),\, \overline{u}_0 - U(\sigma),\, V(\sigma),\, \overline{v}_0 - V(\sigma)\} \geq \frac{n}{16}\right\}, \tag{7.28}$$

the set of configurations where the vertices in the sets from (7.21) are "balanced". Then there exists a coupling $(X_t, \widetilde{X}_t)_{t\in\mathbb{N}}$ of the Glauber dynamics starting at σ and $\widetilde{\sigma}$ such that:

(i) *$S_t = \widetilde{S}_t$ for all $t \in \mathbb{N}$.*

(ii) *If $R_t \coloneqq U_t - \widetilde{U}_t$ and $R_0 \geq 0$, then $R_t \geq 0$ for all $t \in \mathbb{N}$, and*

$$\mathbb{E}_{\sigma,\widetilde{\sigma}}\left[R_{t+1} - R_t \mid X_t, \widetilde{X}_t\right] \leq 0. \tag{7.29}$$

(iii) *On the event $\{X_t \in \Xi,\, \widetilde{X}_t \in \Xi\}$, there exists $c > 0$, independent of n, such that*

$$\mathbb{P}_{\sigma,\widetilde{\sigma}}\left(R_{t+1} - R_t \neq 0 \mid X_t, \widetilde{X}_t\right) \geq c. \tag{7.30}$$

Proof. Assume that $U_0 \geq \widetilde{U}_0$ without loss of generality. The usual Glauber dynamics is used to update X_t. A vertex i is selected uniformly at random, and the spin of $X_t(i)$ is updated to $+1$ with probability $p_+(S_t - X_t(i)/n)$, and -1 otherwise. To update $\widetilde{X}_t$, a vertex $\widetilde{i}$ is selected uniformly at random from the set of vertices $\{i : \widetilde{X}_t(i) = X_t(i)\}$. The spin of $\widetilde{X}_t(\widetilde{i})$ is then updated with the same spin that was selected for $X_t(i)$, which is possible since both chains start with the same magnetisation.

i	$\widetilde{i}$	Spin selected	$R_{t+1} - R_t$
$i \in A(X_t)$	$\widetilde{i} \in C(\widetilde{X}_t)$	-1	-1
$i \in D(X_t)$	$\widetilde{i} \in B(\widetilde{X}_t)$	$+1$	-1
$i \in B(X_t)$	$\widetilde{i} \in D(\widetilde{X}_t)$	$+1$	$+1$
$i \in C(X_t)$	$\widetilde{i} \in A(\widetilde{X}_t)$	-1	$+1$
all other combinations			0

Figure 7.2.3: The combination of the vertices i, $\widetilde{i}$, and the spin selected to update $X_t(i)$ determines the increment $R_{t+1} - R_t$.

This defines a coupling such that $S_t = \widetilde{S}_t$. Recall the partitioning of the vertices by X_t and $\widetilde{X}_t$ from Figure 7.2.2. The process R_t increases or decreases by at most one depending on the combination of the vertices and spin selected, according to the table in Figure 7.2.3, and hence $R_t \geq 0$. Since $S_t = \widetilde{S}_t$, it follows that $R_t = U_t - \widetilde{U}_t = V_t - \widetilde{V}_t$. Recalling that $|A(X_t)| = U_t$, $|B(X_t)| = \overline{u}_0 - U_t$, $|C(X_t)| = \overline{v}_0 - V_t$, and $|D(X_t)| = V_t$ (and analogously for $\widetilde{X}_t$), we have

$$\begin{aligned}\mathbb{P}_{\sigma,\widetilde{\sigma}}\left(R_{t+1} - R_t = -1 \mid X_t, \widetilde{X}_t\right) = {} & \left(\frac{U_t}{n}\right)\left(\frac{\overline{v}_0 - V_t + R_t}{\overline{v}_0 - V_t + U_t}\right) p_-(S_t - n^{-1}) \\ & + \left(\frac{V_t}{n}\right)\left(\frac{\overline{u}_0 - U_t + R_t}{\overline{u}_0 - U_t + V_t}\right) p_+(S_t - n^{-1}),\end{aligned} \tag{7.31}$$

$$\mathbb{P}_{\sigma,\widetilde{\sigma}}\left(R_{t+1}-R_t=+1 \mid X_t,\widetilde{X}_t\right)=\left(\frac{\overline{u}_0-U_t}{n}\right)\left(\frac{V_t-R_t}{\overline{u}_0-U_t+V_t}\right)p_+(S_t-n^{-1}) \\ \left(\frac{\overline{v}_0-V_t}{n}\right)\left(\frac{U_t-R_t}{\overline{v}_0-V_t+U_t}\right)p_-(S_t-n^{-1}). \tag{7.32}$$

Therefore, the process R_t has non-positive drift (and so it is a supermartingale):

$$\mathbb{E}_{\sigma,\widetilde{\sigma}}\left[R_{t+1}-R_t \mid X_t,\widetilde{X}_t\right]=\frac{-R_t}{n}[p_+(S_t+n^{-1})+p_-(S_t-n^{-1})]\le 0.$$

Recall that $\frac{n}{4}\le \overline{u}_0,\overline{v}_0\le\frac{3n}{4}$ from (7.23), and that the functions p_+ and p_- are uniformly bounded away from zero from (7.10). On the event $\{X_t\in\Xi,\ \widetilde{X}_t\in\Xi\}$, the vertices in the sets $A(X_t)$, $B(X_t)$, $C(X_t)$, and $D(X_t)$ (and analogously for $\widetilde{X}_t$) are "balanced". Therefore, using (7.31) and (7.32) again implies that there exists a constant $c>0$, independent of n, such that $\mathbb{P}_{\sigma,\widetilde{\sigma}}\left(R_{t+1}-R_t\neq 0 \mid X_t,\widetilde{X}_t\right)\ge c$. □

To prove the upper bound of cutoff, it remains to control the probabilities of reaching or remaining in the various "nice" sets. Recall that $t_n=\frac{n\log n}{2(1-\beta)}$. For convenience, write $t_n(\alpha)=t_n+\alpha n$. Fix $(\overline{u}_0,\overline{v}_0)\in\Lambda_0$, and arbitrary $(\widetilde{u},\widetilde{v})\in\Lambda$. Let σ_0 be any configuration with $(U(\sigma_0),V(\sigma_0))=(\overline{u}_0,\overline{v}_0)$, and $\widetilde{\sigma}$ be any configuration with $(U(\widetilde{\sigma}),V(\widetilde{\sigma}))=(\widetilde{u},\widetilde{v})$. Recalling the definition of Ξ from (7.28), define

$$H_1(t)\coloneqq\{\tau_{\mathrm{mag}}\le t\},\quad\text{and}\quad H_2(t_1,t_2)\coloneqq\bigcap_{t=t_1}^{t_2}\{X_t\in\Xi,\ \widetilde{X}_t\in\Xi\}. \tag{7.33}$$

After the magnetisation phase, $\mathbb{P}_{\sigma_0,\widetilde{\sigma}}\left(H_1(t_n(\alpha))^c\right)=O(\alpha^{-1/2})$ from Lemma 7.8. Consider the set $A_0=\{i:\sigma_0(i)=+1\}$, noting that $|A_0|=\overline{u}_0$. Recall that $M_t(A_0)=\frac{1}{2}\sum_{i\in A_0}X_t(i)$ (and similarly for $\widetilde{M}_t(A_0)$). Assume that $U_{t_n(\alpha)}\ge\widetilde{U}_{t_n(\alpha)}$ (if not, simply reverse the roles of X_t and $\widetilde{X}_t$ in the following analysis).

