THE UNIVERSITY OF MELBOURNE

DOCTORAL THESIS

The Coupling Time for the Ising Heat-Bath Dynamics & Efficient Optimization for Statistical Inference

Author:
Timothy HYNDMAN

Supervisors:
Prof. Peter TAYLOR
Prof. Aurore DELAIGLE
Assoc. Prof. Tim GARONI

Submitted in total fulfilment of the requirements of the degree of Doctor of Philosophy

Operations Research
School of Mathematics and Statistics

November 2018

THE UNIVERSITY OF MELBOURNE

Abstract

Faculty of Science
School of Mathematics and Statistics

Doctor of Philosophy

The Coupling Time for the Ising Heat-Bath Dynamics & Efficient Optimization for Statistical Inference

by Timothy Hyndman

The title page must be followed by an abstract of 300–500 words in English. The Thesis Abstract is written here (and usually kept to just this page). The page is kept centered vertically so can expand into the blank space above the title too.

Declaration of Authorship

This is to certify that:

1. the thesis comprises only my original work towards the PhD except where indicated in the Preface,
2. due acknowledgement has been made in the text to all other material used,
3. the thesis is fewer than 100 000 words in length, exclusive of tables, maps, bibliographies and appendices.

Signed: ______________________________

Date: ______________________________

Preface

If applicable, a Preface page includes a statement of:

- Work carried out in collaboration indicating the nature and proportion of the contribution of others and in general terms the portions of the work which the candidate claims as original
- Work submitted for other qualifications
- Work carried out prior to PhD candidature enrolment
- any third party editorial assistance, either paid or voluntary (as limited to the Editing of Research Theses by Professional Editors guidelines) and/or
- Where a substantially unchanged multi-author paper is included in the thesis a statement prepared by the candidate explaining the contributions of all involved. A signed copy by all authors must be included with the submission form.

"Thanks to my solid academic training, today I can write hundreds of words on virtually any topic without possessing a shred of information, which is how I got a good job in journalism."

Dave Barry

Acknowledgements

The acknowledgements and the people to thank go here, don't forget to include your project advisor…

Contents

List of Figures

For/Dedicated to/To my…

Chapter 1

Introduction to this Thesis

Initially, this thesis was intended to be made up entirely of the contents of Part II, along with what we hoped would be several significant further contributions to the study. However, the practicalities of a deadline, along with the challenging nature of the research, meant that the decision was made to augment this thesis with an essentially separate section of study. This is what makes up Part I.

The reader should view these two parts as standalone topics, to be read independently. However, they are not without any commonality. Both are within the realm of stochastic mathematics, Part I being a study of a random variable constructed from a stochastic process, and Part II being a study of probability distributions that maximize certain statistical objective functions.

Part I

The Coupling Time for the Ising Heat-Bath Dynamics

Chapter 2

Introduction

2.1 The Ising model

The Ising model is named after Ernst Ising who studied it in his 1924 thesis [1] under the supervision of Wilhelm Lenz, who introduced the model in [2]. It was originally motivated by the phenomenon of ferromagnetism but it has since found application to apply to numerous other situations in both physics and other fields [1].

The Ising model occupies a prominent position in the statistical physics literature. This is largely due to the existence of a phase transition; a sharp transition in the large scale behaviour of the model as a parameter crosses a critical value. The transition was first shown to exist by Rudolph Peierls [4] in what was the first proof of the existence of a phase transition for any model with purely local interactions in statistical mechanics. Additionally, the Ising model is both relatively simple, and also mathematically tractable in some non-trivial cases [5]. These qualities are rare among models with a phase transition and so the Ising model has become somewhat of a staple for both studying phase transitions and testing new statistical mechanics techniques.

The model is a probability distribution on spin configurations - assignments of $+1$ and -1 spins to each vertex in a finite graph $G = (V, E)$. The set of all possible configurations is

$$\Omega = \{-1, +1\}^V \tag{2.1}$$

and for a particular configuration, $\sigma \in \Omega$, we refer to the spin of a particular vertex $i \in V$ as $\sigma[i]$. Each configuration has an associated energy, given by

$$H_{G,\beta,h}(\sigma) = -\beta \sum_{ij \in E} \sigma[i]\sigma[j] - h \sum_{i \in V} \sigma[i] \tag{2.2}$$

[1] See [3, notes of Section 1.4.2] for a list of references concerning this.

where $\beta \in [0, \infty)$ is the inverse temperature, and $h \in \mathbb{R}$ is the magnetic field.

The Gibbs measure is the distribution on Ω that characterises the Ising model and it is defined by

$$\pi_{G,\beta,h}(\sigma) \propto \exp(-H_{G,\beta,h}(\sigma)). \tag{2.3}$$

In everything that follows, we will be concerned only with the zero-field ($h = 0$) Ising model. This gives us the slightly simpler form for the Gibbs measure,

$$\pi_{G,\beta}(\sigma) \propto \exp\left(\beta \sum_{ij \in E} \sigma[i]\sigma[j]\right), \qquad \sigma \in \{-1, 1\}^V. \tag{2.4}$$

2.1.1 The phase transition

An in depth study of the Ising phase transition and its associated critical temperature will not be needed for this work. However, we will still wish to refer to it occasionally and so here we give a workable description of the phase transition on lattices.

Consider the Gibbs measure with zero-field (2.4) in the limits $\beta \downarrow 0$ and $\beta \uparrow \infty$. It is easy to see that in the former limit, the measure is uniform across all configurations and in the latter limit, the measure assigns all weight to the constant configurations $\sigma^- = (-1, -1, \dots, -1)$ and $\sigma^+ = (+1, +1, \dots, +1)$. This leads to the following overly simplistic description of the phase transition. It is an abrupt change in distribution that occurs as we increase the temperature; from distributions concentrated on states whose spins mostly agree, to distributions producing states which have roughly equal numbers of plus and minus spins.

To be slightly more concrete we define quantities called the magnetization and magnetization density. The *magnetization* on a volume $\Lambda \subseteq V$ is defined as

$$M_\Lambda(\sigma) = \sum_{i \in \Lambda} \sigma[i]. \tag{2.5}$$

Normalizing this gives the *magnetization density*, $M_\Lambda(\sigma)/|\Lambda|$. On the d-dimensional torus with side length L, $G(L) = (\mathbb{Z}/L\mathbb{Z})^d$, the quantity

$$m(\beta) = \lim_{L \to \infty} \mathbb{E}_\beta \left| \frac{M_{G(L)}(\sigma)}{|G(L)|} \right| \tag{2.6}$$

depends on the inverse temperature β. When $d = 1$, $m(\beta) = 0$ for any β and there is no phase transition. However, when $d > 1$, there exists some critical $\beta_c(d)$ such that $m(\beta) = 0$ for $\beta < \beta_c(d)$ and $m(\beta) > 0$ for $\beta > \beta_c(d)$ [3]. This $\beta_c(d)$ is the critical inverse temperature at which we observe a phase transition.

2.2 Coupling from the past

One of the central challenges regarding the Ising model is how to efficiently sample from the Gibbs measure. Calculating the normalizing constant for (2.4), known as the partition function, is a #P-complete problem [6]. As such a direct approach to sampling is expected to be computationally intractable in general, and so other methods must be employed instead. One such method is Markov Chain Monte Carlo (MCMC). This involves constructing a Markov chain whose states are elements of Ω and whose stationary distribution is given by (2.4). One can then obtain a sample by running this Markov chain for long enough that the output has distribution sufficiently close to (2.4). One difficulty in using MCMC is that one does not know a priori what constitutes "long enough". In principal, bounds on this time can be obtained, but in practise, proving these bounds can be very challenging.

An alternative to classical MCMC called Coupling from the Past (CFTP) was introduced by Propp and Wilson [7]. Unlike MCMC, CFTP not only has an automatically determined running time, but it has the additional advantage of outputting exact samples from the stationary distribution. This does not come without a cost - CFTP has a random running time. Therefore, a key question towards evaluating the effectiveness of CFTP is understanding the distribution of its running time, that is, the *coupling time*.

In Chapters 3 and 4, we will investigate the coupling time for the Ising heat-bath Glauber dynamics, both on the cycle, at any temperature, in Chapter 3, and on any vertex transitive graph, at sufficiently high temperatures, in Chapter 4. Our main result in each chapter will be proving that, when appropriately scaled, the distribution of the coupling time essentially converges to a Gumbel as the size of the graph increases.

Prior to this work, not much has been written about the coupling time for the heat-bath dynamics. In [8, Conjecture 7.1], the authors conjecture that on the d-dimensional lattice with side length L, $\mathbb{Z}_L^d$, the coupling time of the Ising heat-bath process converges to a Gumbel distribution as $L \to \infty$ at all temperatures above the critical temperature. This is then followed by numerical evidence supporting their conjecture. We may also use the results by Propp and Wilson in [7, Section 5] to relate the *mixing time* of the heat-bath Glauber dynamics with tail bounds of the coupling time.

Given a parameter ϵ, a Markov Chain Y_t has mixing time

$$t_{\text{MIX}}(\epsilon) = \inf \{t : d(t) \le \epsilon\} \tag{2.7}$$

where

$$d(t) = \max_{y_0 \in \Omega} ||\mathbb{P}(Y_t \in \cdot | Y_0 = y_0) - \pi||_{\text{TV}} \tag{2.8}$$

and where the total variation distance between two distributions ν_1 and ν_2 is defined as

$$||\nu_1 - \nu_2||_{\text{TV}} = \max_{A \in \Omega} |\nu_1(A) - \nu_2(A)| = \frac{1}{2} \sum_{\sigma \in \Omega} |\nu_1(\sigma) - \nu_2(\sigma)|. \tag{2.9}$$

The results of [7, Theorem 5], within the setting of the discrete time Ising heat-bath process, state that the coupling time, T, satisfies

$$\frac{\mathbb{P}[T > k]}{n+1} \leq \bar{d}(k) \leq \mathbb{P}[T > k] \tag{2.10}$$

where

$$\bar{d}(k) = \max_{\mu_1, \mu_2} ||\mu_1^k - \mu_2^k||_{\text{TV}} \tag{2.11}$$

and μ^k is the distribution of the Markov chain at time k when started from a random state from distribution μ. The relationship between $\bar{d}$ and the mixing time is given by the result $d(t) \leq \bar{d}(t) \leq 2d(t)$ [9, Lemma 4.10].

On the complete graph, a complete characterisation of the mixing time as a function of temperature is obtained in [10]. On other graphs, the mixing time is treated in [11] for sufficiently high temperature. More recently, a series of papers by Lubetzky and Sly ([12], [13], [14], and [15]) have established much sharper results concerning the mixing time on a wide class of graphs. Their methods form a key part of our proof and will be discussed further in Section 2.3.

2.2.1 Ising heat-bath Glauber dynamics

The continuous-time heat-bath Glauber dynamics for the Ising model is a Markov chain whose states are elements of Ω and whose stationary distribution is given by (2.4). For a given graph $G = (V, E)$, and a given inverse temperature, β, we can describe the dynamics as follows.

Initialize every vertex in V with a spin (for example, we could start in the all-plus configuration). To each vertex in V we give an independent rate-one Poisson clock. For $\sigma \in \Omega$ and $i \in V$, define the probability

$$p_i(\sigma) = \frac{e^{\beta S_i(\sigma)}}{e^{\beta S_i(\sigma)} + e^{-\beta S_i(\sigma)}} \tag{2.12}$$

where

$$S_i(\sigma) = \sum_{j \sim i} \sigma[j] \tag{2.13}$$

is the sum of the spins of the neighbours of i, and $j \sim i$ denotes that j is connected to i with some edge $ij \in E$. Let σ_t denote the spin configuration at time t. When the clock of vertex i rings at some time t, we update $\sigma_t[i]$ to $+1$ with probability $p_i(\sigma_t)$, and to -1 otherwise.

The probability $p_i(\sigma)$ is constructed so that it gives the probability that vertex i is $+1$ if we sample it from π (2.4) conditioned on every other vertex having its spin fixed by σ. Note that this causes the dynamics to have π as its stationary distribution.

2.2.2 The coupling time

We now describe the two coupled chains from which we define the coupling time of the Ising heat-bath Glauber dynamics. It will prove convenient to first describe the discrete time chains along with their coupling and then discuss how to extend this coupling to the continuous time chain. In order to define the discrete time coupling, we introduce a random mapping representation.

Define $f : \Omega \times V \times [0, 1] \mapsto \Omega$ via $f(\sigma, i, u) = \sigma'$ where $\sigma'[j] = \sigma[j]$ for $j \neq i$ and

$$\sigma'[i] = \begin{cases} 1, & u \leq p_i(\sigma), \\ -1, & u > p_i(\sigma). \end{cases} \tag{2.14}$$

We note that f is monotonic, in the following sense. We define a partial ordering on Ω by writing that $\sigma \preceq \omega$ if $\sigma, \omega \in \Omega$ are such that $\sigma[i] \leq \omega[i]$ for all $i \in V$ (and similarly for $\sigma \succeq \omega$). Then for any fixed $i \in V$ and $u \in [0, 1]$, if $\sigma \preceq \omega$ then $f(\sigma, i, u) \preceq f(\omega, i, u)$.

Let $(\mathscr{V}_k, U_k)_{k\geq 1}$ be an i.i.d. sequence of copies of $(\mathscr{V}, U)$. Define top and bottom discrete time chains, $(\mathscr{T}_t)^{\text{DIS}}_{t\in\mathbb{N}}$ and $(\mathscr{B}_t)^{\text{DIS}}_{t\in\mathbb{N}}$, with initial states

$$\mathscr{T}_0^{\text{DIS}} = (1, 1, \dots, 1) \tag{2.15}$$

$$\mathscr{B}_0^{\text{DIS}} = (-1, -1, \dots, -1) \tag{2.16}$$

that update according to $\mathscr{T}_{t+1}^{\text{DIS}} = f(\mathscr{T}_t^{\text{DIS}}, \mathscr{V}_k, U_k)$ and $\mathscr{B}_{t+1}^{\text{DIS}} = f(\mathscr{B}_t^{\text{DIS}}, \mathscr{V}_k, U_k)$.

We call the coupled process, $(\mathscr{B}_t^{\text{DIS}}, \mathscr{T}_t^{\text{DIS}})_{t\in\mathbb{N}}$, *the discrete Ising heat-bath coupling.* From the monotonicity of f, $\mathscr{T}_t^{\text{DIS}} \succeq \mathscr{B}_t^{\text{DIS}}$, for all $t \geq 0$.

There are two ways we can think about extending this process to our continuous-time chain. The first way is to "continuize it at rate n" [9]. To do this we use the discrete process as defined above but the time between each update is an independent exponential with rate n. That is, the continuous-time top and bottom chains are defined as

$(\mathscr{B}_t, \mathscr{T}_t) = (\mathscr{B}^{\text{DIS}}_{N_t}, \mathscr{T}^{\text{DIS}}_{N_t})$ where N_t is an independent rate n Poisson process. We call $(\mathscr{B}_t, \mathscr{T}_t)_{t\geq 0}$ simply, *the Ising heat-bath coupling.*

It is perhaps not immediately obvious that the continuous time top and bottom chains have the same dynamics we described in Section 2.2.1. This leads us to the second way of extending the discrete coupling to continuous time. Instead of updating the whole chain at rate n and choosing a vertex to update on the kth update via $\mathcal{V}_k$, we can think of each vertex in the chain as having its own independent rate 1 Poisson clock that tells it when to update. To clarify, whenever the Poisson clock of any vertex i rings at time t, we perform the kth update of the chain by setting $\mathcal{V}_k = i$ and updating as in the discrete case via $\mathscr{B}_t \leftarrow f(\mathscr{B}_t, \mathcal{V}_k, U_k)$ and $\mathscr{T}_t \leftarrow f(\mathscr{T}_t, \mathcal{V}_k, U_k)$.

From the memoryless property of the exponential, the sequence $\mathcal{V}_1, \mathcal{V}_2, \ldots$ that is generated is i.i.d. uniform on V. Since we have n vertices updating at rate 1, the whole chain is updating at rate n, and so our two methods of extending the discrete coupling to continuous time are equivalent.

This leads to a more descriptive explanation of the continuous time coupling: the top and bottom chains share the same rate-one Poisson clocks at each vertex, and upon updating that vertex, we share the same uniform random variable U between the two chains to determine whether to update to a plus or minus according to (2.14).

The *coupling time* of the Ising heat-bath process is the random variable

$$T = \inf\{t : \mathscr{T}_t = \mathscr{B}_t\}. \tag{2.17}$$

This is the main object of interest for our analysis. Note that the coupling time is not just a property of the Ising heat-bath process, but also of the coupling we have chosen. In Section 3.1 we will make a change to the coupling we use to make the analysis easier. Some care will need to be taken to verify that the coupling time is not affected by this change.

2.2.3 Equivalence of discrete and continuous coupling time

So far we have stated that the running time of CFTP has the same distribution as the coupling time. In fact, we have glossed over one important detail. Namely, CFTP is exclusively run in discrete time, and our coupling time is defined by the continuous time dynamics. Therefore, for our motivation to be reasonable, we would like to show some sort of equivalence between the distributions of the discrete and continuous coupling times. We do this via Proposition 2.1.

Proposition 2.1. *Let $(N_n)_{n\in\mathbb{N}}$ be a sequence of positive integer-valued random variables, and $(m_n)_{n\in\mathbb{N}}$ be a non-decreasing sequence of integers such that $N_n \geq m_n$ for all n and $\lim_{n\to\infty} m_n = \infty$. Let $T(n)$ be the random time it takes for a rate λ Poisson clock to go off n times. That is, $T(n) \sim \mathrm{Erlang}(n, \lambda)$.*

Let a_n and b_n be positive deterministic sequences such that $b_n/a_n \to \infty$ and

$$\frac{b_n^2}{a_n^2} \log \frac{b_n}{a_n} = o(m_n). \tag{2.18}$$

Define

$$Y_n = \frac{T(N_n) - b_n}{a_n} \tag{2.19}$$

and

$$Z_n = \frac{N_n - \lambda b_n}{\lambda a_n}. \tag{2.20}$$

Let X be a random variable with continuous distribution function. Then $Y_n \xrightarrow{d} X$ if and only if $Z_n \xrightarrow{d} X$.

To prove Proposition 2.1 we first require the following Lemma.

Lemma 2.2. *Let $T(k)$ be the sum of k i.i.d. rate λ exponentials. For all $\epsilon > 0$,*

$$\mathbb{P}\left(\left|\frac{T(k)\lambda}{k} - 1\right| \geq \epsilon\right) \leq 2\exp\left(-k\epsilon^2/4\right) \tag{2.21}$$

Proof. For all $\epsilon > 0$,

$$\mathbb{P}\left(\left|\frac{T(k)\lambda}{k} - 1\right| \geq \epsilon\right) = \mathbb{P}\left(\frac{T(k)\lambda}{k} \leq 1 - \epsilon\right) + \mathbb{P}\left(\frac{T(k)\lambda}{k} \geq 1 + \epsilon\right). \tag{2.22}$$

Since $T(k)$ is the sum of k i.i.d. rate λ exponentials, its moment generating function is

$$M_k(t) = \left(\frac{\lambda}{\lambda - t}\right)^k, \qquad t < \lambda, \tag{2.23}$$

(see [16, Example 21.3]). Using a Chernoff bound, for all $0 < t < \lambda$, $\epsilon > 0$,

$$\begin{aligned}
\mathbb{P}\left(\frac{T(k)\lambda}{k} \geq 1 + \epsilon\right) &= \mathbb{P}\left(T(k) \geq \frac{k}{\lambda}(1+\epsilon)\right) &(2.24)\\
&\leq \left(\frac{\lambda}{\lambda - t}\right)^k \exp\left(-\frac{tk}{\lambda}(1+\epsilon)\right) &(2.25)\\
&= \exp\left(k\left(\ln(\lambda/(\lambda - t)) - t(1+\epsilon)/\lambda\right)\right). &(2.26)
\end{aligned}$$

Taking $t = \epsilon\lambda/(1+\epsilon)$, which for any $\epsilon > 0$ satisfies $t \in (0, \lambda)$ as required, we have that for all $\epsilon > 0$,

$$\mathbb{P}\left(\frac{T(k)\lambda}{k} \geq 1+\epsilon\right) \leq \exp\left(k(\log(1+\epsilon)-\epsilon)\right). \tag{2.27}$$

Similarly, for all $t < 0$, $\epsilon > 0$,

$$\mathbb{P}\left(\frac{T(k)\lambda}{k} \leq 1-\epsilon\right) = \mathbb{P}\left(T(k) \leq \frac{k}{\lambda}(1-\epsilon)\right) \tag{2.28}$$

$$\leq \left(\frac{\lambda}{\lambda-t}\right)^k \exp\left(-\frac{tk}{\lambda}(1-\epsilon)\right) \tag{2.29}$$

$$= \exp\left(k\left(\ln(\lambda/(\lambda-t)) - t(1-\epsilon)/\lambda\right)\right). \tag{2.30}$$

Since $T(k) > 0$ almost surely we have

$$\mathbb{P}\left(\frac{T(k)\lambda}{k} \leq 1-\epsilon\right) = 0 \tag{2.31}$$

when $\epsilon \geq 1$. Conversely, suppose $0 < \epsilon < 1$ and take $t = -\epsilon\lambda/(1-\epsilon) < 0$. Then

$$\mathbb{P}\left(\frac{T(k)\lambda}{k} \leq 1-\epsilon\right) \leq \exp\left(k(\log(1-\epsilon)+\epsilon)\right), \tag{2.32}$$

$$\leq \exp(k(\log(1+\epsilon)-\epsilon)). \tag{2.33}$$

Since $\log(1+\epsilon)-\epsilon$ is well defined for all $\epsilon > 0$ we then have, for any $\epsilon > 0$,

$$\mathbb{P}\left(\frac{T(k)\lambda}{k} \leq 1-\epsilon\right) \leq \exp(k(\log(1+\epsilon)-\epsilon)). \tag{2.34}$$

Overall,

$$\mathbb{P}\left(\left|\frac{T(k)\lambda}{k} - 1\right| \geq \epsilon\right) \leq 2\exp\left(k(\log(1+\epsilon)-\epsilon)\right) \tag{2.35}$$

$$\leq 2\exp\left(-k\epsilon^2/4\right) \tag{2.36}$$

for all $\epsilon > 0$. □

We now prove the main proposition.

Proof of Proposition 2.1. In order to prove either direction, it is sufficient to show (see [16, Theorem 25.4]) that for any $\epsilon > 0$,

$$\lim_{n\to\infty} \mathbb{P}(|Y_n - Z_n| > \epsilon) = 0. \tag{2.37}$$

First note that

$$|Y_n - Z_n| = \left|\frac{T(N_n) - b_n}{a_n} - \frac{N_n - \lambda b_n}{\lambda a_n}\right| \tag{2.38}$$

$$= \left|\frac{T(N_n)}{a_n} - \frac{N_n}{\lambda a_n}\right| \tag{2.39}$$

$$= \left|\frac{T(N_n)\lambda}{N_n} - 1\right| \frac{N_n}{\lambda a_n}. \tag{2.40}$$

So for any $\epsilon > 0$

$$\mathbb{P}(|Y_n - Z_n| > \epsilon) \leq \mathbb{P}\left(\left|\frac{T(N_n)\lambda}{N_n} - 1\right| > \epsilon\frac{a_n}{4b_n}\right) + \mathbb{P}\left(\frac{N_n}{\lambda a_n} > \frac{4b_n}{a_n}\right). \tag{2.41}$$

We will show that both of the terms on the right hand side vanish as $n \to \infty$. We start with the first of these.

Since $N_n \geq m_n$,

$$\mathbb{P}\left(\left|\frac{T(N_n)\lambda}{N_n} - 1\right| > \epsilon\frac{a_n}{4b_n}\right) \leq \mathbb{P}\left(\sup_{k \geq m_n}\left\{\left|\frac{T(k)\lambda}{k} - 1\right| > \epsilon\frac{a_n}{4b_n}\right\}\right) \tag{2.42}$$

$$= \mathbb{P}\left(\bigcup_{k \geq m_n}\left\{\left|\frac{T(k)\lambda}{k} - 1\right| > \epsilon\frac{a_n}{Cb_n}\right\}\right) \tag{2.43}$$

$$\leq \sum_{k=m_n}^{\infty} \mathbb{P}\left(\left|\frac{T(k)\lambda}{k} - 1\right| > \epsilon\frac{a_n}{4b_n}\right). \tag{2.44}$$

To apply Lemma 2.2, we need that $\epsilon a_n/(4b_n) < 1$. However, since $a_n/b_n \to 0$, we can ensure this holds by taking n large enough. Continuing,

$$\sum_{k=m_n}^{\infty} \mathbb{P}\left(\left|\frac{T(k)\lambda}{k} - 1\right| > \epsilon\frac{a_n}{4b_n}\right) \leq 2\sum_{k=m_n}^{\infty} \exp\left(-k\epsilon^2\frac{a_n^2}{64b_n^2}\right) \tag{2.45}$$

$$= 2\frac{\exp\left(-\epsilon^2 a_n^2(m_n - 1)/\left(64b_n^2\right)\right)}{\exp\left(\epsilon^2 a_n^2/\left(64b_n^2\right)\right) - 1}. \tag{2.46}$$

Since $x \leq \exp(x) - 1$ for $x \geq 0$, for sufficiently large n,

$$2\frac{\exp\left(-\epsilon^2 a_n^2(m_n - 1)/\left(64b_n^2\right)\right)}{\exp\left(\epsilon^2 a_n^2/\left(64b_n^2\right)\right) - 1} \leq \frac{128b_n^2}{a_n^2\epsilon^2}\exp\left(-\epsilon^2 a_n^2(m_n - 1)/\left(64b_n^2\right)\right) \tag{2.47}$$

$$\leq 256\frac{b_n^2}{a_n^2\epsilon^2}\exp\left(-m_n\epsilon^2 a_n^2/\left(64b_n^2\right)\right) \tag{2.48}$$

By (2.18), this goes to zero as $n \to \infty$.

To bound the second term in (2.41), we will treat the two directions of the proof separately. Firstly, assume that $Z_n \xrightarrow{d} X$. Then note that

$$\mathbb{P}\left(\frac{N_n}{\lambda a_n} > \frac{4b_n}{a_n}\right) = \mathbb{P}\left(Z_n > 3\frac{b_n}{a_n}\right) \tag{2.49}$$

and since $b_n/a_n \to \infty$, and X has a continuous distribution function,

$$\lim_{n\to\infty} \mathbb{P}\left(\frac{N_n}{\lambda a_n} > \frac{4b_n}{a_n}\right) = 0, \tag{2.50}$$

(see [16, Theorem 14.2, Lemma 2]) and so (2.37) holds.