First, we show that at $t_n(\alpha)$, the expected value of $|R_t|=|U_t-\widetilde{U}_t|$ is of order $\sqrt{n}$. Note that $2M_t(A_0)=U_t-(\overline{u}_0-U_t)$, and so $U_t=M_t(A_0)+\frac{\overline{u}_0}{2}$. Similarly, $\widetilde{U}_t=\widetilde{M}_t(A_0)+\frac{\overline{u}_0}{2}$. Therefore, $|U_t-\widetilde{U}_t|=|M_t(A_0)-\widetilde{M}_t(A_0)|\le|M_t(A_0)|+|\widetilde{M}_t(A_0)|$. Taking expectations and using Lemma 7.5 (iv) (c) shows that

$$\mathbb{E}_{\sigma_0,\widetilde{\sigma}}\left[|U_{t_n(\alpha)}-\widetilde{U}_{t_n(\alpha)}|\right]\le\mathbb{E}_{\sigma_0}\left[|M_{t_n(\alpha)}(A_0)|\right]+\mathbb{E}_{\widetilde{\sigma}}\left[|\widetilde{M}_{t_n(\alpha)}(A_0)|\right]=O(\sqrt{n}). \tag{7.34}$$

After t_n, we claim that in the next αn steps, (X_t) and $(\widetilde{X}_t)$ concentrate on Ξ. Let

$$Y\coloneqq\sum_{t_n(\alpha)\le t\le t_n(2\alpha)}\mathbb{1}_{\{|M_t(A_0)|>n/64\}}. \tag{7.35}$$

By Lemma 7.5 (iv) (b) and Markov's inequality, $\mathbb{P}_{\sigma_0,\widetilde{\sigma}}\left(|M_t(A_0)|>n/64\right)=O(n^{-1})$. Thus, $\mathbb{E}_{\sigma_0,\widetilde{\sigma}}[Y]=O(1)$. Since the increments of $M_t(A_0)$ are in $\{-1,0,+1\}$, it follows that if

$|M_{t_0}(A_0)| \geq n/32$ for some t_0, then $|M_t(A_0)| > n/64$ for all t in any interval of length $n/64$ containing t_0. Hence,

$$\mathbb{P}_{\sigma_0,\widetilde{\sigma}}\left(\bigcup_{t=t_n(\alpha)}^{t_n(2\alpha)}\left\{|M_t(A_0)| \geq n/32\right\}\right) \leq \mathbb{P}_{\sigma_0,\widetilde{\sigma}}\left(Y > n/64\right) \leq \frac{64\,\mathbb{E}_{\sigma_0,\widetilde{\sigma}}\left[Y\right]}{n} = O(n^{-1}). \tag{7.36}$$

Recall that $n/4 \leq \overline{u}_0 \leq 3n/4$. If either $U_{t_0} \leq n/16$, or $\overline{u}_0 - U_{t_0} \leq n/16$, for some t_0, then $|2M_{t_0}(A_0)| = |U_{t_0} - (\overline{u}_0 - U_{t_0})| \geq n/8$. (Similar results also hold by considering V_t, $\widetilde{U}_t$, and $\widetilde{V}_t$ instead.) Thus,

$$\mathbb{P}_{\sigma_0,\widetilde{\sigma}}\left(H_2(t_n(\alpha), t_n(2\alpha))^c\right) \leq \mathbb{P}_{\sigma_0,\widetilde{\sigma}}\left(\bigcup_{t=t_n(\alpha)}^{t_n(2\alpha)}\left\{|M_t(A_0)| \geq n/32\right\}\right) = O(n^{-1}). \tag{7.37}$$

Given that the events $H_1(t_n(\alpha))$ and $H_2(t_n(\alpha), t_n(2\alpha))$ hold (writing H_1 and H_2 below for simplicity), the assumptions of Lemma 7.10 are met. Therefore, the two-coordinate chains of (X_t) and $(\widetilde{X}_t)$ can be coupled during $t_n(\alpha) \leq t \leq t_n(2\alpha)$, such that the process $R_t = U_t - \widetilde{U}_t$ is a supermartingale with variance uniformly bounded away from zero. Hence, by Lemma 7.6, there exists a constant $c_1 > 0$ such that

$$\mathbb{P}_{\sigma_0,\widetilde{\sigma}}\left(\{\tau_2 > t_n(2\alpha)\} \cap H_1 \cap H_2 \mid X_{t_n(\alpha)}, \widetilde{X}_{t_n(\alpha)}\right) \leq \frac{c_1|R_{t_n(\alpha)}|}{\sqrt{n\alpha}}. \tag{7.38}$$

By taking expectations and using (7.34), we deduce that for some constant $c_* > 0$, $\mathbb{P}_{\sigma_0,\widetilde{\sigma}}\left(\{\tau_2 > t_n(2\alpha)\} \cap H_1 \cap H_2\right) \leq \frac{c_*}{\sqrt{\alpha}}$. Hence, together with (7.37), the coupling time of the two-coordinate chains is upper bounded by

$$\begin{aligned}\mathbb{P}_{\sigma_0,\widetilde{\sigma}}\left(\tau_2 > t_n(2\alpha)\right) &\leq \mathbb{P}_{\sigma_0,\widetilde{\sigma}}\left(\{\tau_2 > t_n(2\alpha)\} \cap H_1 \cap H_2\right) \\ &\quad + \mathbb{P}_{\sigma_0,\widetilde{\sigma}}\left(H_2^c\right) + \mathbb{P}_{\sigma_0,\widetilde{\sigma}}\left(H_1^c\right) \leq \frac{c_*}{\sqrt{\alpha}} + O(n^{-1}).\end{aligned} \tag{7.39}$$

To summarise, combining Lemma 7.7, Lemma 7.9, (7.26), and (7.27) yields

$$d_n(t_n + (\theta_0 + 2\alpha)n) \leq \mathbb{P}_{\sigma_0,\widetilde{\sigma}}\left(\tau_2 > t_n + 2\alpha n\right) \leq \frac{c_*}{\sqrt{\alpha}} + O(n^{-1}). \tag{7.40}$$

Therefore, $\lim_{\alpha\to\infty} \limsup_{n\to\infty} d_n(t_n + \alpha n) = 0$, as desired.

Proof of Theorem 7.4 – lower bound of cutoff

To obtain a lower bound for the distance to stationarity of the Glauber dynamics $(X_t)_{t\in\mathbb{N}}$, it will be sufficient to lower bound the distance to stationarity of its projection, the magnetisation chain $(S_t)_{t\in\mathbb{N}}$ (see [38, Lemma 7.9]).