Conversely, assume that $Y_n \xrightarrow{d} X$. Note that, if $T(N_n)/a_n \leq c_n/2$ and $|T(N_n)\lambda/N_n - 1| \leq 1/2$, then $N_n/(\lambda a_n) \leq c_n$. So taking $c_n = 4b_n/a_n$ we have for any $c_n > 0$,

$$\mathbb{P}\left(\frac{N_n}{\lambda a_n} > \frac{4b_n}{a_n}\right) \leq \mathbb{P}\left(\frac{T(N_n)}{a_n} > \frac{2b_n}{a_n}\right) + \mathbb{P}\left(\left|\frac{T(N_n)\lambda}{N_n} - 1\right| > \frac{1}{2}\right) \tag{2.51}$$

$$= \mathbb{P}\left(Y_n > \frac{b_n}{a_n}\right) + \mathbb{P}\left(\left|\frac{T(N_n)\lambda}{N_n} - 1\right| > \frac{1}{2}\right). \tag{2.52}$$

As above, since $b_n/a_n \to \infty$, and X has a continuous distribution function, the first term vanishes as $n \to \infty$. The second disappears since

$$\mathbb{P}\left(\left|\frac{T(N_n)\lambda}{N_n} - 1\right| > \frac{1}{2}\right) \leq \mathbb{P}\left(\left|\frac{T(N_n)\lambda}{N_n} - 1\right| > \epsilon\frac{a_n}{4b_n}\right) \tag{2.53}$$

for sufficiently large n. □

Remark 2.3. We apply Claim 2.1 to the coupling time of the Glauber heat-bath dynamics in the following way. Take N_n to be the discrete coupling time on a graph of size n. The continuous time coupling time is given by $T(N_n)$. Note that $N_n \geq m_n = n$ since each vertex must be updated at least once for coupling to occur. Finally Theorems 3.1 and 4.2 establish the limiting distribution of the continuous-time coupling time using scaling and shifting sequences a_n and b_n whose ratio is

$$\frac{b_n}{a_n} = \log n \tag{2.54}$$

and thus (2.18) is satisfied. This means that, appropriately scaled, the discrete-time coupling time has the same limiting distribution as the continuous-time coupling time.

2.2.4 Summary of CFTP

We are now in a position to give a brief summary of the CFTP method, as it applies to the Ising heat-bath coupling. It should be noted that we include this summary of

CFTP for completeness. None of the details regarding the implementation of CFTP are required outside of this section. It serves only as motivation for the study of the coupling time.

Let $f : \Omega \times V \times [0,1] \mapsto \Omega$ and $(\mathcal{V}, U)$ be as defined in Section 2.2.2. Let $(\mathcal{V}_k, U_k)$ be an i.i.d. sequence of copies of $(\mathcal{V}, U)$ and define

$$f_{-k} = f(\cdot, \mathcal{V}_k, U_k). \tag{2.55}$$

We construct the composition

$$F_{-k} = f_0 \circ f_{-1} \circ \cdots \circ f_{-k+1} \tag{2.56}$$

and define the *backwards coupling time* to be

$$T_{\text{BACK}} = \min\{k \in \mathbb{N} : F_{-k}(\mathscr{B}_0) = F_{-k}(\mathscr{T}_0)\}. \tag{2.57}$$

The state $F_{-T_{\text{BACK}}}(\mathscr{B}_0) = F_{-T_{\text{BACK}}}(\mathscr{T}_0)$ is the output of the CFTP algorithm, and was shown by Propp and Wilson [7] to be an exact sample from the chain's stationary distribution. To gain some intuition as to why this is so, observe that by the monotonicity of f, if $F_{-k}(\mathscr{B}_0) = F_{-k}(\mathscr{T}_0)$, then $F_{-k}(\sigma) = F_{-k}(\mathscr{B}_0)$ for any $\sigma \in \Omega$. If we let σ_π be a random sample from the stationary distribution π, then $F_{-k}(\mathscr{B}_0) = F_{-k}(\mathscr{T}_0) = F_{-k}(\sigma_\pi)$ must also have distribution π, which in our case is given by (2.4).

If we reverse the composition to construct

$$F_k = f_k \circ f_{k-1} \circ \cdots \circ f_1 \tag{2.58}$$

we can define the usual discrete time coupling time as

$$T_{\text{DIS}} = \min\{k \in \mathbb{N} : F_k(\mathscr{B}_0) = F_k(\mathscr{T}_0)\}. \tag{2.59}$$

The forwards coupling time, T_{DIS}, has the same distribution as the backwards coupling time, T_{BACK} [7], although in general, $F_{T_{\text{DIS}}}(\mathscr{B}_0) = F_{T_{\text{DIS}}}(\mathscr{T}_0)$ does not have distribution (2.4).

In practise, one runs the CFTP algorithm by starting both the top and bottom chains from some point in the past to time zero. This is repeated for increasingly more distant times in the past until both chains agree at time 0. The sequence of times at which one restarts this process need not be $-1, -2, -3, \ldots$, rather, any monotonic natural sequence $a_1, a_2, \ldots$ can be used. See [9], [17], and [18] for further discussion.

2.3 Information percolation

A cornerstone to the proofs contained in Chapters 3 and 4 is the framework of information percolation, introduced by Lubetzky and Sly in [14]. In this paper, Lubetzky and Sly managed to achieve much sharper results, in much more generality, regarding the mixing time for the Glauber dynamics for the Ising model than had been achieved before. In this section we provide a brief summary of their results before laying out the basic framework, in the context of the Ising heat-bath dynamics, that will be required for Chapters 3 and 4.

2.3.1 Information percolation and cutoff for the stochastic Ising model

Cutoff is the central phenomenon of study in Lubetzky and Sly's 2016 paper titled, 'Information percolation and cutoff for the stochastic Ising model'. A family of Markov chains (Y_t) indexed by n is said to exhibit cutoff if

$$t_{\mathrm{MIX}}(\epsilon) = (1 + o(1))t_{\mathrm{MIX}}(\epsilon'), \tag{2.60}$$

for any fixed $0 < \epsilon, \epsilon' < 1$ (recall (2.7) for the definition of t_{MIX}). A *cutoff window* is a sequence $w_n = o(t_{\mathrm{MIX}}(1/4))$ where

$$t_{\mathrm{MIX}}(\epsilon) = t_{\mathrm{MIX}}(1 - \epsilon) + \mathcal{O}(w_n) \tag{2.61}$$

for any $0 < \epsilon < 1$.

Historically, proving cutoff has proven to be highly challenging. In a survey on the topic, Diaconis [19] wrote 'proof of a cutoff is a difficult, delicate affair, requiring detailed knowledge of the chain, such as all eigenvalues and eigenvectors'. It is therefore worth noting the significant gap between the strength of the results regarding cutoff achieved using information percolation, and those that existed previously.

Previous to [14], the best result known for general graphs was that cutoff occurs with a $\mathcal{O}(1)$ window in the simple case when $\beta = 0$ [20]. However, no results were known for $\beta > 0$, despite a conjecture by Peres in 2009 [9, Section 23.2] that cutoff occurs on any sequence of transitive graphs when the mixing time is of order $\log n$ (as one would expect when $\beta < c_0$ for some $c_0 > 0$ that depends on the sequence of graphs). On lattices, the first results to appear were due to Lubetzky and Sly in 2013 who established cutoff up to the critical temperature for dimensions $d \leq 2$ with a $\mathcal{O}(\log \log n)$ window [12].

Using information percolation, Lubetzky and Sly proved the existence of cutoff for the continous time Glauber dynamics for the Ising model with an $\mathcal{O}(1)$ window on $\mathbb{Z}^d$ for all temperatures up to the critical temperature. In a companion paper [15], they extended this result to include any graph with maximum degree d provided that $\beta < \kappa/d$ for some absolute constant κ. Recently, information percolation has also been used to establish cutoff for the Swendsen-Wang dynamics on the lattice [21], suggesting that the technique is effective on a broader class of problems than simply Glauber dynamics for Ising.

2.3.2 The framework

At its core, information percolation is a way of tracking how the dependencies of the final spins of the Glauber heat-bath dynamics percolate through the graph over time. These dependencies are traced backwards through time from some designated time t^* on the space-time slab $V \times [0, t^*]$ to create the update history (see Figure 2.1 for example). These histories are made in such a way so that, if for every $j \in V$ no path exists connecting (i, t^*) to $(j, 0)$, then the spin of i does not depend on the initial state (and thus at time t^* vertex i takes $+1$ and -1 spins with equal probability by symmetry). The main constructs used to create this history are the update sequence, and the update support function which we will now define.

2.3.2.1 The update sequence

Recalling our random mapping representation from Section 2.2.2, we can encode an update of our coupled process with the tuple $(\mathcal{V}, U, S)$, where S is the time of the update, $\mathcal{V}$ is the vertex that is updated, and U is the value of the uniform random variable that tells us whether $\mathcal{V}$ is a plus or minus according to (2.14). The *update sequence* along an interval $(t_0, t_1]$ is the set of these tuples with $t_0 < S \leq t_1$, ordered by S decreasing from t_1.

Let $(Y_t)_{t\geq 0}$ be a copy of the continuous-time heat-bath Glauber dynamics starting in some state $Y_0 \in \Omega$. So $Y_t = \mathscr{T}_t$ if $Y_0 = (1, 1, \dots, 1)$ and $Y_t = \mathscr{B}_t$ if $Y_0 = (-1, -1, \dots, -1)$. Given the state of Y at time t_0, Y_{t_0}, the update sequence along $(t_0, t_1]$ contains all the information we need to contruct Y_{t_1}. In particular, given the update sequence along the interval $(0, t_1]$, Y_{t_1} is a deterministic function of Y_0.

2.3.2.2 The update support function

Given the update sequence along the interval $(t_1, t_2]$, the *update support function*, $\mathscr{F}(A, t_1, t_2)$, is the minimal set of vertices whose spins at time t_1 determine the spins of the vertices in A at time t_2. That is, $i \in \mathscr{F}(A, t_1, t_2)$ if and only if there exist states $Y_{t_1}, Y'_{t_1} \in \{-1, +1\}^V$

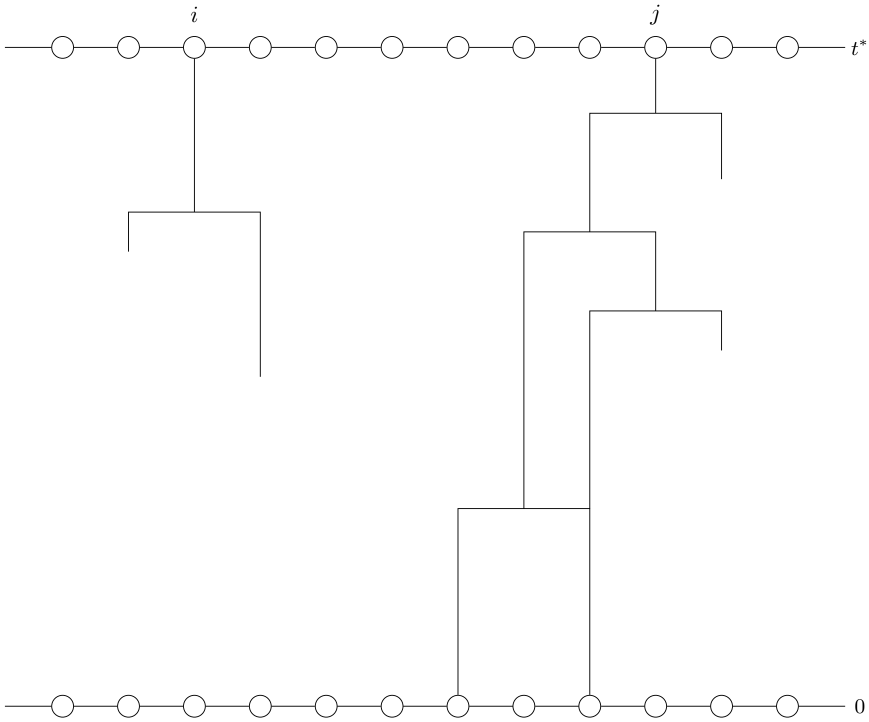


Figure 2.1 – *A section of the space-time slab $V \times [0, t^*]$ along with a typical appearance of the update histories for two vertices on the cycle. Time runs vertically from bottom to top, and the vertices are represented by circles, laid out horizontally. If there is a path in the update history of v between points (u, t) and (v, t^*), then the spin of v at time t^* depends on the spin of u at time t. In this example, since there is no path from vertex (i, t^*) to time 0, the final spin at i does not depend on the initial configuration whereas the final spin at j does.*

that differ only at i and such that when we construct Y_{t_2} and Y'_{t_2} using the update sequence, $Y_{t_2} \neq Y'_{t_2}$.

In particular, if $\mathscr{F}(i, 0, t) = \emptyset$ then the spin at vertex i at time t does not depend on the initial state and so for our coupled chains, $\mathscr{B}_t[i] = \mathscr{T}_t[i]$. As a consequence of the monotonicity of our coupling, we can make the stronger statement that $\mathscr{T}_t[i] = \mathscr{B}_t[i]$ if and only if $\mathscr{F}(i, 0, t) = \emptyset$ which of course means that

$$\mathbb{P}[\mathscr{T}_t[i] \neq \mathscr{B}_t[i]] = \mathbb{P}[\mathscr{F}(i, 0, t) \neq \emptyset]. \tag{2.62}$$

For ease of notation, we will often use the shorthand

$$\mathcal{H}_i(t) := \mathscr{F}(i, t, t^*) \tag{2.63}$$

where t^* is some target time that should be clear from context. We call this the *update support* of vertex i at time t. Tracing $\mathcal{H}_i(t)$ backwards in time from t^* produces a subgraph of $\Omega \times [0, t^*]$ which we write as $\mathcal{H}_i$ and which we call the *update history* of vertex i. To be slightly more precise, to produce $\mathcal{H}_i$ we connect (j, t) to (j, t') if $j \in \mathcal{H}_i(t)$

and there are no updates of j along $(t', t]$ and we connect (j, t) to (j', t) if there was an update at (j, t), $j \in \mathcal{H}_i(t)$, $j' \notin \mathcal{H}_i(t)$, and $j' \in \mathcal{H}_i(t + \epsilon)$ for all sufficiently small $\epsilon > 0$.

Similarly, we also use

$$\mathcal{H}_A(t) := \mathscr{F}(A, t, t^*) \tag{2.64}$$

for the update history of a vertex set A at time t and $\mathcal{H}_A$ for the update history of vertex set A. Note that

$$\mathcal{H}_A(t) = \bigcup_{i \in A} \mathcal{H}_i(t) \tag{2.65}$$

and

$$\mathcal{H}_A = \bigcup_{i \in A} \mathcal{H}_i. \tag{2.66}$$

2.3.2.3 The update function

It is usually non-trivial to construct the update support function from the update sequence. So in order to give some intuition to the definitions above, we describe another function which contains the update support and which is simple to construct. We define the *update function*, $\mathscr{F}_{\mathrm{UPD}}(A, t_1, t_2)$, to be the set of all vertices that A can 'reach' through the update function. That is, $i \in \mathscr{F}_{\mathrm{UPD}}(A, t_1, t_2)$ if and only if there exists a subsequence of the updates, $(\mathcal{V}_k, U_k, S_k)$, such that $t_1 < S_1 < S_2, \ldots, \leq S_m$ and $i, \mathcal{V}_1, \mathcal{V}_2, \ldots, \mathcal{V}_m$ is a path connecting i to some vertex $\mathcal{V}_m \in A$.

Just as we traced the update support backwards through time to create the update history, we can also trace $\mathscr{F}_{\mathrm{UPD}}(i, t, t^*)$ backwards through time to create the analogous *update trace*, $\mathcal{G}_i$. It is clear that $\mathscr{F}(A, t_1, t_2) \subseteq \mathscr{F}_{\mathrm{UPD}}(A, t_1, t_2)$ since a vertex i can only affect the spins of A if there is a path of updates connecting it to A. Likewise we also have that $\mathcal{G}_i \subseteq \mathcal{H}_i$.

Consider how we can construct the update trace of a vertex i from some target time t^*. We have at our disposal the update sequence along $(0, t^*]$ which is placed in order of decreasing time. If vertex i does not appear in the update sequence then we create a temporal edge between (i, t^*) and $(i, 0)$ and our update history is complete. Otherwise, we create a temporal edge between (i, t^*) and (i, t_i) where t_i is the last time vertex i was updated. At this point we add spatial edges from (i, t_i) to (j, t_i) for each $j \sim i$. Then, we iterate this process for i and each of its neighbours starting at time t_i until every edge has reached time 0. In Figure 2.2 we have followed this procedure to show an example update trace for a single vertex on the cycle.

We turn now to discussing how the update history differs from the update trace. We first note that an update to vertex i removes it from the update support as we move

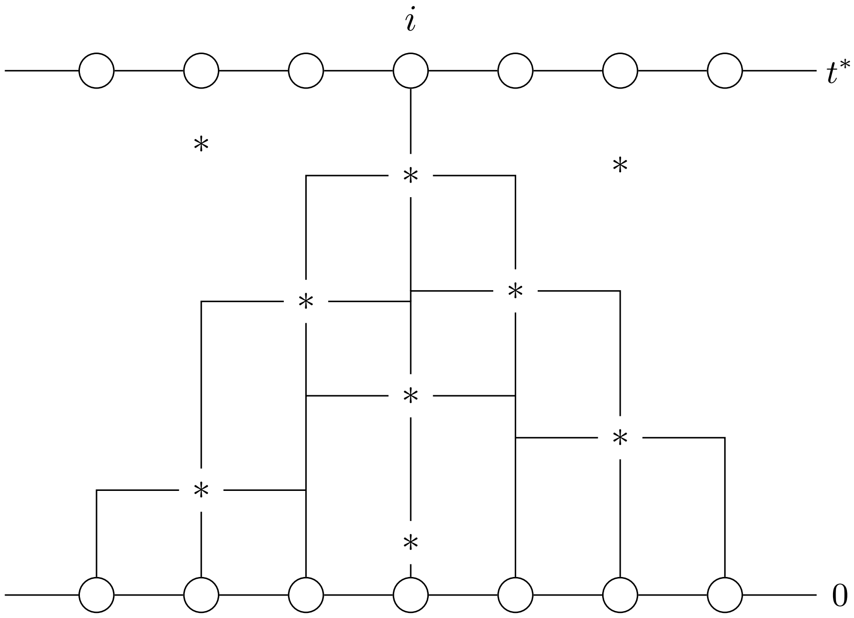


Figure 2.2 – *The update trace of i. Each update $(\mathcal{V}, U, t)$ in the update sequence is represented by a $*$ at $(\mathcal{V}, t)$.*

backwards in time. This is because the updated spin at i is a function only of its neighbours (2.14). The second difference which we will now spend some time discussing is that it is possible for updates to occur that do not depend on neighbouring spins. These updates therefore cause temporal edges leading up to them to terminate without branching out to the neighbouring vertices. These type of updates are called *oblivious updates*. (These are not the only differences between the update history and the update trace; there are other ways in which vertices can be removed from the update support. See Figure 2.4 for an example).

2.3.2.4 Oblivious updates

Roughly speaking, an update to a vertex is oblivious if we do not need to know the configuration of its neighbours to determine the spin of that vertex. More precisely, an update, $(\mathcal{V}, U, t)$, is oblivious if and only if

$$f(\sigma, \mathcal{V}, U)[\mathcal{V}] = f(\sigma', \mathcal{V}, U)[\mathcal{V}] \tag{2.67}$$

for all $\sigma, \sigma' \in \Omega$, where f is as defined in (2.14).

Consider how these updates occur under our random mapping representation. Let Δ_i denote the degree of a vertex i. Recalling (2.12),

$$\frac{e^{-\beta\Delta_i}}{e^{\beta\Delta_i} + e^{-\beta\Delta_i}} \leq p_i(\sigma) \leq \frac{e^{\beta\Delta_i}}{e^{\beta\Delta_i} + e^{-\beta\Delta_i}}, \tag{2.68}$$

with equality holding for the lower and upper limits when the neighbours have spins all minus and all plus respectively. So for a particular update $(\mathcal{V}, U, t)$, if $U \leq \frac{e^{-\beta\Delta_{\mathcal{V}}}}{e^{\beta\Delta_{\mathcal{V}}}+e^{-\beta\Delta_{\mathcal{V}}}}$ then $\mathcal{V}$ is updated to a plus regardless of the configuration of its neighbours. Hence $(\mathcal{V}, U, t)$ is an oblivious update. Similarly, if $U > \frac{e^{\beta\Delta_{\mathcal{V}}}}{e^{\beta\Delta_{\mathcal{V}}}+e^{-\beta\Delta_{\mathcal{V}}}}$ then $\mathcal{V}$ is updated to a minus regardless of the configuration of its neighbours and hence $(\mathcal{V}, U, t)$ is an oblivious update. It is easy to see that these are the only types of oblivious updates.

Given an update at vertex i, the probability that this update is oblivious is

$$\theta_i = 1 - \left(\frac{e^{\beta\Delta_i}}{e^{\beta\Delta_i}+e^{-\beta\Delta_i}} - \frac{e^{-\beta\Delta_i}}{e^{\beta\Delta_i}+e^{-\beta\Delta_i}}\right) \tag{2.69}$$

$$= 1 - \tanh(\beta\Delta_i). \tag{2.70}$$

If G is a Δ-regular graph (as will be the case in the following chapters) then we can drop the subscript and write $\theta = 1 - \tanh(\beta\Delta)$ for the probability of an oblivious update at each vertex.

As noted earlier, oblivious updates cause temporal edges leading to them in the update history to terminate. If $j \in \mathcal{H}_i(t)$, then an oblivious update (j, u, t) removes j from $\mathcal{H}_i(t)$ without adding any of its neighbours. In Figure 2.3 we construct the update history from a single vertex i using the same update sequence as in Figure 2.2 but instead of representing each update with just a $*$, we give a little more information in the following way. Note that on the cycle, the function defined in (2.14) can be rewritten as

$$\sigma'[i] = \begin{cases} 1 & U \leq \theta/2, \\ \sigma[i-1] \vee \sigma[i+1] & \theta/2 < U \leq 1/2, \\ \sigma[i-1] \wedge \sigma[i+1] & 1/2 < U \leq 1-\theta/2, \\ -1 & U > \theta/2. \end{cases} \tag{2.71}$$

We can therefore represent each update $(\mathcal{V}, U, t)$ in the update sequence by placing at $(\mathcal{V}, t)$ one of the symbols $+$, $\vee$, $\wedge$, or $-$ choosen according to U. We then trace back from time t^*, branching to either side when we encounter a $\vee$ or $\wedge$, and terminating whenever we encounter a $+$ or $-$.

It is worth remarking that oblivious updates are not necessarily the only updates that can shrink the size of the update history of i. In Figure 2.4 we use an example from [14] that shows the update support collapsing down to a single vertex from a non-oblivious update. However, for our analysis, oblivious updates will be the only such updates we will be concerned with. Indeed, in Chapter 3 we will use a different coupling so that these are the only updates that shrink the size of the update history, and in Chapter 4

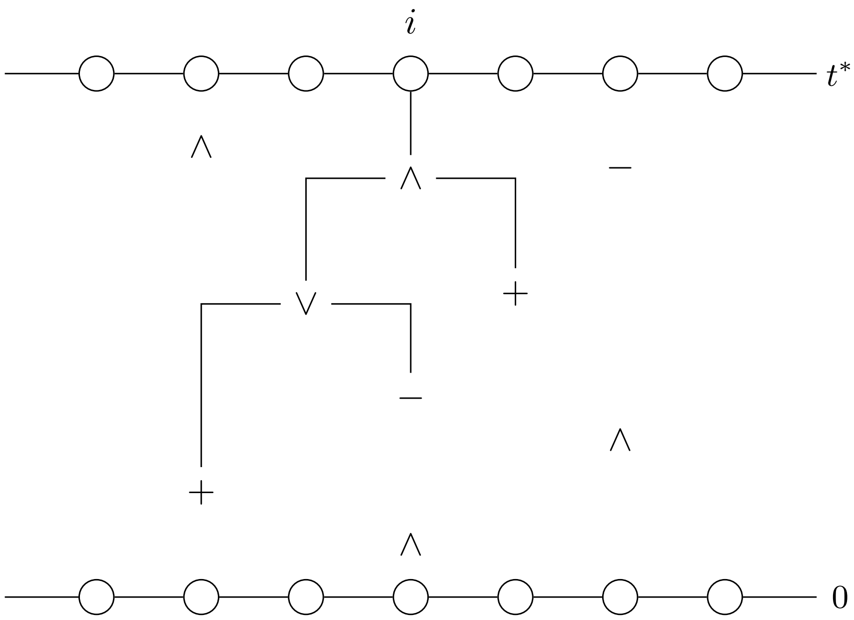


Figure 2.3 – *The update sequence for a section of the cycle and the corresponding update history from vertex i. For this particular update sequence, i takes a final spin of $+1$ regardless of the initial configuration.*

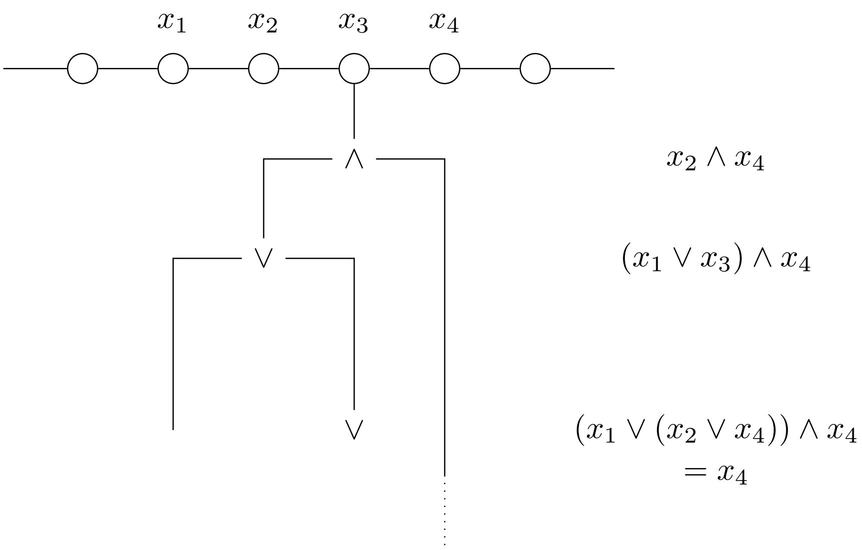


Figure 2.4 – *[Example taken from [14]]. A non-oblivious update that shrinks the size of the update history. On the right is written the final spin of x_3 as a function of the configuration at that time. The update $x_3 \mapsto x_2 \vee x_4$ causes the entire function to collapse to x_4, and so removes x_1 and x_3 from the update history.*

we will use an alternative construction that bounds the true update history, in which all updates are either oblivious or branch out to all Δ neighbours.

Chapter 3

The Coupling Time on the Cycle

In this chapter we consider the Ising heat-bath Glauber dynamics (as described in Section 2.2.1) on the cycle $G_n = (\mathbb{Z}/n\mathbb{Z})$. The object of interest is the coupling time, T_n, which was defined in Section 2.2.2. To simplify the analysis we study another random variable, defined in Section 3.1, which has the same distribution as T_n. The main result is Theorem 3.1 which establishes that T_n converges in distribution to a Gumbel distribution at all temperatures. This confirms, for $d = 1$, a conjecture by Collevecchio et al. that the coupling time of the Ising heat-bath process on the lattice $G_L = (\mathbb{Z}/L\mathbb{Z})^d$ converges to a Gumbel distribution as $L \to \infty$ for all $\beta < \beta_c$ [8, Conjecture 7.1] (We treat higher dimensions, and more generally any vertex transitive graphs, in Chapter 4). Of course, in one dimension, all temperatures are part of the high temperature regime [3], and our result holds for any inverse-temperature β.

There is some intuition behind why we might expect that the coupling time converges to a Gumbel distribution. It is based on the belief that when the temperature is in the high-temperature regime, the dynamics behave qualitatively as if $\beta = 0$. In the $\beta = 0$ case, the spins update independently of their neighbours, and thus the top and bottom chains can be coupled so that they agree on each vertex that has been updated. The coupling time is then precisely the time it takes for each vertex to be updated. This corresponds to the coupon collector's problem, which is known to have a Gumbel limit [22].

As mentioned in Section 2.2.4, the coupling time is of practical interest since its distribution is the same as that of the running time of the coupling from the past (CFTP) algorithm. Our result shows that when running the Glauber heat-bath dynamics for the Ising model on a large enough cycle, the running time of CFTP can be approximated by a Gumbel distribution. We note that even though one is typically more interested in the Ising model on lattices of dimension at least two (so that there exists a phase transition), the one dimensional case proves to be a useful test case for the proof techniques. Furthermore, the applicability of Theorem 3.1 to the full high temperature regime justifies a treatment separate to that of the higher dimensional case considered in Chapter 4, the proof of which holds only for sufficiently high temperatures.

Theorem 3.1. *Let T_n be the coupling time for the continuous-time Ising heat-bath Glauber dynamics for the zero-field ferromagnetic Ising model on the cycle $(\mathbb{Z}/n\mathbb{Z})$. Then for any inverse-temperature β, there exists a subsequence (T_m) of (T_n) such that*

$$\lim_{m\to\infty} \mathbb{P}\left[T_m < \frac{z+\ln m}{\theta}\right] = e^{-C_\theta e^{-z}} \tag{3.1}$$

where $\theta = 1 - \tanh(2\beta)$ and C_θ is a positive constant satisfying

$$\frac{1}{2\sqrt{\frac{4}{\theta}-1}-1} \leq C_\theta \leq 1. \tag{3.2}$$

The proof of Theorem 3.1 will be given in Section 3.3 after the essential preliminaries are presented. In Section 3.1 we describe some modified dynamics and show that the coupling time we construct from these has the same distribution as the coupling time defined in Section 2.2.2. Then in Section 3.2 we outline the overall approach to the proof and define some essential quantities. Finally, Section 3.4 contains additional lemmas that are used in Section 3.3.

3.1 A new coupling on the cycle

On the cycle, we will use a different coupling of $\mathscr{T}_t$ and $\mathscr{B}_t$ via a new random mapping representation that will replace the update rule in (2.14). The new update rules simplify our update histories greatly by ensuring that each of the update histories never contain more than one vertex at any one time. However, we must be cautious. The coupling time is not just a property of the heat-bath dynamics, but also of the specific coupling we chose. Hence, we first verify that switching to our new rules does not change the distribution of T_n.

The new update rules are defined by using almost the same construction as in Section 2.2.2. The one difference is that we replace (2.14) as follows. When vertex i updates, instead of comparing U to the probability $p_i(\sigma)$ to determine the new spin, we instead

$\mathscr{T}_t = \cdot$ / $\mathscr{B}_t = \cdot$ \ $\mathbb{P}[(\mathscr{T}_t[i]', \mathscr{B}_t[i]') = \cdot\,]$	(1,1)	(1,-1)	(-1,-1)
$(\dots, 1, \mathscr{T}_t[i], 1, \dots)$ $(\dots, 1, \mathscr{B}_t[i], 1, \dots)$	$1-\theta$	0	$\frac{\theta}{2}$
$(\dots, 1, \mathscr{T}_t[i], 1, \dots)$ $(\dots, 1, \mathscr{B}_t[i], -1, \dots)$	$\frac{1}{2}$	$\frac{1-\theta}{2}$	$\frac{\theta}{2}$
$(\dots, 1, \mathscr{T}_t[i], 1, \dots)$ $(\dots, -1, \mathscr{B}_t[i], 1, \dots)$	$\frac{1}{2}$	$\frac{1-\theta}{2}$	$\frac{\theta}{2}$
$(\dots, 1, \mathscr{T}_t[i], 1, \dots)$ $(\dots, -1, \mathscr{B}_t[i], -1, \dots)$	$\frac{\theta}{2}$	$1-\theta$	$\frac{\theta}{2}$
$(\dots, 1, \mathscr{T}_t[i], -1, \dots)$ $(\dots, 1, \mathscr{B}_t[i], -1, \dots)$	$\frac{1}{2}$	0	$\frac{1}{2}$
$(\dots, 1, \mathscr{T}_t[i], -1, \dots)$ $(\dots, -1, \mathscr{B}_t[i], -1, \dots)$	$\frac{\theta}{2}$	$\frac{1-\theta}{2}$	$\frac{1}{2}$
$(\dots, -1, \mathscr{T}_t[i], 1, \dots)$ $(\dots, -1, \mathscr{B}_t[i], 1, \dots)$	$\frac{1}{2}$	0	$\frac{1}{2}$
$(\dots, -1, \mathscr{T}_t[i], 1, \dots)$ $(\dots, -1, \mathscr{B}_t[i], -1, \dots)$	$\frac{\theta}{2}$	$\frac{1-\theta}{2}$	$\frac{1}{2}$
$(\dots, -1, \mathscr{T}_t[i], -1, \dots)$ $(\dots, -1, \mathscr{B}_t[i], -1, \dots)$	$\frac{\theta}{2}$	0	$1-\theta$

Table 3.1 – *Probabilities of updating from $(\mathscr{T}_t, \mathscr{B}_t)$ to $(\mathscr{T}_t', \mathscr{B}_t')$ given vertex i updates at time t.*

chose a new spin σ_i' via

$$\sigma_i' = \begin{cases} +1 & U < \theta/2, \\ \sigma_{i-1} & \theta/2 \leq U < 1/2, \\ \sigma_{i+1} & 1/2 \leq U < 1-\theta/2, \\ -1 & U \geq 1-\theta/2. \end{cases} \tag{3.3}$$

where $U \in [0,1]$ is an independent uniform random variable as before. It is easy to see that these update rules give rise to the same transition rates as those in (2.14). To show that the coupling time is unchanged, it is sufficient to verify that the joint jump probabilities of $(\mathscr{T}_t[i], \mathscr{B}_t[i])$ are unchanged for each possible configuration of spins of vertices $i-1$ and $i+1$. There are only nine possible configurations for the two neighbours of i in the top and bottom chain since $\mathscr{B}_t[i] \leq \mathscr{T}_t[i], \forall t$. Likewise, there are only three possible configurations for the updated spins $(\mathscr{T}_t[i]', \mathscr{B}_t[i]')$. Hence, given vertex i updates at time t, we can easily calculate all the required jump probabilities as shown in Table 3.1. These are unchanged whether using (2.14) or (3.3) and so the new rules do not change the coupled dynamics.

[MAKE INTO LEMMA?]

3.1.1 Update histories on the cycle

Under the update rules in (3.3), each time a vertex is updated, it is either an oblivious update with probability θ, or it takes the spin of a uniformly chosen neighbour. Unlike the histories considered earlier (for example Figure 2.3), this time a non-oblivious update does not cause the history to branch out to both its neighbours. Rather, given a non-oblivious update to some vertex v, we only need to know the spins of one of its neighbours to update it (the left spin if $U < 1/2$ and the right if $U \geq 1/2$). So the history simply moves either right or left without branching. As before, encountering an oblivious update causes $\mathcal{H}_i$ to terminate.

Let us give an explicit construction for the update histories just described. Let N_t be the rate n Poisson process used to continuize the discrete process and let $(\mathcal{V}_k, U_k)_{k\geq 1}$ be the discrete noise sequence (as in Section 2.2.2). For each vertex $i \in V$ define the thinned processes,

$$K_t^i = \#\{k \leq N_t : \mathcal{V}_k = i, U_k \in [0, \theta/2) \cup (1-\theta/2, 1]\} \tag{3.4}$$

$$L_t^i = \#\{k \leq N_t : \mathcal{V}_k = i, U_k \in [\theta/2, 1/2)\} \tag{3.5}$$

$$R_t^i = \#\{k \leq N_t : \mathcal{V}_k = i, U_k \in [1/2, 1-\theta/2)\}. \tag{3.6}$$

The process K_t^i is Poisson with rate θ and gives the times when vertex i has an oblivious update. The processes L_t^i and R_t^i are Poisson with rate $(1-\theta)/2$ and give the times when σ_i is replaced by σ_{i-1} and σ_{i+1} respectively.

The collection of K_i^t, L_t^i and R_t^i for every $i \in V$ forms an encoding of the update sequence and may be represented graphically as follows. Place an $\times$ at every $(i, K_t^i) \in V \times [0, t^*]$. Draw a directed edge $(i, L_t^i) \to (v-1, L_t^i)$ for every L_t^i and draw a directed edge $(i, R_t^i) \to (v+1, R_t^i)$ for every R_t^i. To construct the update histories, trace back in time from t^* from each vertex, making turns along directed horizontal edges, and killing the process at any $\times$. An example of this graphical representation along with the update histories is shown in Figure 3.1.

["SHOULD PROBABLY COMMENT ON WHY, IE WHY THEY HAVE THE RIGHT DISTRIBUTION"]

[POINT OUT THIS IS SAME AS GRAPHICAL REPRESENTATION OF NOISY VOTER MODEL]

Following the history of a single vertex i backwards in time from t^*, we trace out a continuous-time random walk which moves left at rate $(1-\theta)/2$, and moves right at rate $(1-\theta)/2$. The walk survives until it encounters a point from K_t^j, $j = \mathcal{H}_i(t)$, at which

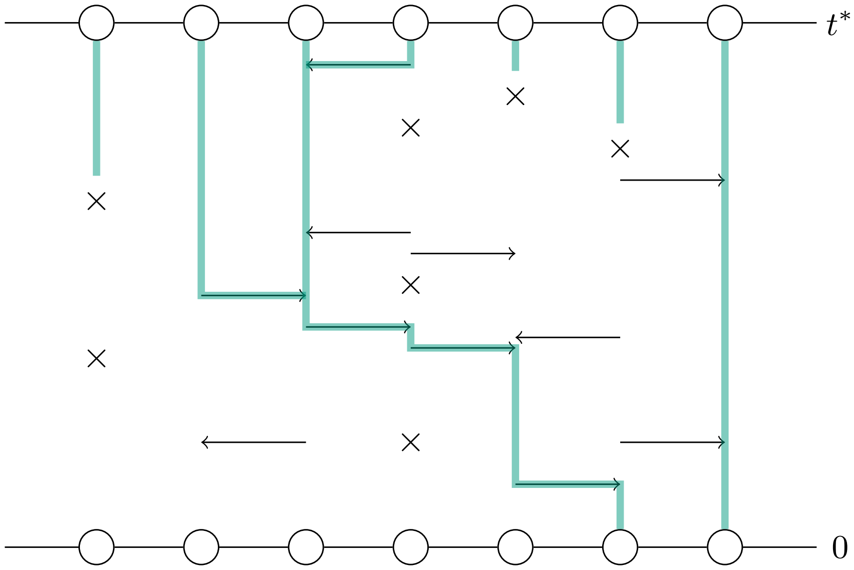


Figure 3.1 – *The update sequence for a section of the cycle represented by the thinned Poisson processes K_t^i ($\times$), L_t^i (left arrows), and R_t^i (right arrows) for each vertex. The corresponding update histories are overlaid in blue.*

point it terminates. From the memoryless property of the exponential waiting times of K_t^j, and since K_t^j is independent of K_t^k for any two vertices $j \neq k$, we have that the time until a single history dies is exponential with rate θ. This immediately gives us the following probability which we will use repeatedly in what follows. Recalling (2.62),

$$\mathbb{P}\left[\mathscr{B}_{t^*}[i] \neq \mathscr{T}_{t^*}[i]\right] = \mathbb{P}\left[\mathcal{H}_i(0) \neq \emptyset\right] = e^{-\theta t^*}. \tag{3.7}$$

3.2 Problem set-up

["MIGHT IT BE CLEARER TO HAVE A SECTION SOMEWHERE, MAYBE CH2 WHERE THE GENERAL COMPOUND POISSON RESULTS COULD BE STATED IN GENERAL? SINCE THIS STUFF IS BASICALLY IDENTICAL ON Z_n OR AN ARBITRARY GRAPH, IT MAY BE A LITTLE OUT OF PLACE HERE??"]

In order to prove Theorem 3.1, we will actually prove a stronger statement using Theorem 3.2. The general idea is that at some fixed time t^* we will count the number of vertices at which the bottom and top chains differ. This number is a random variable, which we will call W, and we can bound the total variation distance between its distribution and that of an appropriate compound Poisson distribution. As a special case, we can then bound the probability that W is zero. Of course, if W is zero then the top and bottom chains must have coupled and so we can use this to establish Theorem 3.1.

Bounding the total variation distance between W and the compound Poisson will be done using compound Poisson approximation as described in [23]. This paper reviews a number of different methods by which approximations may be made. The specific method that we will employ is based on Stein's method for the compound Poisson distribution, introduced in [24].

We now make precise the ideas stated above. Fix z and a time of interest,

$$t^* = (z + \ln n)/\theta. \tag{3.8}$$

For any fixed $z \in \mathbb{R}$, $t^* > 0$ for all sufficiently large n. We only consider such n in what follows. For each vertex $i \in V$, define the indicator

$$X_i = \begin{cases} 1 & \mathscr{B}_{t^*}[i] \neq \mathscr{T}_{t^*}[i], \\ 0 & \mathscr{B}_{t^*}[i] = \mathscr{T}_{t^*}[i] \end{cases} \tag{3.9}$$

and set $W = \sum_{i \in V} X_i$. Note that from (3.7) we get

$$\mathbb{P}[X_i = 1] = e^{-\theta t^*} = \frac{e^{-z}}{n}. \tag{3.10}$$

For each $i \in V$, decompose W into $W = X_i + U_i + Z_i + W_i$ where

$$U_i = \sum_{j \in B_i} X_j, \qquad Z_i = \sum_{j \in C_i} X_j, \qquad W_i = \sum_{j \in D_i} X_j. \tag{3.11}$$

and using $d(i,j)$ to denote the distance between vertices i and j on the cycle, B_i, C_i, and D_i are the vertex sets

$$B_i = \{j \neq i : d(i,j) \leq b_n\}, \tag{3.12}$$

$$C_i = \{j \notin B_i \cup \{i\} : d(i,j) \leq c_n\}, \tag{3.13}$$

$$D_i = V \setminus (B_i \cup C_i \cup \{i\}). \tag{3.14}$$

We have some freedom in choosing b_n and c_n; they are chosen to control the asymptotics of various quantities to be defined later. For this chapter, we will choose $b_n = \ln(n)$ and $c_n = \ln(n)^2$.

[GOING THROUGH TIM'S COMMENTS UP TO HERE]

We now define the quantities

$$\lambda = \sum_{i \in V} \mathbb{E}\left[\frac{X_i}{X_i + U_i} I[X_i + U_i \geq 1]\right], \tag{3.15}$$

$$\mu_l = \frac{1}{l\lambda} \sum_{i \in V} \mathbb{E}\left[X_i I[X_i + U_i = l]\right], \qquad l \geq 1, \tag{3.16}$$

which will be the parameters of the approximating compound Poisson distribution to W. We also define

$$\delta_1 = \sum_{i \in V} \sum_{k \geq 0} \mathbb{P}[X_i = 1, U_i = k] \mathbb{E}\left|\frac{\mathbb{P}[X_i = 1, U_i = k | W_i]}{\mathbb{P}[X_i = 1, U_i = k]} - 1\right|, \tag{3.17}$$

$$\delta_4 = \sum_{i \in V} \left(\mathbb{E}[X_i Z_i] + \mathbb{E}[X_i]\mathbb{E}[X_i + U_i + Z_i]\right), \tag{3.18}$$

which we desire to be small for the compound Poisson approximation to be good.

The following theorem (reworked from [23]) bounds the distance between the distributions of W and the approximating compound Poisson.

Theorem 3.2 ([23])**.** *Let W, λ, $\boldsymbol{\mu}$, δ_1 and δ_4 be as defined above. Then*

$$d_{\mathrm{TV}}(\mathscr{L}(W), \mathrm{CP}(\lambda, \boldsymbol{\mu})) \leq (\delta_1 + \delta_4)e^{\lambda}. \tag{3.19}$$

Note that W is zero precisely when $T < t^*$ and so the events $\{W = 0\}$ and $\{T \leq t^*\}$ are the same. Furthermore, $\boldsymbol{\mu}$ as defined above in (3.16), is supported only on the positive integers and so $\mathbb{P}[\mathrm{CP}(\lambda, \boldsymbol{\mu}) = 0] = e^{-\lambda}$. These observations lead to the following corollary of Theorem 3.2.

Corollary 3.3. *Let T_n be the coupling time of the continuous-time heat-bath Glauber dynamics for the zero-field Ising model at inverse-temperature β on the cycle $(\mathbb{Z}/n\mathbb{Z})$ and let δ_1, δ_4 and λ be as defined above. Then*

$$\left|\mathbb{P}\left[T_n \leq \frac{z + \ln(n)}{\theta}\right] - e^{-\lambda}\right| \leq (\delta_1 + \delta_4)e^{\lambda}, \tag{3.20}$$

where $\theta = 1 - \tanh(2\beta)$.

3.3 Proof of Theorem 3.1

In this section we use Corollary 3.3 to prove Theorem 3.1 by bounding λ and showing that δ_1 and δ_4 go to zero as $n \to \infty$. This is done in Lemmas 3.4, 3.6, and 3.7. The proofs of these require some additional lemmas concerning properties of the update histories which have been deferred to Section 3.4.

We begin by bounding λ. Note that bounding λ is enough to show that there is a subsequence of graphs on which λ converges as required by Theorem 3.1.

Lemma 3.4. *Using the above set-up*

$$\limsup_{n\to\infty} \lambda \leq e^{-z} \tag{3.21}$$

and

$$\liminf_{n\to\infty} \lambda \geq C_\theta e^{-z} \tag{3.22}$$

for some

$$C_\theta \in \left[\frac{1}{2\sqrt{\frac{4}{\theta}-1}-1}, 1\right]. \tag{3.23}$$

Proof. Beginning with the definition of λ, we have

$$\lambda = \sum_{i\in V} \mathbb{E}\left[\frac{X_i}{X_i+U_i} I[X_i+U_i \geq 1]\right] \tag{3.24}$$

$$= \sum_{i=1}^{n} \mathbb{P}(X_i = 1)\mathbb{E}\left[\frac{1}{1+U_i}|X_i = 1\right] \tag{3.25}$$

$$= \sum_{i=1}^{n} \frac{e^{-z}}{n}\mathbb{E}\left[\frac{1}{1+U_i}|X_i = 1\right] \tag{3.26}$$

$$= e^{-z}\mathbb{E}\left[\frac{1}{1+U_i}|X_i = 1\right] \tag{3.27}$$

where we have used that X_i is either zero or one, (3.10), and the transitivity of the graph in each step respectively. Clearly

$$\mathbb{E}\left[\frac{1}{1+U_i}|X_i = 1\right] \leq 1. \tag{3.28}$$

By Jensen's inequality

$$\mathbb{E}\left[\frac{1}{1+U_i}|X_i = 1\right] \geq \frac{1}{\mathbb{E}[1+U_i|X_i=1]} \tag{3.29}$$

$$= \frac{1}{1+\mathbb{E}[U_i|X_i=1]}. \tag{3.30}$$

Now

$$\mathbb{E}[U_i|X_i=1] = \sum_{j\in B_i} \mathbb{P}[X_j=1|X_i=1] \tag{3.31}$$
$$= \sum_{k=1}^{\lfloor b_n \rfloor} \sum_{|j-i|=k} \mathbb{P}[X_j=1|X_i=1] \tag{3.32}$$
$$= 2\sum_{k=1}^{\lfloor b_n \rfloor} \mathbb{P}[X_{i+k}=1|X_i=1] \tag{3.33}$$

where we have used the symmetry of X_{i+k} and X_{i-k} in the last step. From Lemma 3.10,

$$\mathbb{E}[U_i|X_i=1] \le 2\sum_{k=1}^{\lfloor b_n \rfloor} \left(\frac{e^{-z}}{n} + 2\left(\frac{2-\sqrt{\theta(4-\theta)}}{2-\theta}\right)^k \right) \tag{3.34}$$
$$< 2\sum_{k=1}^{\lfloor b_n \rfloor} \frac{e^{-z}}{n} + 4\sum_{k=1}^{\infty} \left(\frac{2-\sqrt{\theta(4-\theta)}}{2-\theta}\right)^k \tag{3.35}$$
$$= \frac{2\lfloor b_n \rfloor}{n} e^{-z} + 2\left(\sqrt{\frac{4}{\theta}-1}-1\right). \tag{3.36}$$

Finally, as $n \to \infty$ the first term vanishes and

$$\liminf_{n\to\infty} \mathbb{E}\left[\frac{1}{1+U_i}|X_i=1\right] \ge \frac{1}{2\sqrt{\frac{4}{\theta}-1}-1}. \tag{3.37}$$

□

To prove that λ converges we will show that U_i^n converges in distribution to some other random variable. This then allows us to use [WHAT??] to prove that λ converges. The proofs require the construction of some new random variables which we define now.

The update sequence on the cycle may be represented using the thinned processes K_t^i, L_t^i, and R_t^i. We may extend the update sequence into the interval $t<0$ by mirroring these processes backwards in time from $t=0$ [HOW DO I SAY THIS PROPERLY?]. We then can use this extended update sequence to define the update support and update histories on the interval $(-\infty, t^*]$. For each $i \in V$, we define the indicators

$$\bar{X}_i = \begin{cases} 1 & \forall t \in (-\infty, t^*], \mathcal{H}_i(t) \neq \emptyset, \\ 0 & \text{otherwise.} \end{cases} \tag{3.38}$$

and the random variables

$$\bar{U}_i = \sum_{j\in B_i} \bar{X}_j. \tag{3.39}$$

With probability 1, $\bar{X}_i = 0$ and $\bar{U}_i = 0$ since the update histories eventually encounter an oblivious update with probability 1.

The following lemma relates the distribution of U_i with that of $\bar{U}_i$.

Lemma 3.5.

$$\mathbb{P}(\bar{U} \geq k|\bar{X}_i = 1) \leq \mathbb{P}(U_i \geq k|X_i = 1) \leq \mathbb{P}(\bar{U}_i \geq k|\bar{X}_i = 1) + e^{-z}\frac{|B_i|}{n} \tag{3.40}$$

Proof. We start with the lower bound. We first note that conditioning on $X_i = 1$ and $\bar{X}_i = 1$ have the same result along $[0, t^*]$. [WRITE THIS PROPERLY]. So

$$\mathbb{P}(U_i \geq k|X_i = 1) = \mathbb{P}(U_i \geq k|\bar{X}_i = 1). \tag{3.41}$$

For every $j \in V$, if $\bar{X}_j = 1$, then $X_j = 1$. So $\bar{U}_i \leq U_i$ and

$$\mathbb{P}(U_i \geq k|X_i = 1) \geq \mathbb{P}(\bar{U}_i \geq k|\bar{X}_i = 1). \tag{3.42}$$

We now treat the upper bound. Define the event

$$A_j = \{\mathcal{H}_j(0) \neq \emptyset, \mathcal{H}_j \cap \mathcal{H}_i = \emptyset\}. \tag{3.43}$$

and their union

$$A = \bigcup_{j \in B_i} A_j. \tag{3.44}$$

We note that on the event $A^c \cap \{\bar{X}_i = 1\}$, $U_i = \bar{U}_i$. $\square$

The next two lemmas prove that δ_1 and δ_4 go to zero as $n \to \infty$. Since from Lemma 3.4 we know that λ is bounded above by a constant, this is enough for the right hand size of (3.20) to go to zero as $n \to \infty$.