Choose an initial state $s_0 \in \Omega_M$ satisfying $0 < s_0 < \frac{1-\beta}{3}$. Let $t_n^*(\alpha) := t_n - \frac{\alpha n}{1-\beta}$, and $\eta := 1 - \frac{1-\beta}{n}$. The main idea is to use the Taylor expansion $\tanh(x) = x - \frac{x^3}{3} + \frac{2x^5}{15} - O(x^7)$

to control the drift of the magnetisation chain from (7.11). By expanding $\tanh(\beta(s+n^{-1}))$ about βs in $f_n(s)$, and using $\theta_n(s) = O(n^{-2})$, the key inequality is that, for $S_t > 0$,

$$\mathbb{E}_{s_0}\left[|S_{t+1}| \mid S_t\right] \geq \eta\,|S_t| - \frac{|S_t|^3}{2n} - O(n^{-2}).$$

(This holds trivially for $S_t = 0$, and, by symmetry, also for $S_t < 0$.) Using the bounds on the first and second moments of S_t from Lemma 7.5, it can be shown (see [37, (3.24)–(3.27)] for the details) that, for sufficiently large n,

$$\mathbb{E}_{s_0}\left[|S_{t_n^*(\alpha)}|\right] \geq \frac{s_0\,\eta^{t_n^*(\alpha)}}{2} \geq B, \qquad \text{where } B := \frac{s_0 e^{\alpha}}{2n^{1/2}}. \tag{7.41}$$

Suppose that S_{π_n} is the normalised magnetisation under the stationary distribution π_n. By symmetry, $\mathbb{E}\left[S_{\pi_n}\right] = 0$. Using the stationary condition $\pi_n = \pi_n P_M^t$, and interchanging the order of summation, we see that

$$\mathrm{Var}(S_{\pi_n}) = \sum_{s\in\Omega_M} \pi_n(s)\,s^2 = \sum_{z\in\Omega_M}\left(\pi_n(z)\sum_{s\in\Omega_M} P_M^t(z,s)\,s^2\right) = \sum_{z\in\Omega_M}\pi_n(z)\mathbb{E}_z\left[S_t^2\right].$$

Hence, by Lemma 7.5 (iii) and (iv) (a), $\max\{\mathrm{Var}_{s_0}(S_t), \mathrm{Var}(S_{\pi_n})\} \leq \frac{c}{n}$ for some constant $c > 0$. Therefore, by using (7.41) with Chebyshev's inequality (Theorem 1.2), as well as the bound $\mathrm{Var}_{s_0}(|S_t|) \leq \mathrm{Var}_{s_0}(S_t)$, we have

$$\begin{aligned}
\mathbb{P}_{s_0}\left(|S_{t_n^*(\alpha)}| \leq \frac{B}{2}\right) &\leq \mathbb{P}_{s_0}\left(\left|\,|S_{t_n^*(\alpha)}| - \mathbb{E}_{s_0}\left[|S_{t_n^*(\alpha)}|\right]\right| \geq \frac{B}{2}\right) \leq \frac{\mathrm{Var}_{s_0}(|S_{t_n^*(\alpha)}|)}{B^2/4} \leq \frac{16ce^{-2\alpha}}{s_0^2},\\
\mathbb{P}\left(|S_{\pi_n}| \geq \frac{B}{2}\right) &\leq \frac{\mathrm{Var}(|S_{\pi_n}|)}{B^2/4} \leq \frac{16ce^{-2\alpha}}{s_0^2}.
\end{aligned} \tag{7.42}$$

Define the set $A := \left[\frac{-B}{2}, \frac{B}{2}\right]$. The inequalities (7.42) show that the distribution of the magnetisation chain at $t_n^*(\alpha) = \frac{\alpha n}{1-\beta}$ is well separated from π_n, with

$$\left\|\mathbb{P}_{s_0}\left(S_{t_n^*(\alpha)} \in \cdot\right) - \pi_n\right\|_{\mathrm{TV}} \geq \pi_n(A) - \mathbb{P}_{s_0}\left(|S_{t_n^*(\alpha)}| \in A\right) \geq 1 - \frac{32ce^{-2\alpha}}{s_0^2}.$$

The final term in the inequality above also provides a lower bound for $d_n\left(t_n^*(\alpha)\right)$. Hence, $\lim_{\alpha\to\infty}\liminf_{n\to\infty} d_n(t_n - \alpha n) = 1$, as desired.

Remark 7.11. Ding et al. [18, Proposition 3.9] show that the spectral gap of the Glauber dynamics $(X_t)_{t\in\mathbb{N}}$ is the same as the spectral gap of its magnetisation chain $(S_t)_{t\in\mathbb{N}}$. In the high temperature regime $\beta < 1$, they find that the spectral gap $\gamma_*^{(n)} = (1 + o(1))\frac{1-\beta}{n}$ ([18, Theorem 1]). Since we showed that $t_{\mathrm{mix}}^{(n)} = O(n\log n)$ in Theorem 7.2, the product condition $t_{\mathrm{mix}}^{(n)} \cdot \gamma_*^{(n)} \to \infty$ as $n \to \infty$ is satisfied, which is necessary for cutoff from Theorem 6.13. Indeed, there is actually a cutoff, which we have just proved.

7.3 Mixing at the critical temperature

The main result of this section is the following theorem, which establishes the order of the mixing time at the critical temperature $\beta = 1$. This was also proved by Levin et al. [37]. Again, we have chosen to omit some of the technical details from the paper for brevity.

Theorem 7.12. *If $\beta = 1$, then the Glauber dynamics for the Ising model on K_n has a mixing time of order $n^{3/2}$. That is, for some constants c_1, $c_2 > 0$, independent of n,*

$$c_1 n^{3/2} \leq t_{\text{mix}}^{(n)} \leq c_2 n^{3/2}.$$

Proof of Theorem 7.12 – upper bound

The strategy will be to show that two Glauber dynamics can be coupled such that their magnetisations agree in order $n^{3/2}$ steps. After that, we will show that they will coalesce with high probability in an additional order $n \log n$ steps.

Let $(X_t)_{t\in\mathbb{N}}$ be the Glauber dynamics started at σ, with magnetisation chain $(S_t)_{t\in\mathbb{N}}$. Let $\tau_0 \coloneqq \min\{t \in \mathbb{N} : |S_t| \leq 1/n\}$ be the first time that $|S_t|$ "hits zero" (accounting for the case of n odd). We claim that τ_0 is reached with high probability in order $n^{3/2}$ steps.