Lemma 3.6. *Let δ_1 be as defined above in (3.17). Then*

$$\lim_{n \to \infty} \delta_1 = 0. \tag{3.45}$$

Proof. Starting with the definition of δ_1, we have

$$\delta_1 = \sum_{i=1}^{n} \sum_{k=0}^{2\lfloor b_n \rfloor} \mathbb{P}[X_i = 1, U_i = k] \mathbb{E} \left| \frac{\mathbb{P}[X_i = 1, U_i = k|W_i]}{\mathbb{P}[X_i = 1, U_i = k]} - 1 \right|, \tag{3.46}$$

$$= n \sum_{k=0}^{2\lfloor b_n \rfloor} \mathbb{E} \left| \mathbb{P}[X_i = 1, U_i = k|W_i] - \mathbb{P}[X_i = 1, U_i = k] \right| \tag{3.47}$$

by the transitivity of the cycle. Let

$$C_i^c = \{j : |j - i| \leq (c_n + b_n)/2\} \tag{3.48}$$

be the set of vertices within distance $(b_n + c_n)/2$ of i and define the events

$$A_1 = \{\exists j \in B_i \cup \{i\}, \exists t \in [0, t^*] : \mathcal{H}_j(t) \nsubseteq C_i^c\} \tag{3.49}$$

and

$$A_2 = \{\exists j \in D_i, \exists t \in [0, t^*] : \mathcal{H}_j(t) \cap C_i^c \neq \emptyset\} \tag{3.50}$$

as well as their intersection

$$A = A_1 \cap A_2. \tag{3.51}$$

From Lemma 3.8,

$$\mathbb{P}[X_i = 1, U_i = j | A^{\complement}, W_i] = \mathbb{P}[X_i = 1, U_i = j | A^{\complement}]. \tag{3.52}$$

Continuing on from (3.47), we split the probabilities into

$$\begin{aligned}
\delta_1 = n \sum_{k=0}^{2\lfloor b_n \rfloor} \mathbb{E} \Big| &\mathbb{P}[X_i = 1, U_i = k | W_i, A]\mathbb{P}[A|W_i] - \mathbb{P}[X_i = 1, U_i = k | A]\mathbb{P}[A] + &(3.53)\\
&\mathbb{P}(X_i = 1, U_i = k | A^{\complement})(\mathbb{P}[A^{\complement}|W_i] - \mathbb{P}[A^{\complement}]) \Big| \\
&\leq n(2\lfloor b_n \rfloor + 1)\mathbb{E}\left[\mathbb{P}[A|W_i] + \mathbb{P}[A] + \left|\mathbb{P}[A^{\complement}|W_i] - \mathbb{P}[A^{\complement}]\right|\right] &(3.54)\\
&= n(2\lfloor b_n \rfloor + 1)\mathbb{E}\left[\mathbb{P}[A|W_i] + \mathbb{P}[A] + |1 - \mathbb{P}[A|W_i] - (1 - \mathbb{P}[A])|\right] &(3.55)\\
&\leq n(2\lfloor b_n \rfloor + 1)\mathbb{E}\left[\mathbb{P}[A|W_i] + \mathbb{P}[A] + \mathbb{P}[A|W_i] + \mathbb{P}[A])\right] &(3.56)\\
&= 2n(2\lfloor b_n \rfloor + 1)\left(\mathbb{E}[\mathbb{P}[A|W_i]] + \mathbb{P}[A]\right) &(3.57)\\
&= 4n(2\lfloor b_n \rfloor + 1)\mathbb{P}[A]. &(3.58)
\end{aligned}$$

For either A_1 or A_2 to hold, there must exists a history that spreads at least distance $(c_n - b_n)/2$ away from its starting vertex. By a union bound

$$\begin{aligned}
\mathbb{P}[A] &\leq \sum_{j=1}^{n} \mathbb{P}[\mathcal{H}_i \nsubseteq B(i, (c_n - b_n)/2) \times [0, t^*]] &(3.59)\\
&= n\mathbb{P}\left[\bigcup_{u \in [0,t^*]} \mathcal{H}_i(t^* - u) \nsubseteq B(i, (c_n - b_n)/2)\right] &(3.60)
\end{aligned}$$

Combining this with Lemma 3.9, and recalling our choices of $b_n = \ln(n)$ and $c_n = \ln(n)^2$ we get that

$$\delta_1 \leq 8n^2(2\lfloor b_n \rfloor + 1)\exp\left((z+\ln n)/\theta - \ln 2(c_n - b_n)/2\right) \tag{3.61}$$

$$\leq 16\exp(z/\theta)n^{3+1/\theta+\ln 2/2-\ln n\ln 2/2} \tag{3.62}$$

which goes to 0 as $n \to \infty$. $\square$

Lemma 3.7. *Let δ_4 be as defined above in (3.18). Then*

$$\lim_{n\to\infty} \delta_4 = 0. \tag{3.63}$$

Proof. Starting with the definition of δ_4, we have

$$\delta_4 = \sum_{i=1}^{n} \left(\mathbb{E}[X_i Z_i] + \mathbb{E}[X_i]\mathbb{E}[X_i + U_i + Z_i]\right), \tag{3.64}$$

$$= n\mathbb{P}[X_i = 1]\mathbb{E}[Z_i | X_i = 1] + e^{-z}\sum_{j\in\{i\}\cup B_i\cup C_i} \mathbb{E}[X_j], \tag{3.65}$$

$$= e^{-z}\mathbb{E}[Z_i|X_i = 1] + \frac{(2\lfloor c_n \rfloor + 1)e^{-2z}}{n}. \tag{3.66}$$

Now

$$\mathbb{E}[Z_i|X_i = 1] = \sum_{j\in C_i} \mathbb{P}[X_j = 1|X_i = 1], \tag{3.67}$$

$$= 2\sum_{k=\lfloor b_n\rfloor+1}^{\lfloor c_n\rfloor} \mathbb{P}[X_{i+k} = 1|X_i = 1]. \tag{3.68}$$

From Lemma 3.10,

$$\mathbb{E}[Z_i|X_i = 1] \leq 2\sum_{k=\lfloor b_n\rfloor+1}^{\lfloor c_n\rfloor}\left(\frac{e^{-z}}{n} + 2\left(\frac{2-\sqrt{\theta(4-\theta)}}{2-\theta}\right)^k\right), \tag{3.69}$$

$$\leq \frac{2(c_n - b_n + 1)e^{-z}}{n} + 4(c_n - b_n + 1)\left(\frac{2-\sqrt{\theta(4-\theta)}}{2-\theta}\right)^{b_n+1}. \tag{3.70}$$

Altogether,

$$\delta_4 \leq \frac{2(c_n - b_n + 1)e^{-z}}{n} + 4(c_n - b_n + 1)\left(\frac{2-\sqrt{\theta(4-\theta)}}{2-\theta}\right)^{b_n+1} + \frac{(2c_n + 1)e^{-2z}}{n} \tag{3.71}$$

which, recalling that $b_n = \ln(n)$ and $c_n = \ln(n)^2$, goes to 0 as $n \to \infty$. $\square$

3.4 Additional lemmas

This section contains the proofs for a number of lemmas concerning properties of the update histories on the cycle.

Lemma 3.8. *Let i be a vertex in a graph and let*

$$C_i^c = \{j : |j - i| \leq (b_n - c_n)/2\} \tag{3.72}$$

be the set of vertices within distance $(b_n + c_n)/2$ of i. Define the events

$$A_1 = \{\exists j \in B_i \cup \{i\}, \exists t \in [0, t^*] : \mathcal{H}_j(t) \nsubseteq C_i^c\} \tag{3.73}$$

and

$$A_2 = \{\exists j \in D_i, \exists t \in [0, t^*] : \mathcal{H}_j(t) \cap C_i^c \neq \emptyset\} \tag{3.74}$$

as well as their union

$$A = A_1 \cup A_2. \tag{3.75}$$

Then

$$\mathbb{P}[X_i = 1, U_i = j | A^{\complement}, W_i] = \mathbb{P}[X_i = 1, U_i = j | A^{\complement}]. \tag{3.76}$$

Proof. If $A_1^{\complement}$ holds, then the events $\{X_1 = 1\}$ and $\{U_i = j\}$ depend only on the values of the update sequence inside C_i^c. If $A_2^{\complement}$ holds then the events $\{W_i = k\}$, $k \geq 0$, depend only on the values of the update sequence outside of C_i^c. Since the update sequences of each vertex are independent of each other vertex, if $A_2^{\complement}$ holds, conditioning on W_i does not affect the update sequences inside C_i^c and so

$$\mathbb{P}[X_i = 1, U_i = j | A^{\complement}, W_i] = \mathbb{P}[X_i = 1, U_i = j | A^{\complement}]. \tag{3.77}$$

□

The following Lemma bounds how fast updates can percolate through the cycle. The proof is a slight modification of a similar one in [14].

Lemma 3.9. *Consider the update histories on the cycle. Let $B(i, l)$ indicate the set of vertices at integer distance l or smaller from vertex i. The probability that the history of vertex i escapes $B(i, l)$ in time s is bounded by*

$$\mathbb{P}\left[\bigcup_{u \in [0,s]} \mathcal{H}_i(t^* - u) \nsubseteq B(i, l)\right] \leq 2 \exp\left(s - l \ln 2\right). \tag{3.78}$$

Proof. Let $\boldsymbol{w}^- = (i, i-1, \ldots, i-l)$ and $\boldsymbol{w}^+ = (i, i+1, \ldots, i+l)$ denote the sequences of adjacent vertices starting at vertex i and extending distance l to the left and right respectively. For $\mathcal{H}_i$ to contain any vertex outside $B(i,l)$ at a time $u \in [t^* - s, t^*]$ then either each w_k^- was updated at some time $t^* > t_k \geq t^* - s$ with $t_{k-1} > t_k$ or each w_k^+ was updated at some time $t^* > t_k \geq t^* - s$ with $t_{k-1} > t_k$. Call the first event M_- and the second M_+. We have

$$\mathbb{P}[M_-] = \mathbb{P}[M_+] = \mathbb{P}[\mathrm{Po}(s) \geq l] \tag{3.79}$$

where $\mathrm{Po}(s)$ is Poisson with rate s. By a union bound,

$$\mathbb{P}\left[\bigcup_{u\in[0,s]} \mathcal{H}_i(t^* - u) \not\subseteq B(i,l)\right] \leq 2\mathbb{P}[\mathrm{Po}(s) \geq l]. \tag{3.80}$$

The moment generating function of a Poisson random variable with rate s is

$$M(t) = \exp\left(s\left(e^t - 1\right)\right). \tag{3.81}$$

Using a Chernoff bound we have for every $t > 0$,

$$\mathbb{P}[\mathrm{Po}(s) \geq l] \leq \exp\left(s\left(e^t - 1\right) - tl\right). \tag{3.82}$$

Overall we have

$$\mathbb{P}\left[\bigcup_{u\in[0,s]} \mathcal{H}_i(t^* - u) \not\subseteq B(i,l)\right] \leq 2\exp\left(s\left(e^t - 1\right) - tl\right). \tag{3.83}$$

Choosing $t = \ln 2$,

$$\mathbb{P}\left[\bigcup_{u\in[0,s]} \mathcal{H}_i(t^* - u) \not\subseteq B(i,l)\right] \leq 2\exp\left(s - l\ln 2\right). \tag{3.84}$$

□

Lemma 3.10. *Let i and j be the indices of two vertices on the cycle $(\mathbb{Z}/n\mathbb{Z})$ separated by distance k. Then*

$$\mathbb{P}[X_j = 1 | X_i = 1] \leq \frac{e^{-z}}{n} + 2\left(\frac{2 - \sqrt{\theta(4-\theta)}}{2-\theta}\right)^k. \tag{3.85}$$

Proof. There are two ways in which the update history of vertex j can survive until time 0. The update history can survive without intersecting with the update history of vertex i or the update history of vertex j can merge with the update history of vertex i (whose survival we are conditioning on). Breaking up the probability this way we have

$$\begin{aligned}\mathbb{P}[X_j = 1|X_i = 1] = \mathbb{P}[X_j = 1, \mathcal{H}_i \cap \mathcal{H}_j = \emptyset | X_i = 1] \\ + \mathbb{P}[X_j = 1, \mathcal{H}_i \cap \mathcal{H}_j \neq \emptyset | X_i = 1] \quad (3.86)\\ \leq \mathbb{P}[X_j = 1, \mathcal{H}_i \cap \mathcal{H}_j = \emptyset | X_i = 1] + \mathbb{P}[\mathcal{H}_i \cap \mathcal{H}_j \neq \emptyset | X_i = 1]. \quad (3.87)\end{aligned}$$

The result follows from Lemmas 3.11 and 3.12. □

Lemma 3.11. *Let i and j be the indices of two vertices on the cycle $(\mathbb{Z}/n\mathbb{Z})$. Then*

$$\mathbb{P}[X_j = 1, \mathcal{H}_i \cap \mathcal{H}_j = \emptyset | X_i = 1] \leq e^{-z}/n. \quad (3.88)$$

Proof. To begin

$$\begin{aligned}\mathbb{P}[X_j = 1, \mathcal{H}_i \cap \mathcal{H}_j = \emptyset | X_i = 1] = \mathbb{P}[X_i = 1]^{-1}\mathbb{P}[X_i = 1, X_j = 1, \mathcal{H}_i \cap \mathcal{H}_j = \emptyset] \quad (3.89)\\ = \frac{\mathbb{P}[\mathcal{H}_i(0) \cup \mathcal{H}_j(0) \neq \emptyset, \mathcal{H}_i \cap \mathcal{H}_j = \emptyset]}{\mathbb{P}[\mathcal{H}_i(0) \neq \emptyset]}. \quad (3.90)\end{aligned}$$

Define S to be the first time that the histories intersect, or define $S = 0$ if the histories do not intersect before time 0. We construct a new history, $\mathcal{H}'_j$, in the following way. Along the interval $[S, t^*)$, $\mathcal{H}'_j$ is constructed from the update sequence in the exact same way as $\mathcal{H}_j$. However, along the interval $[0, S)$, we replace the update sequence with an another i.i.d. copy of the update sequence. We have constructed $\mathcal{H}'_j$ to be independent of $\mathcal{H}_i$. This is obvious for $t < S$. For $t \geq S$ we note that no updates are a part of both histories. From the memoryless property of the exponential waiting times between updates, we get that each update is independent of the rest, and so the histories are independent too. Note that in the event $\{\mathcal{H}_i \cap \mathcal{H}_j = \emptyset\}$, we have $\mathcal{H}_j = \mathcal{H}'_j$ and so

$$\begin{aligned}\mathbb{P}[\mathcal{H}_i(0) \cup \mathcal{H}_j(0) \neq \emptyset, \mathcal{H}_i \cap \mathcal{H}_j = \emptyset] = \mathbb{P}[\mathcal{H}_i(0) \cup \mathcal{H}'_j(0) \neq \emptyset, \mathcal{H}_i \cap \mathcal{H}_j = \emptyset] \quad (3.91)\\ \leq \mathbb{P}[\mathcal{H}_i(0) \cup \mathcal{H}'_j(0) \neq \emptyset] \quad (3.92)\\ = \mathbb{P}[\mathcal{H}_i(0) \neq \emptyset]\mathbb{P}[\mathcal{H}'_j(0) \neq \emptyset] \quad (3.93)\end{aligned}$$

since $\mathcal{H}_i$ and $\mathcal{H}'_j$ are independent. Since $\mathbb{P}[\mathcal{H}'_j(0) \neq \emptyset] = \mathbb{P}[\mathcal{H}_j(0) \neq \emptyset] = e^{-z}/n$ we get the desired result. □

Lemma 3.12. *Let i and j be the indices of two vertices that are separated by distance k. Then*

$$\mathbb{P}[\mathcal{H}_i \cap \mathcal{H}_j \neq \emptyset | X_j = 1] \leq 2\left(\frac{2 - \sqrt{\theta(4-\theta)}}{2-\theta}\right)^k. \quad (3.94)$$

where $k = |i - j|$.

Proof. We first must deal with the effect that conditioning on $X_j = 1$ has on the probability that the two update histories merge. For the history of vertex j to survive, all updates along the history must be non-oblivious updates. We also note that conditioning on the history of vertex j surviving should not result in an increase in the overall rate of updates (since each update is a chance that the history will die). By this reasoning,

$$\mathbb{P}[\mathcal{H}_i \cap \mathcal{H}_j \neq \emptyset | X_j = 1] \leq \mathbb{P}[\mathcal{H}_i \cap \bar{\mathcal{H}}_j] \tag{3.95}$$

where $\bar{\mathcal{H}}_j$ is an undying random walk that starts at j and going backwards in time from t^* moves right at rate $1/2$ and left at rate $1/2$.

We note that, while $\mathcal{H}_i$ survives, the distance between $\mathcal{H}_i$ and $\bar{\mathcal{H}}_j$ is a birth and death process that starts at $k = |i - j|$ and has birth and death rates, $\lambda = \mu = (2 - \theta)/2$. Let $P(t)$ be such a process and define $s_0 = \inf\{t : P(t) = 0\}$ to be the first time the process reaches zero (this corresponds to the update histories merging). Let s_d be exponentially distributed with rate θ (this corresponds to the update history of vertex i dying). Then

$$\mathbb{P}[\mathcal{H}_i \cap \mathcal{H}_j \neq \emptyset | X_j = 1] \leq \mathbb{P}[\mathcal{H}_i \cap \bar{\mathcal{H}}_j] \leq 2\mathbb{P}_k(s_0 < s_d) \tag{3.96}$$

where $\mathbb{P}_k$ indicates that $P(0) = k$. The factor of two comes from the fact that the update histories may meet by going the other direction around the cycle. We also note that we allow $P(t)$ to continue beyond $t = t^*$, unlike our update histories which stop at time 0. This does not present a problem as the effect of allowing this is to increase the size of our upper bound.

At any time before s_d there are three possibilities for what can happen to P next. Either the next event is a birth with probability $(2 - \theta)/4$, the next event is a death with the same probability or we reach time s_d with probability $\theta/2$. Writing $\zeta_k = \mathbb{P}_k(s_0 < s_d)$ this gives us the recurrence relation

$$\zeta_k = \frac{2-\theta}{4}\zeta_{k-1} + \frac{2-\theta}{4}\zeta_{k+1} \tag{3.97}$$

which is subject to the conditions

$$\zeta_0 = 1 \tag{3.98}$$

$$\zeta_k \leq 1, \, \forall k \in \mathbb{N}. \tag{3.99}$$

This recurrence has characteristic equation

$$x^2 - \frac{4}{2-\theta}x + 1 = 0 \tag{3.100}$$

which has roots

$$r_1 = \frac{2 + \sqrt{\theta(4-\theta)}}{2-\theta} \tag{3.101}$$

$$r_2 = \frac{2 - \sqrt{\theta(4-\theta)}}{2-\theta} \tag{3.102}$$

and so

$$\zeta_k = ar_1^k + br_2^k \tag{3.103}$$

where a and b are constants to be determined from (3.98) and (3.99). We note that $r_1 \geq 1, \forall \theta \in [0,1]$ and so from (3.99) we have that $a = 0$. Finally from (3.98), $b = 1$ and so

$$\zeta_k = \left(\frac{2 - \sqrt{\theta(4-\theta)}}{2-\theta} \right)^k . \tag{3.104}$$

$\square$

Chapter 4

The Coupling Time on Vertex Transitive Graphs

For the most part, this chapter will be similar in structure and content to Chapter 3. The main difference is that we extend the family of graphs on which we consider the Ising heat-bath Glauber dynamics from the cycle to any vertex transitive graph. Again, the main result, Theorem 4.2, concerns the coupling time, T_n, as defined in Section 2.2.2, and establishes that at sufficiently high temperature (that is, for β small enough), the coupling time converges in distribution to a Gumbel distribution.

Restricting β to be sufficiently small is a consequence of the increased generality of this chapter. As mentioned in Chapter 3, in the high-temperature regime we expect the dynamics to be similar to those when $\beta = 0$. When $\beta = 0$ the problem simplifies to the coupon collector's problem, which is known to have a Gumbel limit. However, at the critical temperature, and below in the low-temperature regime, there is no reason to suspect that the dynamics will behave similarly to when $\beta = 0$. So our restriction of β to be small enough for the result to hold is, on at least a descriptive level, somewhat expected.

Our result partially confirms the conjecture by Collevecchio et al. in [8] that the coupling time for the Ising heat-bath process on the d-dimensional lattice, $G_L = (\mathbb{Z}/L\mathbb{Z})^d$, converges to a Gumbel distribution as $L \to \infty$ for all $\beta < \beta_C$. Our result does not hold all the way up until the critical temperature. This is due to the fact that we are considering a larger class of graphs than just the square lattice, and so it is unreasonable to expect such a result to provide sharp bounds for the lattice. A separate treatment of the square lattice in particular may be needed for a result holding all the way up to the critical temperature. Note that this is what Lubetzky and Sly did in [12] to prove the existence of cutoff for the full high-temperature regime. Since our proof is also based on information percolation there is good reason to think that a similar approach could also work to extend our result.

The main part of our proof which requires β to be sufficiently small is in Lemma 4.16. In particular, this Lemma concerns a quantity (see the comments immediately preceding Lemma 4.16) that is very similar to a quantity used in [13] to prove the existence of cutoff. This similarity further encourages future efforts to sharpen our result.

To state the main result we first must define the graphs on which it is valid. Let (G_n) be a sequence of vertex-transitive graphs with fixed degree Δ and n vertices. Let $P_n(k)$ denote the number of vertices at distance k from a vertex i in G_n and let $Q_n(k)$ denote the number of vertices at distance k or less from a vertex i in G_n. Define

$$\mathscr{G} = \left\{ (G_n) : \exists C_2 > 0 \text{ such that } \lim_{n\to\infty} \sum_{k=1}^{\infty} P_n(k)e^{-k} \leq C_2, Q_n(\ln^2(n)) = o(n) \right\}. \quad (4.1)$$

In this chapter we consider sequences of vertex-transitive graphs $(G_n) \in \mathscr{G}$.

It is worth verifying that the set $\mathscr{G}$ contains some graph sequences that are of interest. We start by showing that it contains the d-dimensional discrete tori.

Lemma 4.1. *Define the sequence of length L d-dimensional square lattices on a torus, $(G_L)_{L\geq 1}$, via $G_L = (\mathbb{Z}/\mathbb{Z}L)^d$. Then $(G_L)_{L\geq 1} \in \mathscr{G}$.*

Proof. The torus $G_L = (\mathbb{Z}/\mathbb{Z}L)^d$ is obviously vertex transitive and the sequence $(G_L)_{L\geq 1}$ has fixed degree $\Delta = 2d$. By definition $P_n(k) \leq Q_n(k)$ and we can upper bound $Q_n(k)$ by the number of vertices contained within distance k of the origin on the infinite d-dimensional integer lattice, $Q_\infty(k)$.

To bound $Q_\infty(k)$, consider first the number of vertices contained within distance k of the origin in the closed positive orthant, $Q_\infty(k)^+$. That is, the number of vertices $\boldsymbol{x} = (x_1, \ldots, x_d)$ with $x_i \geq 0$ and such that

$$\sum_{i=1}^{d} x_i \leq k \quad (4.2)$$

We note that we can represent each vertex in the positive orthant within distance k of the origin by k stars separated by d bars, and taking x_i to be the number of stars between bars $i-1$ and i. So for example, if $k = 5$ and $d = 2$, the arrangement

$$\star \mid \star\star \| \star\star \quad (4.3)$$

represents the vertex $\boldsymbol{x} = (1, 2, 0)$. Note that any stars after the last bar are not included (this accounts for $\boldsymbol{x}$ being closer than distance k to the origin). There are $k + d$ choose d ways to arrange the stars and bars and so

$$Q_\infty(k)^+ \leq \binom{k+d}{d} \tag{4.4}$$

and since in d dimensions there are 2^d orthants,

$$P_n(k) \leq Q_n(k) \leq Q_\infty(k) \leq \frac{2^d}{d!}(k+1)(k+2)\ldots(k+d). \tag{4.5}$$

This is a degree d polynomial in k which satisfies the constraints in (4.1). □

Another class of graphs contained in $\mathscr{G}$ are the mth powers of $G_L = (\mathbb{Z}/\mathbb{Z}L)^d$. The mth power of a graph G is a graph, G^m, with the same vertices as G, but in which we make adjacent all vertices whose distance in G is no more than m. Clearly, $Q_n^{G^m}(k) \leq Q_n^G(mk)$, and G^m inherits the transitivity of G. Hence sequences of powers of $G_L = (\mathbb{Z}/\mathbb{Z}L)^d$ are also in $\mathscr{G}$. This is of practical interest since the Ising model on G^m corresponds to the Ising model with m nearest neighbours on G.

We now define a couple of quantities. Firstly, the *magnetization at vertex i at time t* is

$$m_t(i) = \mathbb{E}[\mathcal{T}_t[i]] \tag{4.6}$$

where $(\mathcal{T}_t)_{t\geq 0}$ is the dynamics starting from the all-plus configuration. Note that on transitive graphs, with which this chapter is concerned, we can drop the dependence on i and simply write m_t for the magnetization at any vertex at time t. This quantity is not to be confused with the magnetization on a volume (as mentioned in Section 2.1.1) which has nothing to do with dynamics. The magnetization on a volume is a random variable given by the sum of all the spins on a given volume where the spins are distributed according to the Gibbs measure. In contrast, the magnetization at a vertex, as defined here, is a deterministic function of time that is a property of the Glauber dynamics.

We can now define the time

$$t_c(n) = \inf\left\{t > 0 : m_t = \frac{1}{n}\right\}. \tag{4.7}$$

which is around the time it takes for the top and bottom chains to couple.

Theorem 4.2. *Let $(G_n) \in \mathscr{G}$ be a sequence of vertex-transitive graphs, $G_n = (V, E)$ with $|V| = n$ vertices. Let T_n be the coupling time for the continuous-time Ising heat-bath dynamics for the zero-field ferromagnetic Ising model on G_n. Then for any small* *enough inverse-temperature β, there exists a subsequence (T_m) of (T_n) such that*

$$\lim_{m\to\infty} \mathbb{P}[T_m < z + t_c(m)] = e^{-C_1 e^{-C_2 z}} \tag{4.8}$$

for some

$$C_1 = C_1(\beta, \mathscr{G}) \in (0, 1] \tag{4.9}$$

and

$$C_2 = C_2(\beta, \mathscr{G}) \in [1 - \beta\Delta, 1] \, . \tag{4.10}$$

The proof of Theorem 4.2 will be given in Section 4.3 after the essential preliminaries are presented. In Section 4.1 we describe an alternative construction of the histories that can sometimes be easier to work with. Then in Section 4.2, we outline the overall approach to the proof. The method is very similar to the method used in Chapter 3 but there are some additional problems that are addressed. Finally, we defer results directly concerning the update histories to Section 4.4.

4.1 Information percolation in higher dimensions

In the previous chapter, we showed that on the cycle, there was a coupling that made the update history of a single vertex to be a continuous-time random walk that died at rate θ. On lattices of dimension $d > 2$, we can no longer use this coupling and so the updates histories are significantly more complex.

Recall from Section 2.3.2.2 that given a target time t^*, the update history of a vertex set A at time t, $\mathcal{H}_A(t)$, is the set of vertices whose spins at time t determine the spins of A at time t^*. Developing this history backwards in time from $t = t^*$ produces a subgraph of $\Omega \times [0, t^*]$ which we write as $\mathcal{H}_A$ and call the update history of vertex set A. This history can be constructed using the update sequence along $(t, t^*]$.

In practise, we may choose to construct this history as follows: For each $i \in A$, create a temporal edge between (i, t^*) and (i, t_i) where t_i is the time of the latest update to i (or 0 if i is never updated). Then for each update (i, u, t_i), we either terminate the edge if u is such that the update is oblivious, or we add spatial branches to each of the neighbours of i. We repeat this process recursively for the neighbours of i until every branch has been terminated due to an oblivious update or has reached time 0.

However, it is possible for vertices to be removed from $\mathcal{H}_A(t)$ from updates that are not oblivious (see Figure 2.4). Since our method above for constructing the history does not take this into account, the history it produces will possibly be larger than $\mathcal{H}_A$. To ensure a distinction between the two, the history that results from the above construction we will denote $\hat{\mathcal{H}}_A$, and likewise $\hat{\mathcal{H}}_A(t)$ for the history at time t that results from the above construction. We have that

$$\mathcal{H}_A(t) \subseteq \hat{\mathcal{H}}_A(t) \tag{4.11}$$

and also that $\mathcal{H}_A$ is a subgraph of $\hat{\mathcal{H}}_A$.

4.1.1 The magnetization

One quantity which we used multiple times in Chapter 3 was $\mathbb{P}[X_i = 1]$. Although it was not required earlier, we would now like to make clear that this is in fact the magnetization at time t^*.

Recall that the magnetization at vertex $i \in V$ at time $t > 0$ is defined to be

$$m_t(i) = \mathbb{E}[\mathcal{T}_t[i]] \tag{4.12}$$

where $(\mathcal{T}_t)_{t\geq 0}$ is the dynamics starting from the all-plus configuration. Given a monotonically coupled chain $(\mathcal{B}_t)_{t\geq 0}$, starting in the all minus configuration and such that $\mathcal{T}_t[i] \geq \mathcal{B}_t[i]$ for all $t \geq 0$ and $i \in V$, we can split up this expectation by conditioning on the event $A_t = \{\mathcal{T}_t[i] \neq \mathcal{B}_t[i]\}$. We obtain that

$$\begin{aligned} m_t(i) &= \mathbb{E}[Y_t^+[i]] && (4.13)\\ &= \mathbb{P}\left[A_t\right]\left(\mathbb{P}\left[Y_t^+[i] = 1 | A_t\right] - \mathbb{P}\left[Y_t^+[i] = -1 | A_t\right]\right) && (4.14)\\ &\quad + \mathbb{P}\left[A_t^{\complement}\right]\left(\mathbb{P}\left[Y_t^+[i] = 1 | A_t^{\complement}\right] - \mathbb{P}\left[Y_t^+[i] = -1 | A_t^{\complement}\right]\right). \end{aligned}$$

Now if event $A_t^{\complement}$ holds, $\mathcal{T}_t[i] = \mathcal{B}_t[i]$, and so by symmetry vertex i must take values -1 and $+1$ uniformly. Furthermore, by the monotonicity of our coupling, if A_t holds, we must have that $\mathcal{T}_t[i] = +1$ and $\mathcal{B}_t[i] = -1$. So

$$m_t(i) = \mathbb{P}[A_t]. \tag{4.15}$$

Finally, given a target time t^*, X_i is defined such that $\{X_i = 1\} = A_{t^*}$. So

$$\mathbb{P}[X_i = 1] = m_{t^*}(i). \tag{4.16}$$

We end this section with some results concerning the magnetization, and in particular, the magnetization at time

$$t^* = t_c(n) + z. \tag{4.17}$$

The following comes from [15] and is valid on any graph, not just transitive ones.

Lemma 4.3 ([15], Claim 3.3)**.** *On any graph with maximum degree Δ, for any $t, s > 0$ we have*

$$e^{-2s} \leq \frac{\sum_i m_{t+s}[i]^2}{\sum_i m_t[i]^2} \leq e^{-2(1-\beta\Delta)s}. \tag{4.18}$$

The following corollaries are then straightforward.

Corollary 4.4. *On any vertex transitive graph with degree Δ, for any $t, s > 0$ we have*

$$e^{-s} m_t \leq m_{t+s} \leq m_t e^{-(1-\beta\Delta)s}. \tag{4.19}$$

Corollary 4.5. *On any vertex transitive graph with degree Δ, m_{t^*} can be bounded as follows:*

For $z \geq 0$,

$$\frac{e^{-z}}{n} \leq m_{t^*} \leq \frac{e^{-(1-\beta\Delta)z}}{n}. \tag{4.20}$$

For $z \leq 0$,

$$\frac{e^{-(1-\beta\Delta)z}}{n} \leq m_{t^*} \leq \frac{e^{-z}}{n}. \tag{4.21}$$

Bearing in mind that $m_0 = 1$, we also obtain a bound on $t_c(n)$.

Corollary 4.6. *On any vertex transitive graph with degree Δ, for $\beta < 1/\Delta$*

$$\ln(n) \leq t_c(n) \leq \frac{\ln(n)}{1-\beta\Delta}. \tag{4.22}$$

4.2 Problem set-up

As in Chapter 3, we prove Theorem 4.2 through the use of a stronger statement in Theorem 4.7. The overall approach is almost exactly as described in Section 3.2; at a time t^* we count the number of vertices at which the bottom and top chains differ and show that the distribution of this random variable, which we call W, is close to an appropriately chosen compound Poisson distribution. The main difference in the construction is that we use a different t^* here.

Recalling the definition of $t_c(n)$ in (4.7), fix z and a time of interest $t^* = t_c(n) + z$. From here we define X_i, U_i, W_i, δ_1, δ_4, λ, and μ exactly as in Chapter 3 but using our new definition for t^*. For convenience, we repeat the definitions here.

For each vertex $i \in V$, define indicators

$$X_i = \begin{cases} 1 & \mathscr{B}_{t^*}[i] \neq \mathscr{T}_{t^*}[i], \\ 0 & \mathscr{B}_{t^*}[i] = \mathscr{T}_{t^*}[i] \end{cases} \tag{4.23}$$

and set $W = \sum_{i \in V} X_i$. For each $i \in V$, decompose W into $W = X_i + U_i + Z_i + W_i$ where

$$U_i = \sum_{j \in B_i} X_j, \qquad Z_i = \sum_{j \in C_i} X_j, \qquad W_i = \sum_{j \in D_i} X_j. \tag{4.24}$$

and B_i, C_i, and D_i are the vertex sets

$$B_i = \{j \neq i : |j - i| \leq b_n\}, \tag{4.25}$$

$$C_i = \{j \notin B_i \cup \{i\} : |j - i| \leq c_n\}, \tag{4.26}$$

$$D_i = V \setminus (B_i \cup C_i \cup \{i\}). \tag{4.27}$$

As in Chapter 3, we have some freedom in choosing b_n and c_n but we will again choose $b_n = \ln(n)$ and $c_n = \ln(n)^2$.

We now define the quantities

$$\lambda = \sum_{i \in V} \mathbb{E}\left[\frac{X_i}{X_i + U_i} I[X_i + U_i \geq 1]\right], \tag{4.28}$$

$$\mu_l = \frac{1}{l\lambda} \sum_{i \in V} \mathbb{E}\left[X_i I[X_i + U_i = l]\right], \qquad l \geq 1, \tag{4.29}$$

which will be the parameters of the approximating compound Poisson distribution to W. We also define

$$\delta_1 = \sum_{i \in V} \sum_{k \geq 0} \mathbb{P}[X_i = 1, U_i = k] \mathbb{E}\left|\frac{\mathbb{P}[X_i = 1, U_i = k | W_i]}{\mathbb{P}[X_i = 1, U_i = k]} - 1\right|, \tag{4.30}$$

$$\delta_4 = \sum_{i \in V} \left(\mathbb{E}[X_i Z_i] + \mathbb{E}[X_i]\mathbb{E}[X_i + U_i + Z_i]\right), \tag{4.31}$$

which we desire to be small for the compound Poisson approximation to be good.

The following theorem (reworked from [23]) bounds the distance between the distributions of W and the approximating compound Poisson.

Theorem 4.7 ([23]). *Let W, λ, $\boldsymbol{\mu}$, δ_1 and δ_4 be as defined above. Then*

$$d_{\mathrm{TV}}(\mathscr{L}(W), \mathrm{CP}(\lambda, \boldsymbol{\mu})) \leq (\delta_1 + \delta_4)e^{\lambda}. \tag{4.32}$$

As per the discussion proceeding Theorem 3.2, we obtain the following as a corollary to Theorem 4.7.

Corollary 4.8. *Let T_n be the coupling time of the continuous-time heat-bath Glauber dynamics for the zero-field Ising model at inverse-temperature β on the graph G_n and let δ_1, δ_4 and λ be as defined above. Then*

$$\left|\mathbb{P}\left[T_n \leq z + t_c(n)\right] - e^{-\lambda}\right| \leq (\delta_1 + \delta_4)e^{\lambda}, \tag{4.33}$$

where $\theta = 1 - \tanh(2\beta)$.

4.3 Proof of Theorem 4.2

In this section we use Corollary 4.8 to prove Theorem 4.2 by bounding λ and showing that δ_1 and δ_4 go to zero as $n \to \infty$. This is done in Lemmas 4.9, 4.10, and 4.11. The proofs of these require some additional lemmas concerning properties of the update histories which have been deferred to Section 4.4.

We begin by bounding λ. Note that bounding λ is enough to show that there is a subsequence of graphs on which λ converges as required by Theorem 4.2.

Lemma 4.9. *Using the above set-up,*

$$\limsup_{n\to\infty} \lambda \leq \max(e^{-z}, e^{-(1-\beta\Delta)z}) \tag{4.34}$$

and there exists a constant $C \in (0, 1)$ such that

$$\liminf_{n\to\infty} \lambda \geq C \min(e^{-z}, e^{-(1-\beta\Delta)z}). \tag{4.35}$$

Proof. Starting with the definition of λ, we have

$$\begin{aligned}
\lambda &= \sum_{i\in V} \mathbb{E}\left[\frac{X_i}{X_i + U_i} I[X_i + U_i \geq 1]\right] && (4.36)\\
&= \sum_{i=1}^{n} \mathbb{P}(X_i = 1)\mathbb{E}\left[\frac{1}{1 + U_i} | X_i = 1\right] && (4.37)\\
&= nm_{t^*}\mathbb{E}\left[\frac{1}{1 + U_i} | X_i = 1\right] && (4.38)
\end{aligned}$$

where we have used that X_i is zero-one, (4.16), and the transitivity of the graph. Clearly

$$\mathbb{E}\left[\frac{1}{1 + U_i} | X_i = 1\right] \leq 1 \tag{4.39}$$

and so $\lambda \leq nm_{t^*} \leq \max(e^{-z}, e^{-(1-\beta\Delta)z})$.

By Jensen's inequality

$$\mathbb{E}\left[\frac{1}{1+U_i}|X_i = 1\right] \geq \frac{1}{\mathbb{E}[1+U_i|X_i = 1]} \tag{4.40}$$

$$= \frac{1}{1+\mathbb{E}[U_i|X_i = 1]}. \tag{4.41}$$

so in order to find a lower bound for λ we will find an upper bound to $\mathbb{E}[U_i|X_i = 1]$. Now by Lemma 4.14, there exists a $C_1 > 0$ such that for small enough β,

$$\mathbb{E}[U_i|X_i = 1] = \sum_{j \in B_i} \mathbb{P}[X_j = 1|X_i = 1] \tag{4.42}$$

$$\leq |B_i|m_{t^*} + C_1 \sum_{k=1}^{\lfloor b_n \rfloor} \sum_{|j-i|=k} e^{-k}. \tag{4.43}$$

From (4.1),

$$\mathbb{E}[U_i|X_i = 1] \leq |B_i|m_{t^*} + C_1 \sum_{k=1}^{\lfloor b_n \rfloor} P_n(k)e^{-k} \tag{4.44}$$

$$\leq |B_i|m_{t^*} + C_1 \sum_{k=1}^{\infty} P_n(k)e^{-k} \tag{4.45}$$

$$\leq C_z|B_i|/n + C_2 \tag{4.46}$$

for some $C_2 > 0$, $C_z = \max(e^{-z}, e^{-(1-\beta\Delta)z})$. As $n \to \infty$, the first term vanishes and we are left with

$$\liminf_{n\to\infty} \lambda \geq \frac{1}{1+C_2} nm_{t^*} \tag{4.47}$$

$$\geq C \min(e^{-z}, e^{-(1-\beta\Delta)z}) \tag{4.48}$$

for some $C \in (0, 1)$. □

Lemma 4.10. *Let δ_1 be as defined above in (4.30). Then*

$$\lim_{n\to\infty} \delta_1 = 0. \tag{4.49}$$

Proof. Starting with the definition of δ_1, we have

$$\delta_1 = \sum_{i=1}^{n}\sum_{k=0}^{|B_i|} \mathbb{P}[X_i = 1, U_i = k]\mathbb{E}\left|\frac{\mathbb{P}[X_i = 1, U_i = k|W_i]}{\mathbb{P}[X_i = 1, U_i = k]} - 1\right| \tag{4.50}$$

$$= n\sum_{k=0}^{|B_i|} \mathbb{E}\left|\mathbb{P}[X_i = 1, U_i = k|W_i] - \mathbb{P}[X_i = 1, U_i = k]\right| \tag{4.51}$$

by the transitivity of the graph. Let

$$C_i^c = \{j : |j - i| \leq (c_n + b_n)/2\} \tag{4.52}$$

be the set of vertices within distance $(b_n + c_n)/2$ of i and define the events

$$A_1 = \{\exists j \in B_i \cup \{i\}, \exists t \in [0, t^*] : \mathcal{H}_j(t) \nsubseteq C_i^c\} \tag{4.53}$$

and

$$A_2 = \{\exists j \in D_i, \exists t \in [0, t^*] : \mathcal{H}_j(t) \cap C_i^c \neq \emptyset\} \tag{4.54}$$

as well as their intersection

$$A = A_1 \cap A_2. \tag{4.55}$$

From Lemma 3.8,

$$\mathbb{P}[X_i = 1, U_i = j|A^{\complement}, W_i] = \mathbb{P}[X_i = 1, U_i = j|A^{\complement}]. \tag{4.56}$$

Continuing on from (4.51), we split the probabilities into

$$\begin{aligned}
\delta_1 &= n\sum_{k=0}^{|B_i|} \mathbb{E}\Big|\mathbb{P}[X_i = 1, U_i = k|W_i, A]\mathbb{P}[A|W_i] - \mathbb{P}[X_i = 1, U_i = k|A]\mathbb{P}[A] + \qquad (4.57)\\
&\qquad \mathbb{P}(X_i = 1, U_i = k|A^{\complement})(\mathbb{P}[A^{\complement}|W_i] - \mathbb{P}[A^{\complement}])\Big| \\
&\leq n(|B_i| + 1)\mathbb{E}\left[\mathbb{P}[A|W_i] + \mathbb{P}[A] + \left|\mathbb{P}[A^{\complement}|W_i] - \mathbb{P}[A^{\complement}]\right|\right] \qquad (4.58)\\
&= n(|B_i| + 1)\mathbb{E}\left[\mathbb{P}[A|W_i] + \mathbb{P}[A] + |1 - \mathbb{P}[A|W_i] - (1 - \mathbb{P}[A])|\right] \qquad (4.59)\\
&\leq n(|B_i| + 1)\mathbb{E}\left[\mathbb{P}[A|W_i] + \mathbb{P}[A] + \mathbb{P}[A|W_i] + \mathbb{P}[A])\right] \qquad (4.60)\\
&= 2n(|B_i| + 1)\left(\mathbb{E}[\mathbb{P}[A|W_i]] + \mathbb{P}[A]\right) \qquad (4.61)\\
&= 4n(|B_i| + 1)\mathbb{P}[A] \qquad (4.62)
\end{aligned}$$

For either A_1 or A_2 to hold, there must exists a history that spreads at least distance $(c_n - b_n)/2$ away from its starting vertex. By a union bound

$$\mathbb{P}[A] \leq \sum_{j=1}^{n} \mathbb{P}[\mathcal{H}_i \not\subseteq B(i,(c_n-b_n)/2)\times[0,t^*]] \tag{4.63}$$

$$= n\mathbb{P}\left[\bigcup_{u\in[0,t^*]} \mathcal{H}_i(t^*-u) \not\subseteq B(i,(c_n-b_n)/2)\right] \tag{4.64}$$

Combining this with Lemma 4.13, and recalling our choices of $b_n = \ln(n)$ and $c_n = \ln(n)^2$ we get that

$$\delta_1 \leq 4n^2(|B_i|+1)\exp(t^*\Delta - \ln\Delta(c_n-b_n)/2) \tag{4.65}$$

$$\leq 4n^{2+\Delta/(1-\beta\Delta)}(|B_i|+1)\exp(\Delta z)\exp(-\ln\Delta(c_n-b_n)/2) \tag{4.66}$$

which goes to 0 as $n\to\infty$.

□

Lemma 4.11. *Let δ_4 be as defined above in (4.31). Then*

$$\lim_{n\to\infty}\delta_4 = 0. \tag{4.67}$$

Proof. Starting with the definition of δ_4, we have

$$\delta_4 = \sum_{i=1}^{n}\left(\mathbb{E}[X_iZ_i] + \mathbb{E}[X_i]\mathbb{E}[X_i+U_i+Z_i]\right) \tag{4.68}$$

$$= n\mathbb{E}[X_iZ_i] + nm_{t^*}^2\,(1+|B_i|+|C_i|) \tag{4.69}$$

$$= nm_{t^*}\mathbb{E}[Z_i|X_i=1] + nm_{t^*}^2\,(1+|B_i|+|C_i|) \tag{4.70}$$

$$\leq C_z\mathbb{E}[Z_i|X_i=1] + \frac{C_z^2}{n}\,(1+|B_i|+|C_i|) \tag{4.71}$$

where

$$C_z = \max(e^{-z}, e^{-(1-\beta\Delta)z}). \tag{4.72}$$

Now from (4.1), the second term above vanishes as $n\to\infty$. So we turn our attention to the first term. From Lemma 4.14, there exists a $C>0$ such that for small enough β,

$$\mathbb{E}[Z_i|X_i=1] = \sum_{j\in C_i}\mathbb{P}[X_j=1|X_i=1] \tag{4.73}$$

$$\leq |C_i|\left(m_{t^*} + Ce^{-b_n}\right) \tag{4.74}$$

$$\leq |C_i|\left(\frac{C_z}{n} + \frac{C}{n}\right) \tag{4.75}$$

which goes to zero as $n \to \infty$. □

4.4 Additional lemmas

The first of our additional lemmas comes from [13, Lemma 3.1]. We have made our statement of the lemma slightly more precise and so we have rewritten both the lemma and proof out here along with our modifications. In particular, we have specified precisely how small β must be for the statement to hold.

The lemma concerns two quantities, $\chi(\mathcal{H}_A)$ and $\mathcal{L}(\mathcal{H}_A)$ which in some sense measure the horizontal and vertical size of $\mathcal{H}_A$ respectively. Define

$$\chi(\mathcal{H}_i) = \#\{((u,t),(v,t)) \in \mathcal{H}_i\} \tag{4.76}$$

which counts the total number of spatial edges in $\mathcal{H}_i$ and define

$$\mathcal{L}(\mathcal{H}_i) = \sum_{i \in V} \int_0^{t^*} I_{(i,t)\in\mathcal{H}_i}\,\mathrm{d}t \tag{4.77}$$

which is the sum of the lengths of all the temporal edges in $\mathcal{H}_i$.

Lemma 4.12 ([13]). *For any $0 \le \eta < 1$, $\lambda \in \mathbb{R}$, $\alpha > -\ln(1-\eta)$, if*

$$\tanh(\beta\Delta) \le \frac{1-\eta-e^{-\alpha}}{e^{(\alpha+\lambda)\Delta} - e^{-\alpha}}, \tag{4.78}$$

then for any $A \subseteq V$,

$$\mathbb{E}[\exp(\lambda\chi(\mathcal{H}_A) + \eta\mathcal{L}(\mathcal{H}_A))] \le \exp(\alpha|A|). \tag{4.79}$$

Proof. We first relax our histories to our alternative construction by observing that

$$\chi(\mathcal{H}_i) \le \chi(\hat{\mathcal{H}}_i), \qquad \mathcal{L}(\mathcal{H}_i) \le \mathcal{L}(\hat{\mathcal{H}}_i). \tag{4.80}$$

Let $W_s = |\hat{\mathcal{H}}_i(t^* - s)|$, let $Y_s = \chi(\hat{\mathcal{H}}_i \cap V \times [t^* - s, t^*])$ count the total number of spatial edges observed in the history by time $t^* - s$ and let $Z_s = \mathcal{L}(\hat{\mathcal{H}}_i \cap V \times [t^* - s, t^*])$.

Initially, $W_0 = 1$, $Y_0 = 0$, and $Z_0 = 0$. Recall that an oblivious update of a vertex causes it to be removed from the history and that a non-oblivious update causes the history to branch out to its Δ neighbours. Oblivious updates occur at rate θW_s and cause W_s to decrease by 1. Non-oblivious updates occur at rate $(1-\theta)W_s$ and cause both W_s and Y_s to increase by no more than Δ. The length, Z_s, grows as $\mathrm{d}Z_s = W_s\,\mathrm{d}s$. Therefore we can create a coupled process $(\bar{W}_s, \bar{Y}_s, \bar{Z}_s)$ such that $\bar{W}_s \geq W_s$, $\bar{Y}_s \geq Y_s$, and $\bar{Z}_s \geq Z_s$ in the following way. We start with $(\bar{W}_s, \bar{Y}_s, \bar{Z}_s) = (|A|, 0, 0)$ and at rate $\theta \bar{W}_s$, $\bar{W}_s$ decreases by 1; at rate $(1-\theta)\bar{W}_s$, both $\bar{W}_s$ and $\bar{Y}_s$ increase by Δ; and $\bar{Z}_s$ grows as $\mathrm{d}\bar{Z}_s = \bar{W}_s \,\mathrm{d}s$.

Let $Q_s = \exp\left(\alpha \bar{W}_s + \lambda \bar{Y}_s + \eta \bar{Z}_s\right)$ where α, λ, and η are some fixed constants yet to be determined, and $\alpha > -\ln(1-\eta)$. We have

$$\frac{d}{ds}\mathbb{E}\left[Q_s|Q_{s_0}\right]\bigg|_{s=s_0} = \left(\eta + \theta(e^{-\alpha}-1) + (1-\theta)(e^{(\alpha+\lambda)\Delta}-1)\right)\bar{W}_{s_0}Q_{s_0} \tag{4.81}$$

which is non-positive when

$$\theta \geq \frac{\eta + e^{(\alpha+\lambda)\Delta} - 1}{e^{(\alpha+\lambda)\Delta} - e^{-\alpha}} \tag{4.82}$$

or in terms of the inverse temperature, β,

$$\tanh(\beta\Delta) \leq \frac{1-\eta-e^{-\alpha}}{e^{(\alpha+\lambda)\Delta} - e^{-\alpha}}. \tag{4.83}$$

Hence Q_s is a supermartingale when (4.83) holds. Define the stopping time

$$\tau = \inf\{s : \bar{W}_s = 0\}. \tag{4.84}$$

At this time, the histories have completely died out and $\bar{Y}_s$ and $\bar{Z}_s$ cannot grow any more. That is, $\bar{Y}_\tau \geq \chi(\mathcal{H}_A)$ and $\bar{Z}_\tau \geq \mathcal{L}(\mathcal{H}_A)$. From optional stopping,

$$\mathbb{E}[\exp(\lambda \bar{Y}_\tau + \eta \bar{Z}_\tau)] \leq \mathbb{E}[Q_0] \tag{4.85}$$

$$= \exp(\alpha|A|). \tag{4.86}$$

□

This next Lemma is based on [14, Lemma 2.1] and bounds the speed at which the histories can spread through the graph.

Lemma 4.13. *Let $B(i,l)$ indicate the set of vertices at distance l or smaller from vertex i. The probability that the history of vertex i escapes $B(i,l)$ in time s is bounded by*

$$\mathbb{P}\left[\bigcup_{u\in[0,s]} \mathcal{H}_i(t^*-u) \not\subseteq B(i,l)\right] \leq \exp\left(s\Delta^2 - l\ln\Delta\right). \tag{4.87}$$

Proof. Let $\mathcal{W} = \{\boldsymbol{w} = (w_1, w_2, \ldots, w_l) : w_1 = i, ||w_{k-1} - w_k|| = 1\}$ be the set of length l sequences of adjacent vertices starting at vertex i. If $\mathcal{H}_i$ contains any vertex outside $B(i,l)$ at a time $u \in [t^*-s, t^*]$ then there must be some sequence $w \in \mathcal{W}$ such that each w_i was updated at some time $t^* > t_i > t^* - s$ and $t_{k-1} > t_k$. Call this event M_w. For any particular sequence w,

$$\mathbb{P}[M_w] = \mathbb{P}[\text{Po}(s) \geq l] \tag{4.88}$$

where $\text{Po}(s)$ is Poisson with rate s. By a union bound over $\mathcal{W}$,

$$\mathbb{P}\left[\bigcup_{u \in [0,s]} \mathcal{H}_i(t^* - u) \nsubseteq B(i,l)\right] \leq \Delta^{l-1}\mathbb{P}[\text{Po}(s) \geq l]. \tag{4.89}$$

The moment generating function of a Poisson random variable with rate s is

$$M(t) = \exp\left(s\left(e^t - 1\right)\right). \tag{4.90}$$

Using a Chernoff bound we have for every $t > 0$,

$$\mathbb{P}[\text{Po}(s) \geq l] \leq \exp\left(s\left(e^t - 1\right) - tl\right). \tag{4.91}$$

Overall we have

$$\mathbb{P}\left[\bigcup_{u \in [0,s]} \mathcal{H}_i(t^* - u) \nsubseteq B(i,l)\right] \leq \Delta^{l-1}\exp\left(s\left(e^t - 1\right) - tl\right) \tag{4.92}$$

$$\leq \exp\left(s\left(e^t - 1\right) + l(\ln\Delta - t)\right). \tag{4.93}$$

Choosing $t = 2\ln\Delta$,

$$\mathbb{P}\left[\bigcup_{u \in [0,s]} \mathcal{H}_i(t^* - u) \nsubseteq B(i,l)\right] \leq \exp\left(s\left(\Delta^2 - 1\right) - l\ln\Delta\right) \tag{4.94}$$

$$\leq \exp\left(s\Delta^2 - l\ln\Delta\right). \tag{4.95}$$

□

Lemma 4.14. *There exists a constant $C > 0$ such that for sufficiently small β,*

$$\mathbb{P}[X_j = 1 | X_i = 1] \leq m_{t^*} + Ce^{-k}. \tag{4.96}$$

where $k = |i - j|$ is the distance between vertices i and j.