When $\beta = 1$, the Taylor expansion of $\tanh(x)$ shows that $\mathbb{E}[S_{t+1}] = S_t - \frac{S_t^3}{3n} + O(n^{-2})$ (see (7.11)). Let α be a constant satisfying $2/3 < \alpha \leq 1$. By analysing the moments of the magnetisation chain in order to control the drift of $|S_t|$, it can be shown (see [37, (4.2)]) that there exists a constant $c_1 = c_1(\alpha) > 0$ such that

$$\mathbb{E}_\sigma\left[|S_{c_1 n^{3-2\alpha}}| \cdot \mathbb{1}_{\{\tau_0 > c_1 n^{3-2\alpha}\}}\right] \leq n^{\alpha-1}. \tag{7.43}$$

Recall that if $|S_t| > \frac{1}{n}$, then $|S_t|$ has non-positive drift from (7.12). Moreover, its holding probability is uniformly bounded away from zero by (7.10). Therefore, using Lemma 7.6 with the stopping time τ_0 implies that for some constant $c_2 > 0$,

$$\mathbb{P}_\sigma\left(\tau_0 > c_1 n^{3-2\alpha} + \gamma n^{2\alpha} \mid X_{c_1 n^{3-2\alpha}}\right) \leq \frac{c_2 n |S_{c_1 n^{3-2\alpha}}|}{\sqrt{\gamma} n^\alpha}. \tag{7.44}$$

Choose $\alpha = 3/4$, so that $3 - 2\alpha = 2\alpha$. Multiplying both sides of (7.44) by $\mathbb{1}_{\{\tau_0 > c_1 n^{3-2\alpha}\}}$, taking expectations, and then using (7.43) shows that

$$\mathbb{P}_\sigma\left(\tau_0 > (c_1 + \gamma) n^{3/2}\right) = O(\gamma^{-1/2}). \tag{7.45}$$

This bound on τ_0 will be used with the following lemma, based off [18, Lemma 3.1].

Lemma 7.13. *Let $(S_t)_{t\in\mathbb{N}}$ and $(\widetilde{S}_t)_{t\in\mathbb{N}}$ be two magnetisation chains that are started from arbitrary states s and $\widetilde{s}$ respectively. Recall that $\tau_{\text{mag}} = \min\{t \in \mathbb{N} : S_t = \widetilde{S}_t\}$, and $\tau_0 = \min\{t \in \mathbb{N} : |S_t| \leq 1/n\}$. Suppose that $\mathbb{P}_\sigma(\tau_0 \geq T) < \epsilon$ for some $T > 0$ and $0 < \epsilon < 1$. Then there exists a constant $c_\epsilon > 0$, and a coupling $(S_t, \widetilde{S}_t)_{t\in\mathbb{N}}$, such that*

$$\mathbb{P}_{s,\widetilde{s}}(\tau_{\text{mag}} \leq c_\epsilon T) \geq 1 - \epsilon.$$

Proof. Assume that $|\widetilde{S}_0| < |S_0|$ without loss of generality. Define $G_1 \coloneqq \{\tau_0 < T\}$, which occurs with probability at least $1 - \epsilon$, by assumption. Denote the first time that $|S_t|$ is adjacent to $|\widetilde{S}_t|$ by $\tau_{\text{abs}} \coloneqq \min\{t \in \mathbb{N} : |S_t| \leq |\widetilde{S}_t| + \frac{2}{n}\}$. By definition, $\tau_{\text{abs}} < \tau_0$. Let S_t and $\widetilde{S}_t$ run independently until $\tau_{\text{abs}} + 1$.

If S_t and $\widetilde{S}_t$ are within one step of each other (i.e. $|S_t - \widetilde{S}_t| \leq \frac{2}{n}$), then they will coalesce in the next step with probability bounded away from zero. This is clear if $S_t = \widetilde{S}_t$. Otherwise, if the two chains are adjacent, then they will coalesce if the chain on the "outside" (i.e. further away from zero) moves in, and the chain on the "inside" remains still. From (7.10), both these probabilities are uniformly bounded away from 0 and 1. Hence, there exists a constant $0 < b < 1$ such that

$$\mathbb{P}_{s,\widetilde{s}}\left(S_{t+1} = \widetilde{S}_{t+1} \mid |S_t - \widetilde{S}_t| \leq \frac{2}{n}\right) > b. \tag{7.46}$$

(For example, $b = \frac{1}{8}[1 - \tanh(\beta)]$ will suffice for sufficiently large n.) Define the event $G_2 \coloneqq \{|S_{\tau_{\text{abs}}+1}| = |\widetilde{S}_{\tau_{\text{abs}}+1}|\}$. Due to the symmetry of $(S_t)_{t\in\mathbb{N}}$ and $(-S_t)_{t\in\mathbb{N}}$, G_2 occurs with probability at least b by (7.46).

Suppose that G_2 holds. Then we claim that S_t and $\widetilde{S}_t$ will coalesce at some time $t \leq \tau_0 + 1$ with probability at least b. This is clear if $S_{\tau_{\text{abs}}+1} = \widetilde{S}_{\tau_{\text{abs}}+1}$. Otherwise, if $S_{\tau_{\text{abs}}+1} = -\widetilde{S}_{\tau_{\text{abs}}+1}$, then by the reflective symmetry of the magnetisation chain again, S_t and $\widetilde{S}_t$ can be coupled such that $S_t = -\widetilde{S}_t$ (i.e. reflection coupling). Hence, at time τ_0, $|S_{\tau_0} - \widetilde{S}_{\tau_0}| = 2|S_{\tau_0}| \leq \frac{2}{n}$. (If n is even, $S_{\tau_0} = \widetilde{S}_{\tau_0}$ already.)

Define the event $G_3 \coloneqq \{S_{\tau_0+1} = \widetilde{S}_{\tau_0+1}\}$. By (7.46), S_{τ_0} and $\widetilde{S}_{\tau_0}$ will coalesce in the next step at time $\tau_0 + 1$ with probability at least b. Since G_2 and G_3 are independent of G_1, $\mathbb{P}_{s,\widetilde{s}}(G_1 \cap G_2 \cap G_3) = \mathbb{P}_{s,\widetilde{s}}(G_1)\,\mathbb{P}_{s,\widetilde{\sigma}}(G_2)\,\mathbb{P}_{s,\widetilde{s}}(G_3 \mid G_2) \geq b^2(1-\epsilon)$. Hence,

$$\mathbb{P}_{s,\widetilde{s}}(\tau_{\text{mag}} \leq T) \geq \mathbb{P}_{s,\widetilde{s}}(G_1 \cap G_2 \cap G_3) \geq b^2(1-\epsilon). \tag{7.47}$$

Finally, to extend the result, the above process can be repeated in blocks of T time units until coalescence. If the process is repeated once, then

$$\mathbb{P}_{s,\widetilde{s}}(\tau_{\text{mag}} > 2T) = \mathbb{P}_{s,\widetilde{s}}\left(\tau_{\text{mag}} > 2T \mid S_t \neq \widetilde{S}_T\right)\mathbb{P}_{s,\widetilde{s}}(\tau > T) \leq (1 - b^2(1-\epsilon))^2.$$

By induction, if we repeat the process $\ell \in \mathbb{Z}_+$ times, $\mathbb{P}_{s,\widetilde{s}}(\tau_{\text{mag}} > \ell T) \leq (1 - b^2(1-\epsilon))^\ell$. Hence $\mathbb{P}_{s,\widetilde{s}}(\tau_{\text{mag}} \leq \ell T) \geq 1 - (1 - b^2(1-\epsilon))^\ell$. Since $1 - b^2(1-\epsilon) < 1$, it follows that we can choose a sufficiently large c_ϵ such that $\mathbb{P}_{s,\widetilde{s}}(\tau_{\text{mag}} > c_\epsilon T) \geq 1 - \epsilon$, as desired. □

Hence, combining (7.45) and Lemma 7.13 implies that $\mathbb{P}_{\sigma,\widetilde{\sigma}}\left(\tau_{\text{mag}} > c_* n^{3/2}\right) < 1/8$ for some constant $c_* > 0$.