Proof. There are two ways in which the update history of vertex j can survive until time 0. The update history can survive without intersecting with the update history of vertex i or the update history of vertex j can merge with the update history of vertex i

(whose survival we are conditioning on). Breaking up the probability this way we have

$$\begin{aligned}\mathbb{P}[X_j = 1|X_i = 1] = \mathbb{P}[X_j = 1, \mathcal{H}_i \cap \mathcal{H}_j = \emptyset|X_i = 1] \\ + \mathbb{P}[X_j = 1, \mathcal{H}_i \cap \mathcal{H}_j \neq \emptyset|X_i = 1] \quad (4.97)\\ \leq \mathbb{P}[X_j = 1, \mathcal{H}_i \cap \mathcal{H}_j = \emptyset|X_i = 1] + \mathbb{P}[\mathcal{H}_i \cap \mathcal{H}_j \neq \emptyset|X_i = 1]. \quad (4.98)\end{aligned}$$

The result follows from Lemmas 4.15 and 4.16. □

Lemma 4.15. *Let i and j be the indices of two vertices on a vertex transitive graph. Then*

$$\mathbb{P}[X_j = 1, \mathcal{H}_i \cap \mathcal{H}_j = \emptyset|X_i = 1] \leq m_{t^*} \quad (4.99)$$

Proof. To begin

$$\begin{aligned}\mathbb{P}[X_j = 1, \mathcal{H}_i \cap \mathcal{H}_j = \emptyset|X_i = 1] = \mathbb{P}[X_i = 1]^{-1}\mathbb{P}[X_i = 1, X_j = 1, \mathcal{H}_i \cap \mathcal{H}_j = \emptyset] \quad (4.100)\\ = \frac{\mathbb{P}[\mathcal{H}_i(0) \cup \mathcal{H}_j(0) \neq \emptyset, \mathcal{H}_i \cap \mathcal{H}_j = \emptyset]}{\mathbb{P}[\mathcal{H}_i(0) \neq \emptyset]}. \quad (4.101)\end{aligned}$$

Define S to be the first time that the histories intersect, or define $S = 0$ if the histories do not intersect before time 0. We construct a new history, $\mathcal{H}'_j$, in the following way. Along the interval $[S, t^*)$, $\mathcal{H}'_j$ is constructed from the update sequence in the exact same way as $\mathcal{H}_j$. However, along the interval $[0, S)$, we replace the update sequence with an another i.i.d. copy of the update sequence. We have constructed $\mathcal{H}'_j$ to be independent of $\mathcal{H}_i$. This is obvious for $t < S$. For $t \geq S$ we note that no updates are a part of both histories. From the memoryless property of the exponential waiting times between updates, we get that each update is independent of the rest, and so the histories are independent too. Note that in the event $\{\mathcal{H}_i \cap \mathcal{H}_j = \emptyset\}$, we have $\mathcal{H}_j = \mathcal{H}'_j$ and so

$$\begin{aligned}\mathbb{P}[\mathcal{H}_i(0) \cup \mathcal{H}_j(0) \neq \emptyset, \mathcal{H}_i \cap \mathcal{H}_j = \emptyset] = \mathbb{P}[\mathcal{H}_i(0) \cup \mathcal{H}'_j(0) \neq \emptyset, \mathcal{H}_i \cap \mathcal{H}'_j = \emptyset] \quad (4.102)\\ \leq \mathbb{P}[\mathcal{H}_i(0) \cup \mathcal{H}'_j(0) \neq \emptyset] \quad (4.103)\\ = \mathbb{P}[\mathcal{H}_i(0) \neq \emptyset]\mathbb{P}[\mathcal{H}'_j(0) \neq \emptyset] \quad (4.104)\end{aligned}$$

since $\mathcal{H}_i$ and $\mathcal{H}'_j$ are independent. Since $\mathbb{P}[\mathcal{H}'_j(0) \neq \emptyset] = \mathbb{P}[\mathcal{H}_j(0) \neq \emptyset] = m_{t^*}$ we get the desired result. □

Lemma 4.16 contains some similarities to the proof contained in [13, Lemma 2.1]. Indeed, the quantity

$$\mathbb{P}[\{\mathcal{H}_i \cap \mathcal{H}_j \neq \emptyset\} \cap \{\mathcal{H}_i(0) \cup \mathcal{H}_j(0) \neq \emptyset\}] \quad (4.105)$$

which appears in (4.111) below, is equivalent to the expression

$$\mathbb{P}[A \in \mathrm{RED}_A^*] \tag{4.106}$$

when $A = \{i, j\}$ using the notation of that paper. We use their method to bound this probability, but add some extra steps for clarity.

Lemma 4.16. *Let i and j be the indices of two vertices separated by distance k. Then there exists a C such that for sufficiently small β,*

$$\mathbb{P}[\mathcal{H}_i \cap \mathcal{H}_j \neq \emptyset | X_i = 1] \leq Ce^{-k}. \tag{4.107}$$

Proof. Firstly,

$$\mathbb{P}[\mathcal{H}_i \cap \mathcal{H}_j \neq \emptyset | X_i = 1] = \frac{\mathbb{P}[\{\mathcal{H}_i \cap \mathcal{H}_j \neq \emptyset\} \cap \{X_i = 1\}]}{\mathbb{P}[X_i = 1]} \tag{4.108}$$

$$\leq m_{t^*}^{-1} \mathbb{P}[\{\mathcal{H}_i \cap \mathcal{H}_j \neq \emptyset\} \cap \{X_i = 1\}] \tag{4.109}$$

$$\leq m_{t^*}^{-1} \mathbb{P}[\{\mathcal{H}_i \cap \mathcal{H}_j \neq \emptyset\} \cap (\{X_i = 1\} \cup \{X_j = 1\})] \tag{4.110}$$

$$= m_{t^*}^{-1} \mathbb{P}[\{\mathcal{H}_i \cap \mathcal{H}_j \neq \emptyset\} \cap \{\mathcal{H}_i(0) \cup \mathcal{H}_j(0) \neq \emptyset\}] \tag{4.111}$$

since

$$\{X_i = 1\} = \{\mathcal{H}_i(0) \neq \emptyset\}\,. \tag{4.112}$$

Proceeding backwards from t^*, define S to be the random time at which $\mathcal{H}_i(t) \cup \mathcal{H}_j(t)$ first reduced to less than two vertices, or define $S = 0$ if the combined histories contain at least two vertices all the way to time 0. Note that

$$\{\mathcal{H}_i(0) \cup \mathcal{H}_j(0) \neq \emptyset\} \subseteq \{\mathscr{F}(v, 0, S) \neq \emptyset\} \tag{4.113}$$

where v is either the single vertex $v = \mathcal{H}_i(S) \cup \mathcal{H}_j(S)$ in the case that the histories coalesce to a single point at S, or any arbitrary single vertex otherwise. We also note that

$$\{\mathcal{H}_i \cap \mathcal{H}_j \neq \emptyset\} \subseteq \{\chi\left((\mathcal{H}_i \cup \mathcal{H}_j) \cap V \times [S, t^*]\right) \geq k - 1\} \tag{4.114}$$

since there must be at least $k - 1$ branching edges for the histories to meet. The event on the right hand side of (4.114) depends only on the update sequence in $[S, t^*]$. The event on the right hand side of (4.113) depends only on the update sequence in $[0, S)$.

Therefore, given S, these events are independent and

$$\begin{aligned}
&\mathbb{P}[\{\mathcal{H}_i \cap \mathcal{H}_j \neq \emptyset\} \cap \{\mathcal{H}_i(0) \cup \mathcal{H}_j(0) \neq \emptyset\} \,|S = t_s] \\
&\quad \leq \mathbb{P}[\mathscr{F}(v, 0, S) \neq \emptyset | S = t_s] \mathbb{P}[\chi\left((\mathcal{H}_i \cup \mathcal{H}_j) \cap V \times [S, t^*]\right) \geq k-1 | S = t_s] && (4.115) \\
&\quad = m_{t_s} \mathbb{P}[\chi\left((\mathcal{H}_i \cup \mathcal{H}_j) \cap V \times [S, t^*]\right) \geq k-1 | S = t_s] && (4.116)
\end{aligned}$$

and so

$$\begin{aligned}
\mathbb{P}[\{\mathcal{H}_i \cap \mathcal{H}_j \neq \emptyset\} \cap \{\mathcal{H}_i(0) \cup \mathcal{H}_j(0) \neq \emptyset\}] &\leq \mathbb{E}[I_{\chi((\mathcal{H}_i \cup \mathcal{H}_j) \cap V \times [S,t^*]) \geq k-1} m_S] && (4.117) \\
&\leq \mathbb{E}[I_{\chi(\mathcal{H}_i \cup \mathcal{H}_j) \geq k-1} m_S]. && (4.118)
\end{aligned}$$

From Corollary 4.4,

$$m_S \leq e^{t^* - S} m_{t^*} \tag{4.119}$$

and since $|\mathcal{H}_i(t) \cup \mathcal{H}_j(t)| \geq 2$ for $t \in (S, t^*]$,

$$t^* - S \leq \mathcal{L}(\mathcal{H}_i \cup \mathcal{H}_j)/2. \tag{4.120}$$

So

$$\begin{aligned}
\mathbb{P}[\mathcal{H}_i \cap \mathcal{H}_j \neq \emptyset | X_i = 1] &\leq m_{t^*}^{-1} m_{t^*} \mathbb{E}[I_{\chi(\mathcal{H}_i \cup \mathcal{H}_j) \geq k-1} e^{\mathcal{L}(\mathcal{H}_i \cup \mathcal{H}_j)/2}] && (4.121) \\
&\leq \mathbb{E}[e^{\chi(\mathcal{H}_i \cup \mathcal{H}_j) - (k-1)} e^{\mathcal{L}(\mathcal{H}_i \cup \mathcal{H}_j)/2}] && (4.122) \\
&= e^{-k+1} \mathbb{E}[e^{\chi(\mathcal{H}_i \cup \mathcal{H}_j) + L(\mathcal{H}_i \cup \mathcal{H}_j)/2}] && (4.123)
\end{aligned}$$

From Lemma 4.12, for any $\alpha > \ln(2)$ if

$$\tanh(\beta \Delta) \leq \frac{1 - 2e^{-\alpha}}{2(e^{\Delta(\alpha+1)} - e^{-\alpha})} \tag{4.124}$$

then

$$\mathbb{E}[e^{\chi(\mathcal{H}_i \cup \mathcal{H}_j) + L(\mathcal{H}_i \cup \mathcal{H}_j)/2}] \leq e^{2\alpha} \tag{4.125}$$

and we get the desired result by choosing $C = \exp(1 + 2\alpha)$. □

Part II

Efficient Optimization for Statistical Inference

Chapter 5

Maximum Likelihood Location Mixtures

5.1 Introduction

Mixtures of distributions have been used to model a wide variety of phenomena, with successful applications in the fields of "astronomy, biology, genetics, medicine, psychiatry, economics, engineering, and marketing, among many other fields in the biological, physical, and social sciences" [25, Section 1.1.1]. Mixture models have been in use for over 100 years. In 1984, Pearson used a mixture of two normal densities to model the distribution of the ratio of forehead to body lengths of a sample of 1000 crabs [26]. His mixture of two normals was able to account for the skewness in the data, which a single normal could not model. Pearson suggested that this signalled the existence of two sub-populations of crabs; each associated with its own normal distribution.

However, we do not require that data comes from a mixture of distributions in any physical sense for mixtures to be a useful modelling tool. One of the traits that has contributed to the extent of the use of mixtures is that they provide a convenient semi-parametric way of modelling unknown distributions. They are particularly useful in situations where a parametric method is too restrictive to satisfactorily model the data, and a fully non-parametric method, such as kernel density estimation, may require evaluating a sum which contains more terms than desired. By way of illustration, Priebe in [27] discussed modelling a log normal density using a mixture of normals. With $n = 10000$ observations, Priebe only required about 30 normals to obtain a good approximation. This is in contrast to a kernel density estimator which in this case would be essentially a mixture of 10000 normals.

A *mixture density* is a probability density function that can be written in the form

$$f_Q(\boldsymbol{x}) = \int_\Omega f(\boldsymbol{x};\boldsymbol{\theta})\,\mathrm{d}Q(\boldsymbol{\theta}) \tag{5.1}$$

where $f(\boldsymbol{x};\boldsymbol{\theta})$ is the *component density*, parametrised by $\boldsymbol{\theta} \in \Omega$, and Q is a probability distribution on Ω, called the *mixing distribution*. In the case that Q is a discrete probability distribution, with probability masses p_j at points $\boldsymbol{\theta}_j$, $j = 1, \ldots, m$, the mixture density in (5.1) is a *finite mixture* and can be written

$$f_Q(\boldsymbol{x}) = f_{\boldsymbol{\theta},\boldsymbol{p}}(\boldsymbol{x}) = \sum_{j=1}^{m} p_j f(\boldsymbol{x};\boldsymbol{\theta}_j). \tag{5.2}$$

In this chapter, we will be concerned only with finite mixtures on the real line, whose component densities are parametrised by a single shifting parameter, θ. This is what we will call a *location mixture*, and it can be written as

$$f_{\boldsymbol{p},\boldsymbol{\theta}} = \sum_{j=1}^{m} p_j f(x - \theta_j) \tag{5.3}$$

for a single component density $f(x)$. Looking only at mixtures of this form is not as restrictive as it may at first seem. Using a finite location mixture of normals, you can approximate any continuous density arbitrarily well [BETTER CITATION] [25].

A common question when using mixtures is the following. Let $\boldsymbol{X} = (X_1, \ldots, X_n)$ be a random sample of size n, where X_i has probability density function $g(\boldsymbol{x})$. Given $\boldsymbol{x} = (x_1, \ldots, x_n)$, an observed random sample of $\boldsymbol{X}$, and a component density $f(\boldsymbol{x};\boldsymbol{\theta})$, how do we choose the mixing distribution Q so that $f_Q(\boldsymbol{x})$ is a good approximation for $g(\boldsymbol{x})$? One answer to this question is to choose Q to maximise the likelihood.

The *likelihood* of a distribution Q given $\boldsymbol{x}$ is

$$L(Q;\boldsymbol{x}, f) = \prod_{i=1}^{n} f_Q(x_i). \tag{5.4}$$

Maximizing $L(Q;\boldsymbol{x}, f)$ may be achieved by instead maximizing the *log likelihood*,

$$l(Q;\boldsymbol{x}, f) = \log L(Q;\boldsymbol{x}, f) = \sum_{i=1}^{n} \log f_Q(x_i), \tag{5.5}$$

since $\log(x)$ is a strictly increasing function. In [28], Lindsay showed that so long as the likelihood is bounded, there exists a maximizing Q that has at most n points of support. This is a useful result because it means that the problem of finding a maximizing Q is an optimization problem having finite dimensions. It also justifies our decision to only consider finite mixtures in this chapter.

The size of the support of the Q that maximizes (5.5) is the primary quantity of interest for this Chapter. We define it as follows. Let $\mathcal{Q}_n$ be the set of all discrete probability distributions on $\mathbb{R}$ with no more than n points of support. Let

$$\hat{Q}_{\boldsymbol{x},f} = \{Q \in \mathcal{Q}_n : \forall Q' \in \mathcal{Q}_n, l(Q;\boldsymbol{x}) \geq l(Q';\boldsymbol{x})\} \tag{5.6}$$

be the set of all global maximizers of (5.5). Then define

$$K_{\boldsymbol{x}} = K(\boldsymbol{x}; f) = \min\{m \leq n : \mathcal{Q}_m \cap \hat{Q}_{\boldsymbol{x},f} \neq \emptyset\} \tag{5.7}$$

which counts the smallest number of probability masses needed to maximize the likelihood.

There are a few considerations to be made here. The first is that there is not necessarily a unique mixing distribution that maximizes (5.5) and so we have defined the quantity of interest as the smallest number of points of support required for a maximizing distribution. However, in many cases, including most cases where the component density is from the exponential family, we know that the maximizing mixing distribution is unique. This question of uniqueness was addressed by Lindsay in [28] and in [29] for univariate component densities in the exponential family. A generalization of these results can be found in [30].

The second, is that our choice to only consider location mixtures is significant in that it guarantees the likelihood to be bounded so long as the component density is bounded. In general, this is not the case. Consider the likelihood function for a mixture of normals parametrized by both the mean and variance. We can create a mixture by placing equally weighted normals with very small variance at each x_i in our sample. This corresponds to a mixing distribution Q which places probability mass $1/n$ at points (x_i, σ^2) for $i = 1, \ldots, n$. As σ^2 approaches zero, the density at each x_i, $f_Q(x_i)$, increases without bound and so the likelihood is unbounded. There are a number of techniques that can be employed to ensure the likelihood is bounded such as restricting the parameter space appropriately or restricting the number of components to some number that grows slowly with n (the sieve method [31]).

The third, is that we should be careful to make a distinction between calculating $K_{\boldsymbol{x}}$ and choosing an appropriate number of components for a mixture model. The former is simply a deterministic property of an optimization problem whereas the latter is an in-depth and difficult problem which involves consideration of properties such as consistency, and the use of techniques such as information criteria and likelihood ratio. This latter problem is discussed in [25, Chapter 6]. In particular, if each X_i comes from a distribution with density $g(x) = f_Q(x)$ (that is, $g(x)$ is a mixture density itself) then we may consider the 'true' number of components and using certain penalised log likelihood criteria, such as the Akaike information criterion (AIC) [32] or the Bayesian information criterion (BIC) [33], does not underestimate this quantity asymptotically [34]. This suggests that we should not expect $K_{\boldsymbol{x}}$ (which is defined from an unpenalised likelihood) to necessarily be reflective of the properties of $g(x)$.

5.2 Summary of Lindsay

To obtain results concerning $K_{\boldsymbol{x}}$ we will make use of the geometrical approach employed by Lindsay in [28] and [29] and summarized later in [35, Chapter 5]. This section is dedicated to laying out the basics of this approach and summarising the results that are most relevant to our work. Here, we assume that the component densities, $f(\boldsymbol{x};\boldsymbol{\theta})$, are bounded, but we do not restrict ourselves to finite location mixtures on the real line.

5.2.1 The likelihood curve

The key to Lindsay's approach is to reformulate the problem from optimizing over all mixing distributions, to optimizing an appropriate objective function over a convex set in $\mathbb{R}^n$. Define the *likelihood vector*

$$\boldsymbol{\gamma}(Q;\boldsymbol{x},f) = (f_Q(x_1),\ldots,f_Q(x_n)) \tag{5.8}$$

and the objective function

$$\mathcal{L}(\boldsymbol{\gamma}) = \sum_{i=1}^{n} \ln(\gamma_i). \tag{5.9}$$

Let

$$\mathcal{M} = \{\boldsymbol{\gamma}(Q;\boldsymbol{x},f) : Q \text{ is a probability distribution}\} \tag{5.10}$$

be the set of all possible values of the likelihood vector. A maximizing mixing distribution may be found by first finding the $\hat{\gamma} \in \mathcal{M}$ that maximizes $\mathcal{L}(\boldsymbol{\gamma})$ and then solving the n equations

$$\boldsymbol{\gamma}(Q;\boldsymbol{x},f) = \hat{\gamma} \tag{5.11}$$

for Q.

To show that $\mathcal{M}$ is a convex set, consider the *unicomponent likelihood vector*

$$\boldsymbol{\gamma}(\boldsymbol{\theta}; \boldsymbol{x}, f) = (f_{\boldsymbol{\theta}}(x_1), \ldots, f_{\boldsymbol{\theta}}(x_n)) \tag{5.12}$$

where $f_{\boldsymbol{\theta}}$ is the mixture density corresponding to the mixing distribution that places all its mass at $\boldsymbol{\theta}$, that is,

$$f_{\boldsymbol{\theta}}(x) = f(x; \boldsymbol{\theta}). \tag{5.13}$$

For any probability distribution, Q, the likelihood vector can be represented by

$$\boldsymbol{\gamma}(Q; \boldsymbol{x}, f) = \int \boldsymbol{\gamma}(\boldsymbol{\theta}; \boldsymbol{x}, f) \, \mathrm{d}Q(\boldsymbol{\theta}) \tag{5.14}$$

which for a finite mixture Q, with weights p_j assigned to parameters $\boldsymbol{\theta}_j$, can be written as

$$\boldsymbol{\gamma}(Q; \boldsymbol{x}, f) = \sum_{j=1}^{m} p_j \boldsymbol{\gamma}(\boldsymbol{\theta}_j; \boldsymbol{x}, f). \tag{5.15}$$

This leads to an alternative characterisation of $\mathcal{M}$. Define the *unicomponent likelihood curve*

$$\Gamma_{\boldsymbol{x},f} = \{\boldsymbol{\gamma}(\boldsymbol{\theta}; \boldsymbol{x}, f) : \boldsymbol{\theta} \in \Omega\}. \tag{5.16}$$

Then

$$\mathcal{M} = \mathrm{conv}(\Gamma_{\boldsymbol{x},f}), \tag{5.17}$$

where we use $\mathrm{conv}(A)$ to denote the convex hull of A. We now state the following Theorem taken directly from [35, Theorem 18] which is a consequence of the convexity of $\mathcal{M}$, and the concavity of $\mathcal{L}(\boldsymbol{\gamma})$.

Theorem 5.1 ([35]). *Suppose that $\Gamma_{\boldsymbol{x},f}$ is closed and bounded and that $\mathcal{M}$ contains at least one point with positive likelihood. Then there exists a unique $\hat{\boldsymbol{\gamma}} \in \partial\mathcal{M}$, the boundary of $\mathcal{M}$, such that $\hat{\boldsymbol{\gamma}}$ maximizes $\mathcal{L}(\boldsymbol{\gamma})$ over $\mathcal{M}$.*

In addition to this, in [35, Theorem 21] is stated the following.

Theorem 5.2 ([35]). *The solution $\hat{\boldsymbol{\gamma}}$ can be represented as $\boldsymbol{\gamma}(\hat{Q}; \boldsymbol{x}, f)$, where $\hat{Q}$ has no more than n points of support.*

This gives us our first bound on $K_{\boldsymbol{x}}$: for component densities that produce a closed and bounded likelihood curve,

$$K_{\boldsymbol{x}} \le n. \tag{5.18}$$

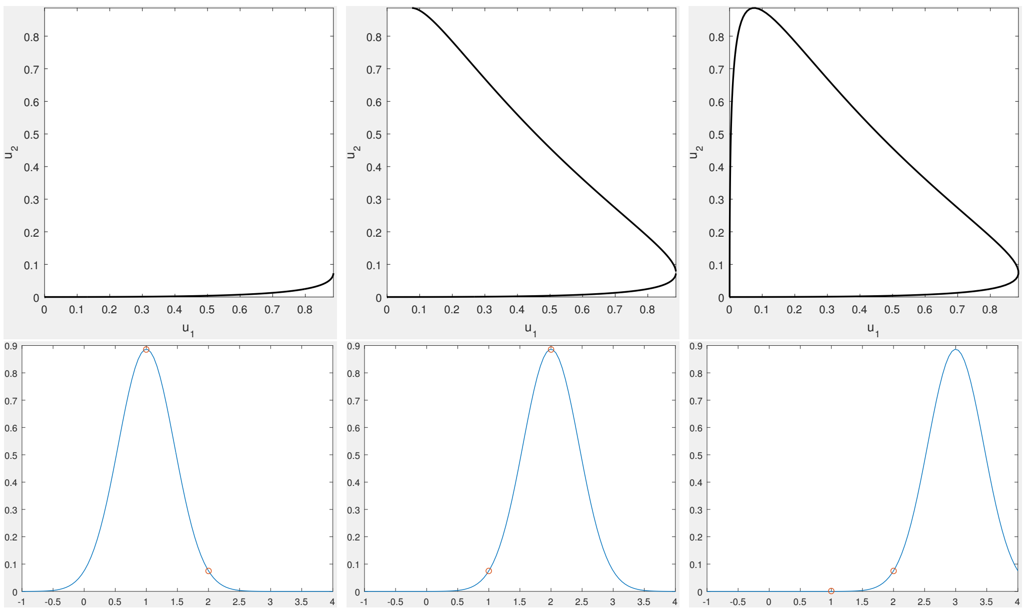
5.2.2 An example likelihood curve

In simple cases, we can plot the likelihood curve, $\Gamma_{\boldsymbol{x},f}$, and objective function, $\mathcal{L}$, along with the maximizing point, $\hat{\boldsymbol{\gamma}}$. In particular, when $n = 2$, $\Gamma_{\boldsymbol{x},f} \subset \mathbb{R}^2$ and if the component density is smoothly parametrized by a single parameter, then we can plot $\Gamma_{\boldsymbol{x},f}$ by tracing out $\boldsymbol{\gamma}(\theta; \boldsymbol{x}, f)$ as we vary θ.

For example, suppose our sample is made up of two points, $\boldsymbol{x} = (x_1, x_2) = (1, 2)$, and suppose our component density is normal with variance $\sigma_2 = 0.45^2$ and parametrized by θ, that is,

$$f(x; \theta) = \frac{1}{0.45\sqrt{2\pi}} \exp\left(-\frac{(x-\theta)^2}{2 \cdot 0.45^2}\right). \tag{5.19}$$

Then the corresponding likelihood curve can be traced out as we increase θ from $-\infty$ to ∞ as shown in Figure 5.1.



(a) $\theta = 1$ (b) $\theta = 2$ (c) $\theta = 3$

Figure 5.1 – *A simple example of a likelihood curve. The component density, $f(x;\theta)$, as defined in (5.19), is shown in blue for $\theta = 1, 2,$ and 3. Each value of θ contributes a point to $\Gamma_{\boldsymbol{x},f}$ whose coordinates are given by $(f(1;\theta), f(2;\theta))$ (represented by the red circles). As we increase θ from $-\infty$ to ∞ we trace out more of $\Gamma_{\boldsymbol{x},f}$.*

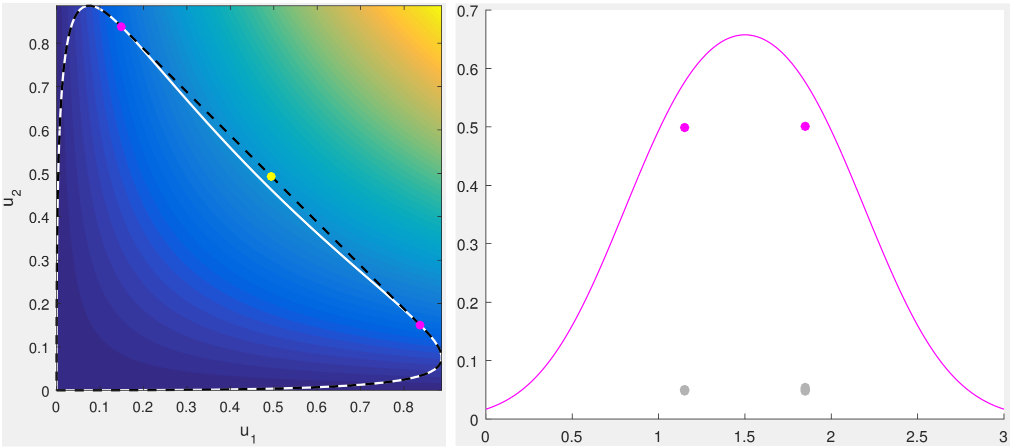
[MAKE THIS FIGURE REPRODUCIBLE?]

The set of possible values of the likelihood vector, $\mathcal{M}$, is given by the convex hull of $\Gamma_{\boldsymbol{x},f}$, the boundary of which is marked along with $\Gamma_{\boldsymbol{x},f}$ in Figure 5.2a and overlaid on a heat map of $\mathcal{L}(\boldsymbol{\gamma})$. The maximizing point, $\hat{\boldsymbol{\gamma}}$, is marked with a yellow dot and lies on the boundary of $\mathrm{conv}(\Gamma_{\boldsymbol{x},f})$ as predicted by Theorem 5.1. This point may be written as the convex combination of two points in $\Gamma_{\boldsymbol{x},f}$,

$$\hat{\boldsymbol{\gamma}} = \sum_{j=1}^{2} p_j \boldsymbol{\gamma}(\theta_j; \boldsymbol{x}, f) \tag{5.20}$$

and we represent the two points, $\boldsymbol{\gamma}(\theta_1; \boldsymbol{x}, f)$ and $\boldsymbol{\gamma}(\theta_2; \boldsymbol{x}, f)$ with magenta dots.

The mixture $\hat{Q}$ that satisfies $\hat{\boldsymbol{\gamma}} = \boldsymbol{\gamma}(\hat{Q}; \boldsymbol{x}, f)$ is the one that places masses p_1 and p_2 at locations θ_1 and θ_2 and it is plotted in Figure 5.2b by two magenta points at (θ_1, p_1) and (θ_2, p_2) along with the overall mixture density $f_{\hat{Q}}(x)$.



(a) **(b)**

Figure 5.2 – *The geometric relationship between the likelihood curve (a) and the maximizing mixture density (b). In (a), the boundary of* $\mathrm{conv}(\Gamma_{\boldsymbol{x},f})$ *is shown as a dashed black line,* $\Gamma_{\boldsymbol{x},f}$ *is the white curve, the heat map shows the objective function (likelihood increases from blue to yellow) and* $\hat{\boldsymbol{\gamma}}$ *is marked with a yellow dot. This point is a convex combination of the two magenta points. These two magenta points correspond to the two probability masses in the maximizing mixing distribution (b).*

[MAKE THIS FIGURE REPRODUCIBLE + FIX]

Remark 5.3. Note that in this example, while $\Gamma_{\boldsymbol{x},f}$ is bounded, it is not closed (it does not contain the limit point $(0, 0)$), counter to the requirements of Theorems 5.1 and 5.2. In fact, any positive density whose support is the whole real line will not contain the limit point $\mathbf{0}$ (where $\mathbf{0}$ represents the zero vector in $\mathbb{R}^n$). However, since $\mathbf{0}$ is clearly not going to be a part of a maximizing mixture, we are safe to apply both theorems if $\Gamma \cup \{\mathbf{0}\}$ is closed. A more detailed discussion concerning how to relax the requirement that the set $\Gamma_{\boldsymbol{x},f}$ be closed can be found in [35, Section 5.2.2.].

5.2.3 Gradient characterization

Another important construction due to Lindsay is the gradient function

$$D_Q(\boldsymbol{\theta}; \boldsymbol{x}, f) = -n + \sum_{i=1}^{n} \frac{f(x_i; \boldsymbol{\theta})}{f_Q(x_i)}. \tag{5.21}$$

This is the derivative of $l(Q; \boldsymbol{x}, f)$ as we move along the path parametrized by

$$(1-p)Q + p\Delta_{\boldsymbol{\theta}} \tag{5.22}$$

evaluated at $p = 0$ and where $\Delta_{\boldsymbol{\theta}}$ is a degenerate distribution that places all of its mass at $\boldsymbol{\theta}$. In [28, Theorem 4.1], Lindsay showed you could characterise the maximizing mixture by three equivalent conditions. The statement of the Theorem here is taken from [35, Theorem 19].

Theorem 5.4 ([35]). *The following three statements are equivalent:*

1. $\hat{Q}$ *maximizes* $l(Q; \boldsymbol{x}, f)$.
2. $\hat{Q}$ *minimizes* $\sup_{\boldsymbol{\theta}} D_Q(\boldsymbol{\theta}; \boldsymbol{x}, f)$.
3. $\sup_{\boldsymbol{\theta}} D_{\hat{Q}}(\boldsymbol{\theta}; \boldsymbol{x}, f) = 0$.

Also contained in [28, Theorem 4.1] was the following result concerning the locations of the support points of $\hat{Q}$. The statement here is taken from [35, Theorem 20].

Theorem 5.5 ([35]). *The support of any maximum likelihood estimator* $\hat{Q}$ *lies in the set*

$$\left\{\boldsymbol{\theta} : D_{\hat{Q}}(\boldsymbol{\theta}; \boldsymbol{x}, f) = 0\right\}. \tag{5.23}$$

If f is parametrized by a single parameter θ and $D_Q(\theta; \boldsymbol{x}, f)$ is twice differentiable in θ, then each interior support point θ^* of the maximizing mixture distribution, $\hat{Q}$, satisfies

$$D_{\hat{Q}}(\theta^*; \boldsymbol{x}, f) = 0, \tag{5.24}$$

$$D'_{\hat{Q}}(\theta^*; \boldsymbol{x}, f) = 0, \tag{5.25}$$

$$D''_{\hat{Q}}(\theta^*; \boldsymbol{x}, f) \leq 0. \tag{5.26}$$

5.2.3.1 Support Hyperplane

There is a geometrical interpretation to these theorems that ties in with the likelihood curve interpretation given above. For any given mixing distribution Q, define the *inverse* *likelihood vector* $\boldsymbol{\gamma}^{-1}(Q;\boldsymbol{x},f) = (1/f_Q(x_1),\ldots,1/f_Q(x_n))$ and the hyperplane

$$\mathcal{H}_Q = \{\boldsymbol{z} : \langle \boldsymbol{\gamma}^{-1}(Q;\boldsymbol{x},f), \boldsymbol{z}\rangle = n\}, \tag{5.27}$$

which contains the usual likelihood vector, $\boldsymbol{\gamma}(Q;\boldsymbol{x},f)$. We may write the gradient function as

$$D(Q;\boldsymbol{x},f) = \langle \boldsymbol{\gamma}^{-1}(Q;\boldsymbol{x},f), \boldsymbol{\gamma}(\boldsymbol{\theta};\boldsymbol{x},f)\rangle - n. \tag{5.28}$$

If $\hat{Q}$ maximizes $l(Q;\boldsymbol{x},f)$ then by statement 3,

$$\langle \boldsymbol{\gamma}^{-1}(\hat{Q};\boldsymbol{x},f), \boldsymbol{\gamma}(\boldsymbol{\theta};\boldsymbol{x},f)\rangle \leq n \tag{5.29}$$

for all $\boldsymbol{\theta} \in \Omega$. This means that $\mathcal{M} = \mathrm{conv}(\Gamma_{\boldsymbol{x},f})$ lies entirely on one side of $\mathcal{H}$ and Theorem 5.5 tell us that if $\boldsymbol{\theta}$ is in the support of $\hat{Q}$, then

$$\langle \boldsymbol{\gamma}^{-1}(\hat{Q};\boldsymbol{x},f), \boldsymbol{\gamma}(\boldsymbol{\theta};\boldsymbol{x},f)\rangle = n \tag{5.30}$$

and so $\boldsymbol{\gamma}(\boldsymbol{\theta};\boldsymbol{x},f) \in \mathcal{H}$. Thus the question of determining the number of support points of $\hat{Q}$ can be answered by finding the number of points at which $\Gamma_{\boldsymbol{x},f}$ touches $\mathcal{H}_{\hat{Q}}$.

[FIGURE]

5.2.4 Additional results on $K_{\boldsymbol{x}}$

Theorem 5.2 bounds $K_{\boldsymbol{x}}$ by n. To get tighter bounds on $K_{\boldsymbol{x}}$, we must make some assumptions about the form of the component densities, or the structure of $\boldsymbol{x}$.

[FOLLOWING IS NOTES FOR MYSELF]

Exponential family [29], more general results including this one and others in [30]. In particular, $n/2$ for discrete densities. May want to cite [36] since it has some very limited results on the support. Read [37] for more citations.

5.3 Empirical Results

In this section we explore $K_{\boldsymbol{x}}$ through empirical results and figures. Throughout this section, and for the remainder of this chapter, we will consider only location mixtures (see (5.3)). All component densities will be unimodal, with a mode at $x = 0$. We list here the component densities that we will use.

Density	$f(x) =$
normal with fixed variance σ^2	$\frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{x^2}{2\sigma^2}\right)$
Cauchy with fixed scale γ	$\frac{1}{\pi\gamma+\pi x^2/\gamma}$

5.3.1 Method

There are two parts to determining $K_{\boldsymbol{x}}$ empirically. The first part is to find the maximum likelihood mixture, $\hat{Q}$, or at least a good approximation Q^*. Given $\boldsymbol{x} \in \mathbb{R}^n$, we use a general purpose nonlinear programming solver to find the maximum likelihood mixture (we use the MATLAB function *fmincon*). We can test that we have indeed reached the global maximum through the use of the derivative function conditions given in Theorem 5.4. If the mixture that is returned by our general solver does not satisfy these conditions then we may use another method, such as the vertex direction method (VDM) or the vertex exchange method (VEM), which is guaranteed to converge to the global maximum but converges slowly (see [38] for a review of different methods). We note that the actual method used is not important so long as the final result satisfies the derivative conditions. In practise, when n is small, a general purpose solver is sufficient, fast, and simple to implement, and we rarely need to use a secondary method.

The second part is to determine the number of points of support of Q^*. We initiate our general purpose solver with a mixture distribution that has more points of support than is needed. Under the optimization, this collapses down to the distribution Q^* which may assign zero weight to some masses, and may concentrate multiple masses on the one point of support. We may simplify such a distribution by removing all masses with zero weight, and by combining all masses which share a point of support into one mass which takes the combined weight of the constituent masses (see Figure 5.3). A naive approach would be to take the number of probability masses in this simplified distribution as the value for $K_{\boldsymbol{x}}$.

However, there are some problems with this approach. In practice, we do not remove masses with exactly zero weight, but rather remove masses with weight $p_j < \epsilon$ for some small $\epsilon > 0$. Similarly, we merge masses i and j with support $|\theta_j - \theta_i| < \delta$ for some small $\delta > 0$. It may be that the true maximum likelihood mixture does contain components that have either very small weight, or are located very close to other masses. In this scenario, we would produce a value for $K_{\boldsymbol{x}}$ that is too small.

Instead, we propose the following. From Theorem 5.5, each point of support, $\hat{\theta}_j$ of the maximizing mixture distribution $\hat{Q}$ is a local maximum of $D_{\hat{Q}}(\theta; \boldsymbol{x}, f)$ with $D_{\hat{Q}}(\theta_j; \boldsymbol{x}, f) = 0$. So, given the mixing distribution Q^* that results from our optimization, we take $K_{\boldsymbol{x}}$

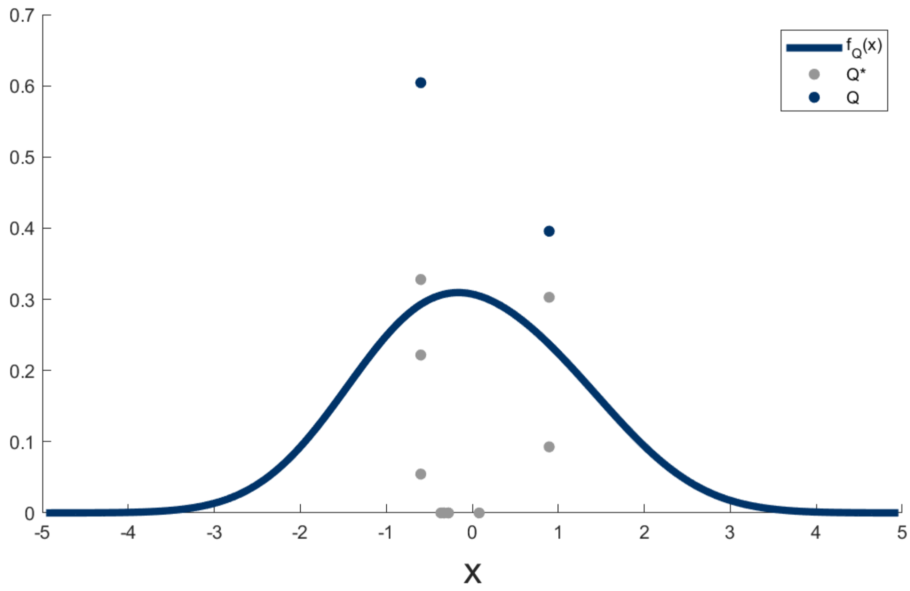


Figure 5.3 – *The unsimplified mixing distribution Q^* that results from using more points of support than required when finding our maximizing mixture, as well as the equivalent simplified distribution, Q, and the resulting mixture density, $f_Q(x)$. Each probability mass with support θ and weight p is represented by a point at (θ, p).*

to be the number of local maximums in $D_{Q^*}(\theta; \boldsymbol{x}, f)$ that take maximum value close to 0.

[WHEN IS SUPPORT ALWAYS EQUAL TO SET OF D(THETA) = 0?]

[EXTEND AS REQUIRED]

5.3.2 Flag graphs

When is n is very small, we can plot $K_{\boldsymbol{x}}$ as a function of $\boldsymbol{x}$. For example, we may take $\boldsymbol{x} = (x_1, x_2)$ across a grid of values and colour each point $\boldsymbol{x}$ according to the value of $K_{\boldsymbol{x}}$. We have done this in Figure 5.4 for a normal component density with fixed variance $\sigma^2 = 1$. We observe a band in which $K_{\boldsymbol{x}} = 1$ and outside of which $K_{\boldsymbol{x}} = 2$.

However, there is some redundancy in this plot which arises because we are using location mixtures. When using a location mixture, if the maximizing mixing distribution for $\boldsymbol{x}$ places weights p_j at locations θ_j, then a maximizing mixing distribution for $\boldsymbol{x}+(c,\dots,c)$ for some constant $c \in \mathbb{R}$ is simply the distribution that places weights p_j at locations $\theta_j + c$. That is, shifting $\boldsymbol{x}$ by c results in the maximizing mixture also shifting by c. A consequence of this fact is that $K_{\boldsymbol{x}}$ is invariant under translations of $\boldsymbol{x}$.

This suggests increases n to 3, and fixing one of the x_i while letting the other two vary. We do this by taking $\boldsymbol{x} = (0, x_2, x_3)$ with x_2 and x_3 varying across an evenly spaced

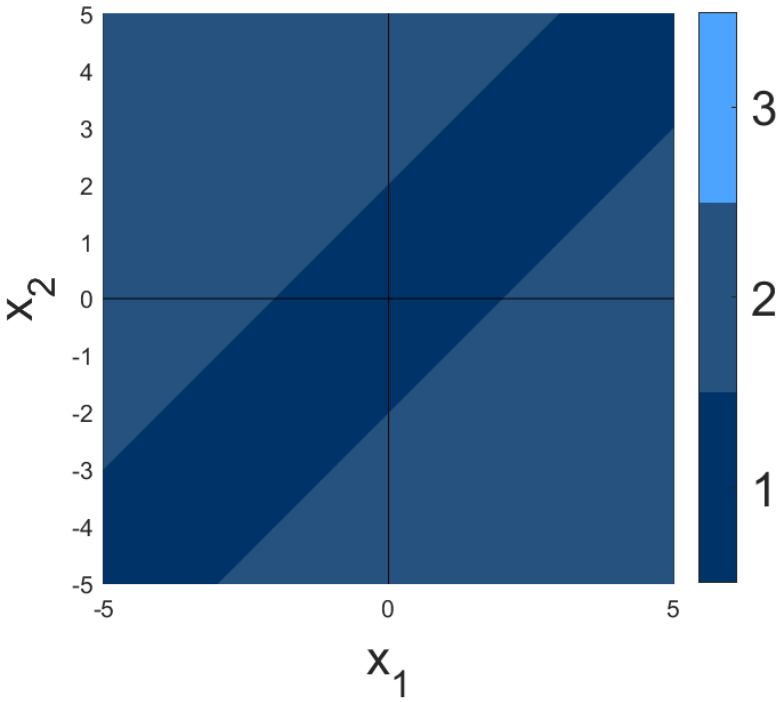


Figure 5.4 – *$K_{\boldsymbol{x}}$ as a function of $\boldsymbol{x} = (x_1, x_2)$, for a normal component density with fixed variance $\sigma^2 = 1$.*

grid. We do this for both a normal component density with variance $\sigma^2 = 1$ (Figure 5.5) and a Cauchy density with scale $\gamma = \sqrt{3}$ (Figure 5.6). We have chosen the scale of the Cauchy so that it has inflection points in the same places as the normal density, namely at $x = \pm 1$. This decision is made in light of Theorem 5.8 which says that for $n = 2$ and for certain component densities f, $K_{\boldsymbol{x}}$ can be determined by comparing $|x_1 - x_2|$ and the distance between the inflection points of f. It seems reasonable to expect that the choice of scale that makes the $n = 2$ figures identical is a good choice for comparing the effects of choosing different component densities in the $n = 3$ case.

While $n = 2, 3$ is not a very realistic scenario when it comes to real data to which we may wish to fit a mixture model, it does help demonstrate a particular way of thinking about the problem of determining $K_{\boldsymbol{x}}$. That is, that $K_{\boldsymbol{x}}$ is simply a function of where $\boldsymbol{x}$ lies in $\mathbb{R}^n$. For a particular choice of component density, we can partition $\mathbb{R}^n$ into sets

$$C_k = \{\boldsymbol{x} \in \mathbb{R}^n | K_{\boldsymbol{x}} = k\}, \qquad k = 1, \dots, n. \tag{5.31}$$

The problem of determining $K_{\boldsymbol{x}}$ is then the same as determining in which of the C_k $\boldsymbol{x}$ lies. If $\boldsymbol{x} = \boldsymbol{X}$ is randomly chosen, then the probability that $K_{\boldsymbol{X}} = k$ is simply the probability that $\boldsymbol{X} \in C_k$. In Section 5.4, we present various results which bound the regions C_k in various settings.

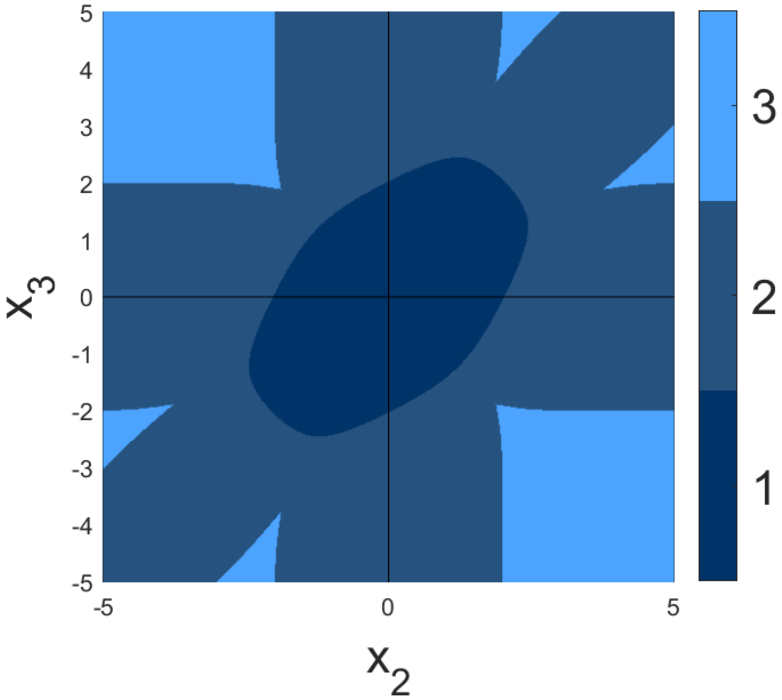


Figure 5.5 – $K_{\boldsymbol{x}}$ *as a function of* $\boldsymbol{x} = (0, x_2, x_3)$ *for a normal component density with fixed variance* $\sigma^2 = 1$.

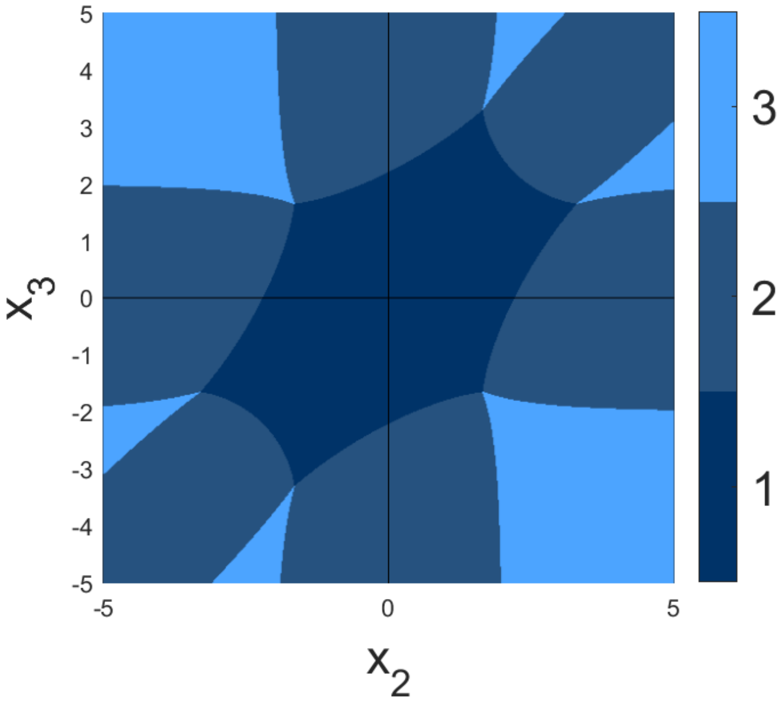


Figure 5.6 – $K_{\boldsymbol{x}}$ *as a function of* $\boldsymbol{x} = (0, x_2, x_3)$ *for a Cauchy component density with fixed scale* $\gamma = \sqrt{3}$.

5.3.3 Other interesting things

We end our empirical results section with

[UP TO HERE]

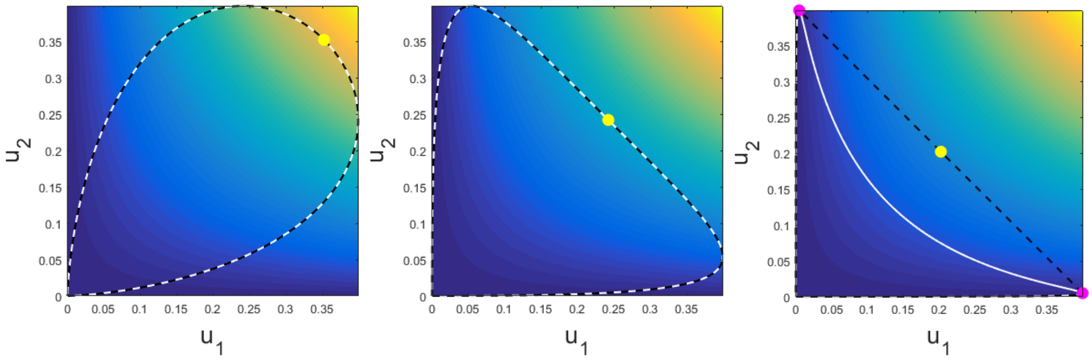
5.4 Results

5.4.1 Results for $n = 2$

In this section, we present Theorem 5.8 which expands upon the results found in [28] and [29].

5.4.2 The likelihood curve for $n = 2$

The shape of $\Gamma_{\boldsymbol{x}}$ can provide us with some insight into the behaviour of $K_{\boldsymbol{x}}$. In Figure 5.7, we give some examples of $\Gamma_{\boldsymbol{x}}$ for $n = 2$ using a normal component density with variance $\sigma^2 = 1$. In particular, we note that the distance between x_1 and x_2 has a



(a) $\boldsymbol{x} = (0, 1)$ **(b)** $\boldsymbol{x} = (0, 2)$ **(c)** $\boldsymbol{x} = (0, 3)$

Figure 5.7 – *The curve* $\Gamma_{\boldsymbol{x}}$ *for three different* $\boldsymbol{x}$ *along with the boundary of* $\text{conv}(\Gamma_{\boldsymbol{x}})$*. The objective function,* $\mathcal{L}(\boldsymbol{\gamma})$*, is represented as a heat map. The optimal point* $\hat{\boldsymbol{\gamma}}$ *is shown in yellow, and where applicable, the points* $\boldsymbol{\gamma}(\theta_j)$ *that make up* $\hat{\boldsymbol{\gamma}}$ *are shown in magenta.*

strong effect on the shape of $\Gamma_{\boldsymbol{x}}$. In Figure 5.7a, the points are distance 1 apart and $\Gamma_{\boldsymbol{x}}$ is the boundary of $\text{conv}(\Gamma_{\boldsymbol{x}})$. In this case, it is clear that $K_{\boldsymbol{x}} = 1$. In Figure 5.7c, the points are distance 3 apart and the optimal point no longer lies on $\Gamma_{\boldsymbol{x}}$. This results in the maximum likelihood mixing distribution needing two points of support and so $K_{\boldsymbol{x}} = 2$. The boundary case, where $\Gamma_{\boldsymbol{x}}$ goes from being a convex curve to having the indentation shown in Figure 5.7c, is shown in Figure 5.7b.

Obtaining results about where these boundaries lie is very difficult in higher dimensions. In [29], Lindsay used the sign of the curvature of $\boldsymbol{\gamma}(\theta; \boldsymbol{x})$ to obtain results for $n = 2$ when the component density is in the exponential family. Here we will present an extension to Lindsay's results by considering densities that satisfy the following assumptions.

A1 (Continuity). $f(x)$ is a continuous density and has the whole real line as its support.

A2 (Differentiability). $f(x)$ is twice differentiable.

A3 (Unimodality). $f(x)$ has a single mode at $x = 0$. I.e. $f'(x) > 0$ for $x < 0$, $f'(0) = 0$, and $f'(x) < 0$ for $x > 0$.

A4 (Symmetry). The density $f(x)$ is symmetric about $x = 0$.

A5. f has only two points of inflection

Definition 5.6. If f satisfies assumptions A1 through to A3, then define $[i^-, i^+]$ to be the largest interval that contains 0 and on which $f''(x) \leq 0$.

That is, i^- and i^+ are inflection points of f. Note that for any f satisfying A4, $i^- = -i^+$. In this case we will write $i = i^+ = -i^-$.

A6. $f'(x) > -f'(x - 2i)$ for $\theta \in (i, \infty)$

Some common densities that satisfy these assumptions include the normal density and the Cauchy density.

Lemma 5.7. *Let $f(x)$ be a density which satisfies assumptions A1 through to A6 and whose inflection points are at $x = i$ and $x = -i$. If $x_2 - x_1 < 2i$ ($x_2 > x_1$) then the equation*

$$-f'(x_1 - \theta) = f'(x_2 - \theta) \tag{5.32}$$

has only one solution.

Proof. We first consider the shape of $f'(x)$. Assumption A3 tells us that $f'(x)$ is positive for $x < 0$ and negative for $x > 0$. The function $f'(x)$ will have turning points at $\pm i$ and from Assumption A5 these will be the only turning points. Hence we have the following picture of $f'(x)$:

$$f'(x) \text{ is } \begin{cases} \text{positive and increasing}, & x \in (-\infty, -i) \\ \text{positive and decreasing}, & x \in (-i, 0) \\ \text{negative and decreasing}, & x \in (0, i) \\ \text{negative and increasing}, & x \in (i, \infty). \end{cases} \tag{5.33}$$

We also note, from A4, that $f'(x)$ is an odd function. Using this and rearranging (5.32) we obtain the equivalent equation

$$g(\theta) = h(\theta) \tag{5.34}$$

where we have put $g(\theta) = f'(\theta)$ and $h(\theta) = -f'(\theta - (x_2 - x_1))$ for ease of notation.