For a coupling $(X_t, \widetilde{X}_t)_{t\in\mathbb{N}}$ of two copies of the Glauber dynamics, denote their coupling time by $\tau_{\text{couple}} \coloneqq \min\{t \in \mathbb{N} : X_t = \widetilde{X}_t\}$. To conclude, the next lemma shows that after the magnetisations of X_t and $\widetilde{X}_t$ match, they will coalesce with high probability after an additional order $n \log n$ steps.

Lemma 7.14 ([37, Lemma 2.9]). *Let $\sigma, \widetilde{\sigma} \in \Omega$ be configurations such that $S(\sigma) = S(\widetilde{\sigma})$. Then there exists a coupling $(X_t, \widetilde{X}_t)_{t\in\mathbb{N}}$ of the Glauber dynamics started at σ and $\widetilde{\sigma}$ respectively, such that for some constant $c_0 = c_0(\beta) > 0$, independent of n,*

$$\mathbb{P}_{\sigma,\widetilde{\sigma}}\left(\tau_{\text{couple}} > c_0 n \log n\right) \leq \frac{1}{n}. \tag{7.48}$$

Proof. The Glauber dynamics is used to update X_t. A vertex i is selected uniformly at random, and the spin of $X_t(i)$ is updated to $+1$ with probability $p_+(S_t - X_t(i)/n)$, and -1 otherwise. If $\widetilde{X}_t(i) = X_t(i)$, then the same update can be performed for $\widetilde{X}_t$.

Otherwise, if $\widetilde{X}_t(i) \neq X_t(i)$, a vertex $\widetilde{i}$ is selected uniformly at random from the set of vertices $\{j : \widetilde{X}_t(j) = X_t(i),\ \widetilde{X}_t(j) \neq X_t(j)\}$. Then $\widetilde{X}_t(\widetilde{i})$ is updated with the same spin as was selected for $X_t(i)$. This is possible since $S(\sigma) = S(\widetilde{\sigma})$, and the coupling maintains $S_t = \widetilde{S}_t$.

Let $D_t = \operatorname{dist}(X_t, \widetilde{X}_t)$ count the different vertices between X_t and $\widetilde{X}_t$. This can only decrease under the chosen coupling. If $X_t(i) \neq \widetilde{X}_t(i)$ and the spin of $X_t(i)$ is flipped, then $D_{t+1} - D_t = -2$. This occurs with probability at least $\min\{p_+, p_-\}$, which is uniformly bounded away from zero by some constant $b > 0$, by (7.10). Otherwise, $D_{t+1} - D_t = 0$.

Hence, $\mathbb{E}_{\sigma,\widetilde{\sigma}}\left[D_{t+1} - D_t\right] \leq -\frac{2bD_t}{n}$. It follows that $Y_t = D_t\,(1-2b/n)^{-t}$ is a non-negative supermartingale (i.e. $\mathbb{E}_{\sigma,\widetilde{\sigma}}\left[Y_{t+1} \mid X_t\right] \leq Y_t$). Therefore,

$$\mathbb{E}_{\sigma,\widetilde{\sigma}}\left[D_t\right] \leq \mathbb{E}_{\sigma,\widetilde{\sigma}}\left[D_0\right]\left(1 - \frac{2b}{n}\right)^t \leq ne^{-2bt/n}. \tag{7.49}$$

Let $t = c_0 n \log n$, where $c_0 = c_0(\beta) > 0$ is a sufficiently large constant such that the upper bound in (7.49) is less than $\frac{1}{n}$. Then by Markov's inequality (Theorem 1.1),

$$\mathbb{P}_{\sigma,\widetilde{\sigma}}\left(\tau_{\text{couple}} > c_0 n \log n\right) = \mathbb{P}_{\sigma,\widetilde{\sigma}}\left(D_{c_0 n \log n} \geq 1\right) \leq \mathbb{E}_{\sigma,\widetilde{\sigma}}\left[D_{c_0 n \log n}\right] \leq \frac{1}{n},$$

as desired. □

Let $H_1(t) = \{\tau_{\text{mag}} \leq t\}$. On the event $H_1(c_* n^{3/2})$ (writing H_1 for simplicity), by Lemma 7.14, there exists a constant c_0 such that

$$\mathbb{P}_{\sigma,\widetilde{\sigma}}\left(\{\tau_{\text{couple}} > c_* n^{3/2} + c_0 n \log n\} \cap H_1 \mid X_{c_* n^{3/2}}, \widetilde{X}_{c_* n^{3/2}}\right) \leq \frac{1}{n}.$$

Taking expectations shows that $\mathbb{P}_{\sigma,\widetilde{\sigma}}\left(\{\tau_{\text{couple}} > c_* n^{3/2} + c_0 n \log n\} \cap H_1\right) \leq 1/n$.

To summarise the proof, we have, by the coupling theorem (Theorem 3.3),

$$\begin{aligned} d_n(c_* n^{3/2} + c_0 n \log n) &\leq \mathbb{P}_{\sigma,\widetilde{\sigma}}\left(\tau_{\text{couple}} > c_* n^{3/2} + c_0 n \log n\right) \\ &\leq \mathbb{P}_{\sigma,\widetilde{\sigma}}\left(H_1^c\right) + \mathbb{P}_{\sigma,\widetilde{\sigma}}\left(\{\tau_{\text{couple}} > c_* n^{3/2} + c_0 n \log n\} \cap H_1\right) \leq \frac{1}{8} + \frac{1}{n}. \end{aligned}$$

For sufficiently large n, the final expression is less than $\frac{1}{4}$. Therefore, this shows that $t_{\text{mix}} \leq c_* n^{3/2} + c_0 n \log n = O(n^{3/2})$, as desired.

Proof of Theorem 7.12 – lower bound

As in the high temperature regime, it will suffice to lower bound the distance to stationarity of the magnetisation chain, which is a projection of the Glauber dynamics. Recall that S_{π_n} denotes the normalised magnetisation under the stationary distribution π_n.

For $\beta = 1$, the following shows that $n^{1/4} S_{\pi_n}$ converges to a limiting distribution.

Proposition 7.15 ([18, Theorem 5.1])**.** *As $n \to \infty$,*

$$n^{1/4} S_{\pi_n} \to Z \exp\left(-\frac{s^4}{12}\right), \qquad s \in \mathbb{R}, \tag{7.50}$$

where the convergence is in distribution, and Z is a normalising constant.

Thus, by Proposition 7.15, we can choose a constant $A > 0$ such that

$$\mathbb{P}\left(|S_{\pi_n}| \leq A n^{-1/4}\right) \geq \frac{3}{4}. \tag{7.51}$$

Let $s_0 := 2An^{-1/4}$. We claim that the distribution of $(S_t)_{t \in \mathbb{N}}$ starting from s_0 remains well separated from π_n after order $n^{3/2}$ steps. The strategy is to construct a modified version of the magnetisation chain, also starting at s_0, that stochastically dominates S_t. By using the Taylor expansion of $\tanh(x)$ for the moments of the magnetisation chain again, the drift of this modified chain towards $An^{-1/4}$ can be bounded from above (see [37, (4.6)–(4.12)] for the full details). Thus, it can be shown that there exists a constant $c_A > 0$ such that

$$\mathbb{P}_{s_0}\left(S_{c_A n^{3/2}} \leq A n^{-1/4}\right) \leq \frac{1}{4}. \tag{7.52}$$

Define the set $B := [-An^{-1/4}, An^{-1/4}]$. By combining (7.51) and (7.52), we deduce that

$$\|\pi_n - \mathbb{P}_{s_0}(S_t \in \cdot)\|_{\text{TV}} \geq \pi_n(B) - \mathbb{P}_{s_0}(S_t \in B) \geq \frac{1}{2}.$$

Since the last term in the inequality above also provides a lower bound for $d_n(c_A n^{3/2})$, it follows that $t_{\text{mix}} \geq c_A n^{3/2}$, as desired.

Remark 7.16. At the critical temperature $\beta = 1$, Ding et al. [18, Theorem 2] show that the spectral gap $\gamma_*^{(n)}$ of the Glauber dynamics $(X_t)_{t \in \mathbb{N}}$ is of order $n^{3/2}$. Recall from Theorem 7.12 that $c_1 n^{3/2} \leq t_{\text{mix}}^{(n)} \leq c_2 n^{3/2}$ for some $c_1, c_2 > 0$. Hence, $t_{\text{mix}}^{(n)} \cdot \gamma_*^{(n)} = O(1)$, which does not tend to infinity, and so there is no cutoff by Theorem 6.13. Therefore, we conclude that the transition to stationarity occurs gradually at around order $n^{3/2}$ steps.

7.4 Exponentially slow mixing at low temperatures

The main result of this final section will be the following theorem, which shows that there is exponentially slow mixing when $\beta > 1$. The conductance technique from Section 4.2 will be used by explicitly identifying a particularly "bad" set that inhibits mixing.

Theorem 7.17. *Suppose that $\beta > 1$ for the Glauber dynamics for the Ising model on K_n. Then there exists real constants $b > 0$ and $r(\beta) > 0$ such that $t_{\text{mix}}^{(n)} \geq be^{nr(\beta)}$.*

Proof. This closely follows [38, Theorem 15.4] with additional details. Denote the set of configurations with exactly k plus spins by $A_k \coloneqq \{\sigma \in \Omega : S(\sigma) = \frac{2k}{n} - 1\}$. By the symmetry of K_n, counting edges shows that $\pi(A_k) = a_k/Z(\beta)$, where

$$a_k = \binom{n}{k} \exp\left(\frac{\beta}{n}\left[\binom{k}{2} + \binom{n-k}{2} - k(n-k)\right]\right). \tag{7.53}$$

By taking logarithms and using Stirling's formula $\log n! = n\log n - n + \frac{1}{2}\log(2\pi n) + o(1)$, it can be verified that $\log(a_k) = n\psi_\beta(c)[1 + o(1)]$, where $c = \frac{k}{n}$, and

$$\psi_\beta(c) = -c\log(c) - (1-c)\log(1-c) + \beta\left[\frac{(1-2c)^2}{2}\right], \quad 0 \leq c \leq 1. \tag{7.54}$$

Here the term $x\log(x)$ is taken to be 0 if $x = 0$. (Note that the first two terms of ψ_β is the binary entropy function.) Since ψ_β is continuous, and $|c - \widetilde{c}| = o(n^{-1})$ (where $\widetilde{c} = \frac{\lfloor cn \rfloor}{n}$), it follows that the error $n|\psi_\beta(c) - \psi_\beta(\widetilde{c})|$ is also $o(n)$. Hence,

$$\log(a_{\lfloor cn \rfloor}) = n\psi_\beta(c)[1 + o(1)], \quad 0 \leq c \leq 1. \tag{7.55}$$

Next, standard calculus on (7.54) shows that $\psi_\beta'(1/2) = 0$ and $\psi_\beta''(1/2) = -4(1-\beta)$. Since $\beta > 1$, $c = \frac{1}{2}$ is a local minimum. Since $\psi_\beta(0) = \frac{\beta}{2}$ and ψ_β is continuous, by the Extreme Value Theorem, there exists a real $c_* < 1/2$ that maximises ψ_β over $[0, 1/2]$.

We claim that $B \coloneqq \{\sigma : S(\sigma) < 0\}$ is a set that has poor conductance. By symmetry, B has the same stationary probability as the set of all configurations with positive normalised magnetisation, and so $\pi(B) \leq \frac{1}{2}$. Since $c_* < 1/2$, $\lfloor c_* n \rfloor$ must be an integer between 0 and $\lfloor n/2 \rfloor$. Therefore, $\pi(B) = \sum_{k=0}^{\lfloor n/2 \rfloor} \pi(A_k) \geq \pi(A_{\lfloor c_* n \rfloor})$.

If n is odd, then $B = \bigcup_{k=0}^{\lfloor n/2 \rfloor} A_k$. Since only one spin is changed at a time, the only way from B to B^c is to start in $A_{\lfloor n/2 \rfloor}$. Thus, recalling the definition of the edge measure from (4.8), $Q(B, B^c) = \sum_{\sigma \in A_{\lfloor n/2 \rfloor}} \pi(\sigma)P(\sigma, B^c) \leq \pi(A_{\lfloor n/2 \rfloor})$. If n is even, then we can use the symmetry of the edge measure, $Q(B, B^c) = Q(B^c, B)$, from Lemma 4.10 (ii) to reach the same conclusion.