If we assume that $0 < x_2 - x_1 < 2i$ then we can consider possible solutions to (5.34) on each of the following intervals.

For $\theta \in (-\infty, 0]$, $g(\theta) \geq 0$ and $h(\theta) < 0$ and so there are no possible solutions.

Likewise, for $\theta \in [x_2 - x_1, \infty)$, $g(\theta) < 0$ and $h(\theta) \geq 0$ and so there are no possible solutions.

For $\theta \in [-i+x_2-x_1, i]$, $g(\theta)$ is decreasing and $h(\theta)$ is increasing and $h(-i+x_2-x_1) = g(i)$ (since f' is odd). Therefore there must be exactly one solution in this interval.

We note that if $x_2 - x_1 \leq i$ then the above intervals cover the real line. In the case that $i < x_2 - x_1 < 2i$ we need to consider these additional intervals.

For $\theta \in (i, x_2 - x_1)$, from assumption A6, $f'(\theta) < -f'(\theta - 2i) < -f'(\theta - (x_2 - x_1))$ since both $-f'(\theta - 2i)$ and $-f'(\theta - (x_2 - x_1))$ are increasing on this interval. Hence there can be no solutions to (5.34) on this interval.

Similarly by symmetry of f, for $\theta \in (0, -i+x_2-x_1)$, $f'(\theta) > -f'(\theta-2i) > -f'(\theta-x_2-x_1)$ and there are no solutions to (5.34) on this interval either.

Since the above intervals cover the real line and since we have shown that there is only one solution in one of these intervals, (5.32) must have only one solution. □

Theorem 5.8. *Let $f(x)$ satisfy assumptions A1 through to A6. Let $\boldsymbol{x} = (x_1, x_2)$ be the sample for which we a finding a maximum likelihood mixture using f as the component density. Then $K_{\boldsymbol{x}} = 1$ if and only if*

$$x_2 - x_1 \leq 2i \tag{5.35}$$

Proof. By the unimodality of f, the points of support of the maximizing mixing distribution must lie between x_1 and x_2. Hence we need only consider the behaviour of $\boldsymbol{\gamma}(\theta; \boldsymbol{x})$ for $\theta \in [x_1, x_2]$. By the symmetry of f, $\hat{\boldsymbol{u}}$ must lie on the line $u_1 = u_2$[1].

First we complete the only if direction of the proof. Assume that $x_2 - x_1 > 2i$. By the symmetry of f, $\boldsymbol{\gamma}(\theta; \boldsymbol{x})$ crosses the $u_1 = u_2$ line at $\theta = (x_1 + x_2)/2$. Now the curvature of $\boldsymbol{\gamma}$ has sign equal to

$$S(\theta) = \begin{vmatrix} \gamma_1'(\theta; \boldsymbol{x}) & \gamma_1''(\theta; \boldsymbol{x}) \\ \gamma_2'(\theta; \boldsymbol{x}) & \gamma_2''(\theta; \boldsymbol{x}) \end{vmatrix} = \begin{vmatrix} -f'(x_1 - \theta) & f''(x_1 - \theta) \\ -f'(x_2 - \theta) & f''(x_2 - \theta) \end{vmatrix}. \tag{5.36}$$

[1]This is obvious but may need a lemma

and so

$$S\left(\frac{x_1+x_2}{2}\right) = \begin{vmatrix} -f'(\frac{x_1-x_2}{2}) & f''(\frac{x_1-x_2}{2}) \\ -f'(\frac{x_2-x_1}{2}) & f''(\frac{x_2-x_1}{2}) \end{vmatrix}. \tag{5.37}$$

Since $x_2-x_1 > 2i$, $\frac{x_1-x_2}{2} > i$ and so $f''((x_1-x_2)/2) > 0$. Similarly, $f''((x_2-x_1)/2) > 0$. We also have that $-f((x_1-x_2)/2) < 0$ and $-f((x_2-x_1)/2) > 0$. Hence $S((x_1+x_2)/2) < 0$ and so $\boldsymbol{\gamma}((x_1+x_2)/2; \boldsymbol{x})$ has negative curvature. The curve Γ must have positive curvature at the points of support and so we cannot have that $K_{\boldsymbol{x}} = 1$.

Now we complete the if direction. Assume that $x_2 - x_1 \le 2i$. By Lemma 5.7, there is only one point at which the curve is pointing perpendicular to the line $u_1 = u_2$ SAY THIS BETTER. By the symmetry of f this occurs when $\gamma(\theta; \boldsymbol{x})$ is crossing the line $u_1 = u_2$. Since f is continuous, the direction that $\gamma(\theta; \boldsymbol{x})$ is moving is also continuous. At $\theta = x_1$, $\gamma(\theta; \boldsymbol{x})$ is pointing straight up and so we have that for $\theta \in [x_1, (x_1+x_2)/2]$, $\gamma(\theta; \boldsymbol{x})$ is travelling in a direction pointing above the line perpendicular to $u_1 = u_2$. For $\theta \in [(x_1+x_2)/2, x_2]$, $\gamma(\theta; \boldsymbol{x})$ points below the line. It is now obvious that $\gamma((x_1+x_2)/2; \boldsymbol{x})$ is the furthest point from the origin that lies on $u_1 = u_2$ and is in the convex hull of $\Gamma_{\boldsymbol{x}}$. Since the likelihood increases as we move away from the origin along the line $u_1 = u_2$ in the positive quadrant, we must have

$$\hat{\boldsymbol{u}} = \gamma((x_1+x_2)/2; \boldsymbol{x})$$

and so $K_{\boldsymbol{x}} = 1$.

□

5.5 Results for general n

5.5.1 Normal Constraints

When our component density is normal with variance σ^2,

$$f(x;\sigma) = \frac{1}{\sigma\sqrt{2\pi}} e^{-x^2/2\sigma^2}, \tag{5.38}$$

equations (5.24) to (5.26) become

$$\frac{1}{n}\sum_{i=1}^{n}\Gamma_k(x_i;\boldsymbol{p},\boldsymbol{\theta}) = 1 \tag{5.39}$$

$$\frac{1}{n}\sum_{i=1}^{n}x_i\Gamma_k(x_i;\boldsymbol{p},\boldsymbol{\theta}) = \theta_k \tag{5.40}$$

$$\frac{1}{n}\sum_{i=1}^{n}(x_i-\theta_k)^2\Gamma_k(x_i;\boldsymbol{p},\boldsymbol{\theta}) \leq \sigma^2 \tag{5.41}$$

where we have written

$$\Gamma_k(x;\boldsymbol{p},\boldsymbol{\theta}) = \frac{f(x-\theta_k;\sigma)}{\sum_{j=1}^{m} p_j f(x-\theta_j;\sigma)}. \tag{5.42}$$

for ease of notation. Using these constraints, we will bound the regions $C_1,\ldots,C_n$ in Theorem 5.14. However, as a gentle introduction we will start with the much simpler problem of just bounding C_1.

Theorem 5.9. *Write $\bar{\boldsymbol{x}}$ for the mean of $\boldsymbol{x}$. If $\boldsymbol{x}\in C_1$ then*

$$\frac{1}{n}\sum_{i=1}^{n}(x_i-\bar{\boldsymbol{x}})^2 \leq \sigma^2. \tag{5.43}$$

Proof. If $\boldsymbol{x}\in C_1$ then the maximizing mixture has one component and so $\Gamma_1(x;\boldsymbol{\theta},\boldsymbol{p}) = 1$. Then (5.40) gives us that $\theta_1 = \bar{\boldsymbol{x}}$ and combining this with (5.41) completes the proof. □

5.5.1.1 Treating $\boldsymbol{x}$ as random

Up until now, we have treated $\boldsymbol{x}$ as fixed, not random, and treated the maximum likelihood problem purely as an optimization one, rather than a statistical one. However, for this section we make the assumption that $\boldsymbol{x}$ is made up of i.i.d. random variables, x_i, which have distribution

$$x_i \sim N(\mu,\sigma_1^2)$$

for $i=1,\ldots,n$. Our component density, f_θ, is normal with variance σ_2^2. From Theorem 5.9,

$$p_u = \mathbb{P}\left(\sum_{i=1}^{n}(x_i-\bar{\boldsymbol{x}})^2 \leq n\sigma_2^2\right)$$

is an upper bound to $\mathbb{P}(\boldsymbol{x} \in C_1)$. Writing s^2 for the unbiased sample variance of $\boldsymbol{x}$

$$\begin{aligned}
p_u &= \mathbb{P}\left(s^2 \leq \frac{n\sigma_2^2}{n-1}\right) \\
&= \mathbb{P}\left(\frac{(n-1)s^2}{\sigma_1^2} \leq \frac{n\sigma_2^2}{\sigma_1^2}\right) \\
&= \mathbb{P}\left(\chi^2_{n-1} \leq \frac{n\sigma_2^2}{\sigma_1^2}\right)
\end{aligned}$$

where χ^2_{n-1} is chi-squared with $n-1$ degrees of freedom.

Remark 5.10. Of particular interest is the case where $\sigma_1 = \sigma_2$. In this case, $p_u \to 1/2$ as $n \to \infty$. While not a new result [CITE SOMETHING HERE], this tells us that the maximum likelihood estimator is not a consistent estimator for the number of components.

5.5.2 Properties of Γ

In order to bound regions where $m \geq 2$, we will need to get a handle on $\Gamma_k(x; \boldsymbol{\theta}, \boldsymbol{p})$. In this section, we list and prove some properties that will be required in Section 5.5.3.

Lemma 5.11.

$$\max_k \left(\Gamma_k(x; \boldsymbol{p}, \boldsymbol{\theta})\right) \geq 1 \tag{5.44}$$

Proof. For each x, there exists a k_0 such that $f(x - \theta_{k_0}; \sigma) \geq f(x - \theta_k; \sigma)$ for all k. It follows that

$$\Gamma_{k_0}(x; \boldsymbol{p}, \boldsymbol{\theta}) = \frac{f(x - \theta_{k_0}; \sigma)}{\sum_{j=1}^m p_j f(x - \theta_j; \sigma)} \geq \frac{f(x - k_0; \sigma)}{\sum_{j=1}^m p_j f(x - \theta_{k_0}; \sigma)} = 1. \tag{5.45}$$

□

Lemma 5.12.

$$\Gamma_k(x; \boldsymbol{p}, \boldsymbol{\theta}) \leq \frac{1}{p_k} \tag{5.46}$$

Proof. Since $f(x) > 0$,

$$\Gamma_k(x; \boldsymbol{p}, \boldsymbol{\theta}) = \frac{f(x - \theta_k; \sigma)}{\sum_{j=1}^m p_j f(x - \theta_j; \sigma)} \leq \frac{f(x - \theta_k; \sigma)}{p_k f(x - \theta_k; \sigma)} = \frac{1}{p_k}. \tag{5.47}$$

□

Lemma 5.13. *Let $\gamma(x)$ be a non-negative function that satisfies*

$$\frac{1}{n} \sum_{i=1}^n \gamma(x_i) = 1. \tag{5.48}$$

Then the θ that minimizes

$$\frac{1}{n}\sum_{i=1}^{n}(x_i-\theta)^2\gamma(x_i) \tag{5.49}$$

is

$$\theta = \frac{1}{n}\sum_{i=1}^{n}x_i\gamma(x_i). \tag{5.50}$$

Proof. CURRENTLY UNUSED. PROOF IN TIM'S NOTEBOOK. □

5.5.3 Bounding C_m

Theorem 5.14. *THM AND PROOF IS BELOW*

Let $\boldsymbol{x} = (x_1,\ldots,x_n)$ and assume that the maximum likelihood mixture for $\boldsymbol{x}$ has no more than m components. Let $\hat{\boldsymbol{\theta}}$ and $\hat{\boldsymbol{p}}$ denote this maximizing mixture. Then by Lemma 5.11, there exists a k^* such that $\Gamma_{k^*}(x_i;\hat{\boldsymbol{p}},\hat{\boldsymbol{\theta}}) \geq 1$ for at least $\lceil \frac{n}{m} \rceil$ different x_i. Let A_{k^*} be the set of all these x_i. Let $A_{\theta_{k^*}}$ be the set of the $|A_{k^*}|$ closest x_i to θ_{k^*}. Then

$$\begin{aligned}\frac{1}{n}\sum_{i=1}^{n}(x_i-\theta_{k^*})^2\Gamma_{k^*}(x_i;\hat{\boldsymbol{p}},\hat{\boldsymbol{\theta}}) &\geq \frac{1}{n}\sum_{i\in A_{\theta_{k^*}}}(x_i-\theta_{k^*})^2 && (5.51)\\ &\geq \frac{1}{n}\left\lceil\frac{n}{m}\right\rceil \operatorname{Var}\left(A_{\theta_{k^*}}\right) \qquad \text{(Biased Variance)} && (5.52)\end{aligned}$$

From (5.41),

$$\operatorname{Var}\left(A_{\theta_{k^*}}\right) \leq \frac{n\sigma^2}{\lceil \frac{n}{m}\rceil}. \tag{5.53}$$

This means that if we cannot find a subset of the x_i that has at least $\frac{n}{m}$ elements and has (biased) variance less than $\frac{n\sigma^2}{\lceil \frac{n}{m}\rceil}$ then we need more than m components in our maximum likelihood mixture.

5.5.4 A particular class of optimization problem

"The results follow from this general theorem which seems obvious."

Theorem 5.15. *Let $(E_m)_{m=1}^{\infty}$ be a sequence of appropriately defined sets and let $(g_m)_{m=1}^{\infty}, g_m : E_m \mapsto \mathbb{R}$ be a sequence of functions that satisfy the following properties*

1. *$\forall \boldsymbol{x} \in \partial E_m, \exists n < m, \boldsymbol{y} \in E_n$ such that $g_m(\boldsymbol{x}) \leq g_n(\boldsymbol{y})$.*

2. *$\exists m_0, \boldsymbol{x}_0 \in E_{m_0}$ such that $\forall m, \boldsymbol{x} \in E_m$, $g_m(\boldsymbol{x}) \leq g_{m_0}(\boldsymbol{x}_0)$.*

Then $\exists m_*, \boldsymbol{x}_* \in E_{m_*} \setminus \partial E_{m_*}$ *such that* $\forall m, \boldsymbol{x} \in E_m$, $g_m(\boldsymbol{x}) \leq g_{m_*}(\boldsymbol{x}_*)$.

Proof. The proof is simple. If $\boldsymbol{x}_0 \notin \partial E_{m_0}$ then we are done. Otherwise, by property 1 we can find a n and $\boldsymbol{y} \in E_n$ such that $g_n(y) = g_{m_0}(\boldsymbol{x}_0)$. If $\boldsymbol{y} \notin \partial E_n$ then we are done, otherwise we repeat the process until we find a $m, \boldsymbol{x}$ pair with $\boldsymbol{x} \notin \partial E_m$. □

5.5.5 Derive Constraints again

WE SHOULD BE ABLE TO DERIVE (5.39) THROUGH (5.41) AGAIN USING THEOREM 5.15.

5.5.6 All points separated by α

Consider the situation in which $|x_i - x_j| > \alpha$ for all $i \neq j$. Intuitively, we would expect that there is some α^* such that if $\alpha > \alpha^*$ then $\boldsymbol{x} \in C_n$.

Theorem 5.16. *If our component density is unimodal and symmetric about zero, and*

$$\frac{f(\alpha/2)}{f(0)} < \frac{1}{n}\left(\frac{n-1}{n}\right)^{n-1}.$$

Then if $|x_i - x_j| > \alpha$ *for all* $i \neq j$, $\boldsymbol{x} \in C_n$.

Proof. Let $\hat{f}_{n-1}$ be the maximum likelihood mixture density of $\boldsymbol{x}$ with no more than $n-1$ components and let L_{n-1} be the corresponding likelihood. Since all the x_i are separated by at least α, there exists an x_{i^*} such that $|x_{i^*} - \theta_j| > \frac{\alpha}{2}$ for all j. Hence

$$\hat{f}_{n-1}(x_{i^*}) < f(\alpha/2)$$

and so

$$L_{n-1} < f(\alpha/2) \prod_{i \neq i^*} \hat{f}_{n-1}(x_i).$$

We will now construct a mixture density that has one more component that $\hat{f}_{n-1}$. We do this by scaling all the components of $\hat{f}_{n-1}$ by a factor of $\frac{n-1}{n}$ and introducing a new component with parameters $(p, \theta) = (\frac{1}{n}, x_{i^*})$. Call this function f_n^* and the corresponding likelihood L_n. Now

$$L_n > \frac{f(0)}{n}\left(\frac{n-1}{n}\right)^{n-1} \prod_{i \neq i^*} \hat{f}_{n-1}(x_i).$$

So if $f(\alpha/2) < \frac{f(0)}{n}\left(\frac{n-1}{n}\right)^{n-1}$ then $L_n > L_{n-1}$ and so $\boldsymbol{x} \in C_n$. $\square$

5.5.7 Discussion about what we hope to acheive

The few original results above (Theorems 5.8 and 5.14) seem to be special cases of what looks to be a much more general rule. Theorem 5.14 seems to be too large by a factor of m when you compare to numerics, and the distance between inflection points in Theorem 5.8 seems to come up again when when you look at images like Figure 5.5 (eg the thickness of the 'bands' is this distance). It is therefore our hope that we can either generalize or add significantly to the Theorems stated so far.

Chapter 6

Deconvolution

6.1 Introduction

From here to the end of Section 6.1.2 is a summary of [39]. We want to find the distribution of a random variable X but only measure

$$W = X + U$$

where U is symmetric (and hence $\phi_U(t)$ is real-valued and even). We also additionally require that $\phi_U(t) \geq 0$. We write

$$\rho_X = \frac{\phi_X}{|\phi_X|}$$

for the phase function of X. Then

$$\begin{aligned} \phi_W &= \phi_X \phi_U && \text{as } X \text{ and } U \text{ are independent,} \\ \frac{\phi_W}{|\phi_W|} &= \frac{\phi_X}{|\phi_X|} \frac{\phi_U}{|\phi_U|}, \\ \rho_W &= \rho_X && \text{as } \phi_U \text{ is real and non-negative.} \end{aligned}$$

Given a probability distribution, there are an infinite number of other distributions that have the same phase function. We make the choice that out of all the distributions with phase function ρ_W, we choose the one that has smallest variance. Hence, we want to find a distribution F_Y that minimizes $\mathrm{Var}(Y)$ such that

$$\rho_Y = \rho_W.$$

6.1.1 Optimization problem

Ideally, we would like to minimize the variance of Y under the constraint that $\rho_Y = \rho_W$. However, we can't do this since we only estimate $\rho_W(t)$ from a random sample of size n and this estimate is bad for large $|t|$. So we instead choose a Y_0 to minimize

$$T(Y) = \int_{-\infty}^{\infty} \left| \hat{\phi}_W(t) - |\hat{\phi}_W(t)|\rho_Y(t) \right|^2 w(t)\,\mathrm{d}t \tag{6.1}$$

where $w(t)$ is some suitably chosen weight function and $\hat{\phi}_W(t)$ is our empirical estimate for $\phi_W(t)$. We then search for Y which minimizes $\mathrm{Var}(Y)$ subject to $T(Y) \leq T(Y_0)$.

We restrict our search to Y discrete with point masses p_j at locations x_j for $j = 1, 2, \ldots, m$. We place our x_j uniformly at random along the interval $[\min W, \max W]$ and choose the p_j to solve the optimization problem described above. Numerical investigations indicate that $m = 5\sqrt{n}$ is a reasonable choice.

6.1.2 Kernel Smoothing

Once we have our discrete distribution Y we can create a continuous density approximation using

$$\hat{f}_Y(x) = \frac{1}{2\pi} \int e^{-itx} \phi_Y(t) \phi_K(ht)\,\mathrm{d}t \tag{6.2}$$

where K is some kernel with bandwidth h. This is exactly equivalent to

$$\hat{f}_Y(x) = \sum_{j=1}^{m} p_j K_h(x - x_j). \tag{6.3}$$

However, we can get a better result by using (6.2) and replacing $\phi_Y(t)$ with an appropriate ridge function for $t \geq t^*$.

6.2 Examples and Relation to Mixture Phenomenon

6.3 General Theorem?

6.4 R Package

Bibliography

[1] Ernst Ising. Beitrag zur theorie des ferromagnetismus. *Zeitschrift für Physik*, 31 (1):253–258, February 1925.

[2] Wilhelm Lenz. Beiträge zum verständnis der magnetischen eigenschaften in festen körpern. *Phys. Z.*, 21:613–615, 1920.

[3] Sacha Friedli and Yvan Velenik. *Statistical Mechanics of Lattice Systems: A Concrete Mathematical Introduction*. Cambridge University Press, November 2017.

[4] R Peierls. On ising's model of ferromagnetism. *Math. Proc. Cambridge Philos. Soc.*, 32(3):477–481, October 1936.

[5] Lars Onsager. Crystal statistics. i. a Two-Dimensional model with an Order-Disorder transition. *Phys. Rev.*, 65(3-4):117–149, February 1944.

[6] Mark Jerrum and Alistair Sinclair. Polynomial-time approximation algorithms for the ising model. *SIAM J. Comput.*, 22(5):1087–1116, 1993.

[7] James Gary Propp and David Bruce Wilson. Exact sampling with coupled markov chains and applications to statistical mechanics. *Random Structures & Algorithms*, 9(1-2):223–252, 1996.

[8] Andrea Collevecchio, Eren Metin Elci, Timothy M Garoni, and Martin Weigel. On the coupling time of the Heat-Bath process for the Fortuin-Kasteleyn Random-Cluster model. *J. Stat. Phys.*, 170(1):22–61, January 2018.

[9] David A Levin, Yuval Peres, and Elizabeth L Wilmer. *Markov chains and mixing times*. American Mathematical Society, 2009.

[10] Jian Ding, Eyal Lubetzky, and Yuval Peres. The mixing time evolution of glauber dynamics for the Mean-Field ising model. *Commun. Math. Phys.*, 289(2):725–764, July 2009.

[11] F Martinelli and E Olivieri. Approach to equilibrium of glauber dynamics in the one phase region. *Commun.Math. Phys.*, 161(3):447–486, April 1994.

[12] Eyal Lubetzky and Allan Sly. Cutoff for the ising model on the lattice. *Invent. Math.*, 191(3):719–755, March 2013.

[13] Eyal Lubetzky and Allan Sly. An exposition to information percolation for the ising model. *Ann. Fac. Sci. Toulouse Math.*, 24(4):745–761, 2015.

[14] Eyal Lubetzky and Allan Sly. Information percolation and cutoff for the stochastic ising model. *J. Amer. Math. Soc.*, 29(3):729–774, 2016.

[15] Eyal Lubetzky and Allan Sly. Universality of cutoff for the ising model. *Ann. Probab.*, 45(6A):3664–3696, November 2017.

[16] Patrick Billingsley. *Probability and Measure*. Wiley Series in Probability and Mathematical Statistics. John Wiley & Sons, 1995.

[17] Olle Häggström. *Finite Markov Chains and Algorithmic Applications*. Cambridge University Press, May 2002.

[18] Mark Jerrum. Mathematical foundations of the markov chain monte carlo method. In Michel Habib, Colin McDiarmid, Jorge Ramirez-Alfonsin, and Bruce Reed, editors, *Probabilistic Methods for Algorithmic Discrete Mathematics*, pages 116–165. Springer Berlin Heidelberg, Berlin, Heidelberg, 1998.

[19] P Diaconis. The cutoff phenomenon in finite markov chains. *Proc. Natl. Acad. Sci. U. S. A.*, 93(4):1659–1664, February 1996.

[20] David Aldous. Random walks on finite groups and rapidly mixing markov chains. In *Séminaire de Probabilités XVII 1981/82*, pages 243–297. Springer Berlin Heidelberg, 1983.

[21] Danny Nam and Allan Sly. Cutoff for the Swendsen-Wang dynamics on the lattice. May 2018.

[22] P Erdős and A Renyi. On a classical problem of probability theory. *Publ. Math. Inst. Hung. Acad. Sci.*, Ser. A 6:215–219, 1961.

[23] A D Barbour and O Chryssaphinou. Compound poisson approximation: a user's guide. *Ann. Appl. Probab.*, 11(3):964–1002, August 2001.

[24] A D Barbour, Louis H. Y. Chen, and Wei-Liem Loh. Compound poisson approximation for nonnegative random variables via stein's method. *Ann. Probab.*, 20(4): 1843–1866, 1992.

[25] Geoffrey McLachlan and David Peel. *Finite Mixture Models*, volume 44. John Wiley & Sons, April 2004.

[26] Karl Pearson. Contributions to the mathematical theory of evolution. *Philos. Trans. R. Soc. Lond. A*, 185:71–110, 1894.

[27] Carey E Priebe. Adaptive mixtures. *J. Am. Stat. Assoc.*, 89(427):796–806, September 1994.

[28] Bruce G Lindsay. The geometry of mixture likelihoods: A general theory. *Ann. Stat.*, 11(1):86–94, March 1983.

[29] Bruce G Lindsay. The geometry of mixture likelihoods, part II: The exponential family. *Ann Stat*, 11(3):783–792, September 1983.

[30] Bruce G Lindsay and Kathryn Roeder. Uniqueness of estimation and identifiability in mixture models. *Can. J. Stat.*, 21(2):139–147, June 1993.

[31] Ulf Grenander. *Abstract inference*. Wiley, New York, 1981.

[32] H Akaike. A new look at the statistical model identification. *IEEE Trans. Automat. Contr.*, 19(6):716–723, December 1974.

[33] Gideon Schwarz. Estimating the dimension of a model. *Ann. Stat.*, 6(2):461–464, March 1978.

[34] Brian G Leroux. Consistent estimation of a mixing distribution. *Ann. Stat.*, 20(3): 1350–1360, September 1992.

[35] Bruce G Lindsay. *Mixture Models: Theory, Geometry, and Applications*. IMS, 1995.

[36] Arseni Seregin. Uniqueness of the maximum likelihood estimator for k-monotone densities. *Proc. Am. Math. Soc.*, 138(12):4511–4515, 2010.

[37] Zhengmin Zhang. *Estimation in Mixture Models*. PhD thesis, Carleton University, 2008.

[38] Dankmar Böhning. A review of reliable maximum likelihood algorithms for semiparametric mixture models. *J. Stat. Plan. Inference*, 47(1):5–28, October 1995.

[39] Aurore Delaigle and Peter Hall. Methodology for non-parametric deconvolution when the error distribution is unknown. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 78(1):231–252, January 2016.