Recalling that $\pi(A_k) = a_k/Z(\beta)$, putting (7.55) into these bounds for $\pi(B)$ and $Q(B, B^c)$ shows that the conductance, defined in (4.9), is upper bounded by

$$\Phi_* \leq \Phi(B) = \frac{Q(B, B^c)}{\pi(B)} \leq \frac{\exp\left(n\psi_\beta(1/2)[1 + o(1)]\right)}{\exp\left(n\psi_\beta(c_*)[1 + o(1)]\right)}. \tag{7.56}$$

Since c_* maximises ψ_β on $[0, 1/2]$, $\psi_\beta(c^*) - \psi_\beta(1/2) > 0$. Let $\epsilon = \frac{\psi_\beta(c_*)-\psi_\beta(1/2)}{2(\psi_\beta(c_*)+\psi_\beta(1/2))} > 0$. Thus, for sufficiently large n, (7.56) implies that the conductance is upper bounded by

$$\exp\big(-n\big[\big(\psi_\beta(c_*) - \psi_\beta(1/2)\big) - \epsilon\big(\psi_\beta(c_*) + \psi_\beta(1/2)\big)\big]\big) = \exp\left(-n\left[\frac{\psi_\beta(c_*) - \psi_\beta(1/2)}{2}\right]\right).$$

Therefore, we can define $r(\beta) := \frac{1}{2}\big(\psi_\beta(c_*) - \psi_\beta(c)\big) > 0$. We have shown that for sufficiently large n, $\Phi_* \le e^{-nr(\beta)}$. Furthermore, we can choose $b > 0$ such that $\Phi_* \le be^{-nr(\beta)}$ for all n. By Theorem 4.12, we conclude that $t_{\text{mix}}^{(n)} \ge be^{nr(\beta)}$, as desired. □

Remark 7.18. (i) The "restricted dynamics" (which moves between configurations with positive magnetisation only) is rapidly mixing with $t_{\text{mix}} \asymp n \log n$, even in this low temperature regime ([37, Theorem 3]). Hence, crossing the "magnetisation bridge" to change the signs (represented by the set B in the proof of Theorem 7.17) is the severe bottleneck of the Glauber dynamics.

(ii) Ding et al. [18, Theorem 3] show that if $\beta > 1$, then the spectral gap $\gamma_*^{(n)}$ of the Glauber dynamics $(X_t)_{t\in\mathbb{N}}$ satisfies $t_{\text{mix}}^{(n)} \cdot \gamma_*^{(n)} = O(1)$. Since the product condition does not hold, Theorem 6.13 implies that there is no cutoff.

Chapter 8

Concluding Remarks

In this thesis, we have surveyed some of the main techniques used to study the mixing times of Markov chains. The development of the probabilistic and analytical tools described in Chapters 3 and 4, as well as from other areas, is an interesting research area. Extending the bounds to more general classes of chains (e.g. non-reversible chains, and chains with uncountable state spaces) would also be valuable.

We have also described some problems requiring a careful understanding of mixing times. In Chapter 5, we discussed the key challenge of finding a rapidly mixing Markov chain in order to present a FPRAS for many "difficult" counting problems, which has led to the development of new tools. It is still an open question whether rapidly mixing chains exist for some problems (such as for sampling proper q-colourings of a graph of maximum degree Δ with $q \geq \Delta + 1$). Furthermore, it is also of practical interest to provide more useful bounds for rapidly mixing chains (after all, n^{20} is still polynomial).

A detailed study of the Glauber dynamics for the mean-field Ising model was conducted in Chapter 7, which was found to either mix rapidly or torpidly, depending on the temperature. At high temperatures, very precise control of the mixing time was required to prove that it exhibits a cutoff (analogous to a "probabilistic phase transition"). The many examples of cutoff provided in Chapter 6 suggests that it may be a more general phenomenon than expected. Proving cutoff for more chains, finding more necessary and sufficient conditions for cutoff, as well as the development of a general theory of cutoff (i.e. what makes it cutoff?) are active research problems.

We conclude with the note that essentially none of the applications of the wonderful MCMC technique (e.g. in statistics, physics, chemistry, biology, etc.) are accompanied by any practically useful bounds on runtime [14]. Hence, there are numerous problems of practical and theoretical interest in the study of the mixing times of Markov chains, which will likely require the refinement of existing tools, as well as new ideas to answer.

References

[1] D. Aldous. "Random Walks on Finite Groups and Rapidly Mixing Markov Chains". In: *Seminar on Probability* 17 (1983), pp. 243–297.

[2] D. Aldous and P. Diaconis. "Shuffling Cards and Stopping Times". In: *The American Mathematical Monthly* 93 (1986), pp. 333–348.

[3] D. Aldous and P. Diaconis. "Strong uniform times and finite random walks". In: *Advances in Applied Mathematics* 8.1 (1987), pp. 69–97.

[4] D. Aldous and J.A. Fill. "Reversible Markov Chains and Random Walks on Graphs". Unfinished monograph, available from https://www.stat.berkeley.edu/users/aldous/RWG/book.pdf. 2002.

[5] R. Basu, J. Hermon, and Y. Peres. "Characterization of Cutoff for Reversible Markov Chains". In: *The Annals of Probability* 45.3 (2017), pp. 1448–1487.

[6] D. Bayer and P. Diaconis. "Trailing the Dovetail Shuffle to its Lair". In: *The Annals of Applied Probability* 2 (1992), pp. 294–313.

[7] R. Bubley and M. Dyer. "Path Coupling: A Technique for Proving Rapid Mixing in Markov Chains". In: *Proceedings 38th Annual Symposium on Foundations of Computer Science*. 1997, pp. 223–231.

[8] R. Bubley and M.E. Dyer. *Path Coupling, Dobrushin Uniqueness, and Approximate Counting*. Technical Report 97.04, School of Computer Studies, University of Leeds. 1997.

[9] S. Chatterjee, P. Diaconis, A. Sly, and L. Zhang. *A Phase Transition for Repeated Averages*. 2020. arXiv: 1911.02756 [math.PR].

[10] G.Y. Chen and L. Saloff-Coste. "The Cutoff Phenomenon for Ergodic Markov Processes". In: *Electronic Journal of Probability* 13 (2008), pp. 26–78.

[11] M. Delcourt, G. Perarnau, and L. Postle. *Rapid mixing of Glauber dynamics for colorings below Vigoda's* 11/6 *threshold*. 2018. arXiv: 1804.04025 [cs.DM].

[12] P. Diaconis. "Group Representations in Probability and Statistics". In: *Lecture Notes-Monograph Series* 11 (1988), pp. 1–192.

[13] P. Diaconis. "The Cutoff Phenomenon in Finite Markov Chains". In: *Proc. Natl. Acad. Sci. USA*. Vol. 93. 1996, pp. 1659–1664.

[14] P. Diaconis. "The Markov Chain Monte Carlo Revolution". In: *Bulletin of the American Mathematical Society* 46 (2009), pp. 179–205.

[15] P. Diaconis and L. Saloff-Coste. "Separation Cut-offs for Birth and Death Chains". In: *Annals of Applied Probability* 16.4 (2006), pp. 2098–2122.

[16] P. Diaconis and M. Shahshahani. "Generating a Random Permutation with Random Transpositions". In: *Z. Wahrscheinlichkeitstheorie verw Gebiete* 57 (1981), pp. 159–179.

[17] P. Diaconis and D. Stroock. "Geometric bounds for eigenvalues of Markov chains". In: *Annals of Applied Probability* 1 (1991), pp. 36–61.

[18] J. Ding, E. Lubetzky, and Y. Peres. "The Mixing Time Evolution of Glauber Dynamics for the Mean-Field Ising Model". In: *Communications in Mathematical Physics* 289.2 (2009), pp. 725–764.

[19] J. Ding, E. Lubetzky, and Y. Peres. “Total Variation Cutoff in Birth-and-Death Chains”. In: *Probability Theory and Related Fields* 146 (2010), pp. 61–85.

[20] M. Dyer, A. Frieze, and R. Kannan. “A Random Polynomial-time Algorithm for Approximating the Volume of Convex Bodies”. In: *38* 38 (1991), pp. 1–17.

[21] M. Dyer and C. Greenhill. “A More Rapidly Mixing Markov Chain for Graph Colourings”. In: *Random Structures & Algorithms* 13 (1998), pp. 285–317.

[22] M. Dyer, C. Greenhill, and M. Ullrich. “Structure and Eigenvalues of Heat-Bath Markov Chains”. In: *Linear Algebra and its Applications* 454 (2014), pp. 57–71.

[23] P. Erdős and A. Rényi. “On a Classical Problem of Probability Theory”. In: *Magyar Tud. Akad. Mat. Kutató Int. Közl.* 6 (1961), pp. 215–220.

[24] W. Feller. *An Introduction to Probability Theory and its Applications*. 3rd ed. Vol. 1. New York: Wiley, 1986.

[25] C. Greenhill. *Making Markov chains less lazy*. 2013. arXiv: 1203.6668 [math.CO].

[26] G.R. Grimmett and D.R. Stirzaker. *Probability and Random Processes*. 3rd ed. New York: Oxford University Press, 2001.

[27] P.R. Halmos. *Finite-Dimensional Vector Spaces*. 1st ed. New York: Springer-Verlag, 1987.

[28] T.P. Hayes. “A Simple Condition Implying Rapid Mixing of Single-Site Dynamics on Spin Systems”. In: *Proceedings of the 47th Annual IEEE Symposium on Foundations of Computer Science*. 2006, pp. 39–46.

[29] M. Jerrum. “A Very Simple Algorithm for Estimating the Number of k-Colorings of a Low-Degree Graph”. In: *Random Structures & Algorithms* 7.2 (1995), pp. 157–165.

[30] M. Jerrum. *Counting, Sampling and Integrating: Algorithms and Complexity*. Birkhäuser, Basel: Lectures in Mathematics ETH Zürich, 2003.

[31] M. Jerrum and A. Sinclair. “Approximating the Permanent”. In: *SIAM Journal on Computing* 18.6 (1989), pp. 1149–1178.

[32] M. Jerrum and A. Sinclair. “Polynomial-time Approximation Algorithms for the Ising Model”. In: *SIAM Journal on Computing* 22.5 (1993), pp. 1087–1116.

[33] M. Jerrum and A. Sinclair. “The Markov Chain Monte Carlo Method: An Approach to Approximate Counting and Integration”. In: *Approximation Algorithms for NP-hard Problems*. Ed. by D. Hochbaum. PWS Publishing Company, 1996. Chap. 12, pp. 482–522.

[34] M. Jerrum, A. Sinclair, and E. Vigoda. “A Polynomial-Time Approximation Algorithm for the Permanent of a Matrix with Nonnegative Entries”. In: *Journal of the ACM* 51.4 (2004), pp. 671–697.

[35] M.R. Jerrum, L.G. Valiant, and V.V. Vazirani. “Random Generation of Combinatorial Structures from a Uniform Distribution”. In: *Theoretical Computer Science* 43 (1986), pp. 169–188.

[36] D.A. Levin. “Cutoff for Markov Chains”. In: *LMS - EPSRC Durham Symposium 2017*. pages.uoregon.edu/dlevin/TALKS/durham.pdf. 2017.

[37] D.A. Levin, M.J. Luczak, and Y. Peres. “Glauber Dynamics for the Mean-Field Ising Model: Cut-off, Critical Power Law, and Metastability”. In: *Probability Theory and Related Fields* 146 (2010), pp. 223–265.

[38] D.A. Levin, Y. Peres, and E.L. Wilmer. *Markov Chains and Mixing Times*. 1st ed. Providence, Rhode Island: American Mathematical Society, 2009.

[39] T. Lindvall. *Lectures on the Coupling Method*. Mineola, New York: Dover, 1992.

[40] E. Lubetzky and A. Sly. “Cutoff Phenomena for Random Walks on Random Regular Graphs”. In: *Duke Mathematical Journal* 153.3 (2010), pp. 475–510.

[41] E. Lubetzky and A. Sly. "Critical Ising on the Square Lattice Mixes in Polynomial Time". In: *Communications in Mathematical Physics* 313 (2013), pp. 815–836.

[42] E. Lubetzky and A. Sly. "Cutoff for the Ising Model on the Lattice". In: *Inventiones Mathematicae* 191 (2013), pp. 719–755.

[43] E. Lubetzky and A. Sly. "Universality of cutoff for the Ising model". In: *The Annals of Probability* 45.6A (2017), pp. 3664–3696.

[44] A.A. Markov. "Extension of the Law of Large Numbers to Dependent Events (Russian)". In: *Bull. Soc. Phys. Math. Kazan* 2.15 (1906), pp. 135–156.

[45] F. Martinelli. "Lectures on Glauber Dynamics for Discrete Spin Models". In: *Lectures on Probability Theory and Statistics*. Ed. by P. Bernard. Berlin, Heidelberg: Springer-Verlag, 1999, pp. 93–191.

[46] R. Montenegro and P. Tetali. *Mathematical Aspects of Mixing in Markov Chains*. Hanover, MA: Now Publishers Inc., 2006.

[47] B. Morris and A. Sinclair. "Random Walks on Truncated Cubes and Sampling 0-1 Knapsack Solutions". In: *SIAM Journal on Computing* 34 (2004), pp. 195–226.

[48] J. Propp and D. Wilson. "Coupling from the Past: a User's Guide". In: *Microsurveys in Discrete Probability* 41 (1998), pp. 181–192.

[49] J.G. Propp and D.B. Wilson. "Exact Sampling with Coupled Markov Chains and Applications to Statistical Mechanics". In: *Random Structures & Algorithms* 9.1-2 (1996), pp. 223–252.

[50] G.O. Roberts and J.S. Rosenthal. "General state space Markov chains and MCMC algorithms". In: *Probability Surveys* 1 (2004), pp. 20–71.

[51] W. Rudin. *Principles of Mathematical Analysis*. 3rd ed. Singapore: McGraw-Hill, 1976.

[52] L. Saloff-Coste. "Lectures on Finite Markov Chains". In: *Lectures on Probability Theory and Statistics*. Ed. by P. Bernard. Vol. 1665. Lecture Notes in Mathematics. Berlin, Heidelberg: Springer, 1997, pp. 301–413.

[53] E. Seneta. *Non-negative Matrices and Markov Chains*. 2nd ed. New York: Springer-Verlag, 1981.

[54] A. Sinclair. "Improved Bounds for Mixing Rates of Markov Chains and Multicommodity Flow". In: *Combinatorics, Probability and Computing* 1.4 (1992), pp. 351–370.

[55] L.G. Valiant. "The Complexity of Enumeration and Reliability Problems". In: *SIAM Journal on Computing* 8.3 (1979), pp. 410–421.

[56] E. Vigoda. "Improved Bounds for Sampling Colorings". In: *Journal of Mathematical Physics* 41.3 (2000), pp. 1555–1569.